TNO Health Research

TNO Institute of Preventive
Health Care

A Reader on Applying Statistics in Public Health and Prevention

August 1992

J.L.A. van Rijckevorsel
C.C.J.H. Bijleveld

Netherlands Institute of Preventive Health Care NIPG-TNO
P.O. Box 124
2300 AC Leiden
the Netherlands

Printed and published by NIPG-TNO
NIPG publication number 92.036

This report can be ordered from NIPG-TNO by transferring *f* 49,50 or equivalent (without costs to the recipient) to account number 56.70.25.500 of HGO-TNO Leiden at the Algemene Bank Nederland, Geversstraat 43, 2341 GA Oegstgeest, the Netherlands. (Within the Netherlands: giro account number of this bank is 9013).
Please state "NIPG publication number 92.036" and your name and address. Fee includes handling and mailing.

CIP data for the Royal Library, the Hague, the Netherlands:

Titel A Reader on Applying Statistics in Public Health & Prevention
Leiden: TNO Institute of Preventive Health Care
Incl. references
ISBN 90-6743-212-1

CONTENTS page

LIST OF CONTRIBUTORS

A. Dijkstra, Deputy-director, TNO Institute of Preventive Health Care, P.O. Box 124, 2300 AC Leiden, The Netherlands.

F. Andries, Department of Work Research, TNO Institute of Preventive Health Care, P.O. Box 124, 2300 AC Leiden, The Netherlands.

C.C.J.H. Bijleveld, Department of Psychometrics, Faculty of Social Sciences, Leiden University, P.O. Box 9555, 2300 AB Leiden, The Netherlands.

S. van Buuren, Department of Statistics and Computer Science, TNO Institute of Preventive Health Care, P.O. Box 124, 2300 AC Leiden, The Netherlands.

J.L.A. van Rijckevorsel, Department of Statistics and Computer Science, TNO Institute of Preventive Health Care, P.O. Box 124, 2300 AC Leiden, The Netherlands.

A.G.C. Vogels, Department of Children and Health, TNO Institute of Preventive Health Care, P.O. Box 124, 2300 AC Leiden, The Netherlands.

M.M. van der Klaauw, Department of Statistics and Computer Science, TNO Institute of Preventive Health Care, P.O. Box 124, 2300 AC Leiden, The Netherlands.

J.K.S. van Ginneken, Netherlands Interdisciplinary Demografic Institute, P.O. Box 11650, 2502 AR The Hague, The Netherlands.

A. Bloemhoff, Department of Occupational Health Research, TNO Institute of Preventive Health Care, P.O. Box 124, 2300 AC Leiden, The Netherlands.

J.E.F.M. Frencken, Department of Children and Health, TNO Institute of Preventive Health Care, P.O. Box 124, 2300 AC Leiden, The Netherlands.

M.A. van 't Hof, Dept. of Medical Statistics, University of Nijmegen, Kapittelweg 54, 6525 EP Nijmegen, The Netherlands.

J.J. Radder, Department of Statistics and Computer Science, TNO Institute of Preventive Health Care, P.O. Box 124, 2300 AC Leiden, The Netherlands.

E. Wortel, Department of Occupational Health Research, TNO Institute of Preventive Health Care, P.O. Box 124, 2300 AC Leiden, The Netherlands.

F.H.G. Marcelissen, Digital Equipment Enterprice B.V., Vonderweg 11, 6500 MD, Eindhoven, The Netherlands.

D.J. van Putten, Department of Occupational Health Research, TNO Institute of Preventive Health Care, P.O. Box 124, 2300 AC Leiden, The Netherlands.

D.N. Geary, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, United Kingdom.

R.J. Gilbert, **E. Hutington**, Unilever Research, Port Sunlight Laboratory, Merseyside, United Kingdom.

B. Goldfarb, Laboratoire de Biostatistique et d'Informatique Medicale, Hôpital Necker, 149 Rue de Sèvres, 75015 Paris, France.

P.A. Burrough, Netherlands Expertise Center for Spatial Information Processing, Rijksuniversiteit Utrecht, Postbus 80.115, 3508 TC Utrecht, The Netherlands.

G. Gallus, Instituto di Biometria E Statistica Medica, Facoltà di Medicina e Chirurgia, Università degli Studi, Via Venezian 1, 20133 Milan, Italy.

D.A. Lievesley, ESRC, Data Archive, University of Essex, Wivenhoe Park, Colchester, Essex, CO 4 3SQ, United Kingdom.

PREFACE

If we would try to classify the contributions to this volume by content matter, apart from statistics, we have roughly speaking four papers on working conditions and occupational health, two on sexually transmissible disease, two on dental health, two on public health tables, one on the geographical distribution of diseases, one on intervention and one on safe behaviour. All papers either directly or indirectly deal with prevention, which is the classical field of The TNO Institute of Preventive Health Care. This reader illustrates the conceptual but also the methodological diversity of the field. Research on Public Health and Prevention does not have its own methodology, it has many, as Dijkstra points out in the first chapter. This reader shows why. Estimating rare health events as Gallus does in chapter 13 is worlds apart from canonical correlations of working conditions by Bloemhoff (chapter 6). Data are different, methods are different and terminology is different. Nevertheless it is all public health and prevention research. In order to structure this we could try to classify research by method, discerning descriptive from inferential methods in the way correspondence analysis differs from logistic regression or multivariate from univariate methods like canonical correlation from the cusum method etc. Perhaps this would make sense to a statistician but it would hardly impress any practioner in the field of public health. An intuitively appealing idea is the division by kind of data: aggregated numbers versus individual based data. Another common distinction is subjective versus objective measurements: questionnaires and self reports versus physiological measurements. Apparently we can make several partly overlapping partitions of the same field. In classification it is customary to compare several different partitions on one set of objects and look for the underlying common most representative partition. Say we partition the contributions to this volume in various not too overlapping ways in order to illustrate how they can be classified by on what they have in common and what not. This leads to a binary table of contributions (rows) by attributes (columns). A cell can contain a one (=1) which means that a contribution falls in one class or a nought (=0) meaning that it does not belong to that class. Because the contributions by Dijkstra and Lievesley are review papers we will not consider them for classification. Additionally we have left out the contribution by Vogels et al. because

it deals with the construction of a statistical data base. This leaves us with ten contributions to classify on 11 attributes.
The attributes, with the number of papers they apply between brackets, are:

Data:
- Including time as a variable/factor (4)
- Questionnaire/self reports (5)
- Aggregated data (5)
- Categorical data (4)

Methodology:
- Exploratory analysis (8)
- Confirmatory analysis (7)

Type of Analysis:
- Multivariate analysis (6)
- Univariate analysis (4)

Miscellaneous:
- Graphical tools (6)
- Comparison of methods (5)

and the contributions, named after the first author are:

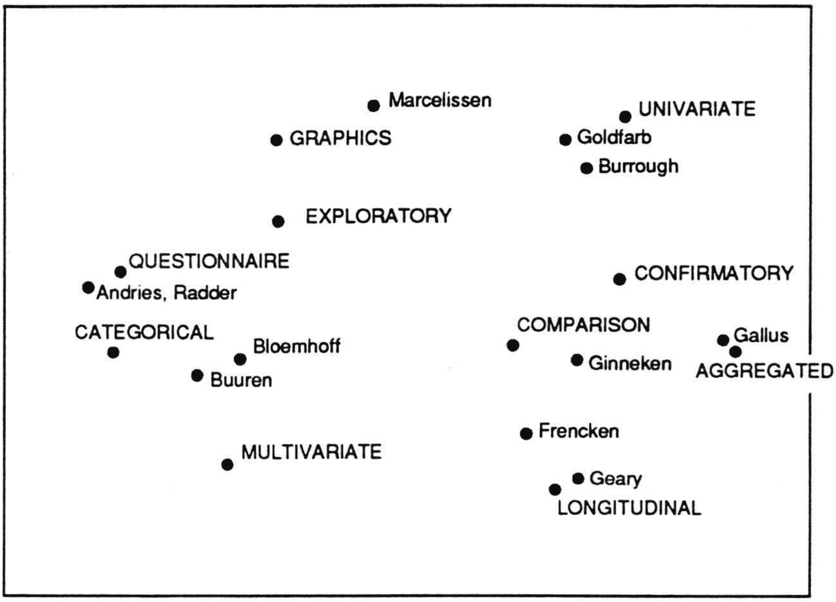
Andries (5)
Buuren (7)
Ginneken (2)
Bloemhoff (6)
Frencken (5)
Radder (5)
Marcelissen (5)
Geary (4)

Goldfarb (5)
Burrough (6)
Gallus (4)

The number of attributes the papers have in common with other papers are between brackets.

If this book would contain a representative sample of the methodologies used in this field then an analysis of this matrix would reveal something of the way this methodology was structured, provided we had chosen the proper classification categories. Both assumptions are hard to prove to be true. Good methodological practice tells us to explore first. Correspondence analysis (see the first chapter by Andries et al.) is a multivariate exploratory tool that looks for the common partition that simultaneously classifies the attributes and the contributions. Well, the quantified exploratory classification matrix by correspondence analysis represent authors and attributes by points in a common space. The first two dimensions of such a space are presented by figure i. Points that are close have a lot in common and points that are distant have not much in common. This means that somebody who wants a quick opinion on the kind of papers this reader contains and who is willing to accept the proposed set of classifying attributes can do with a mere inspection of figure i.

Figure i Contributions to this volume and their attributes



For instance Radder and Andries focus on the exploratory multivariate analysis of questionnaires, Marcelissen is less multivariate but strongly graphically oriented, van Ginneken mainly compares methods for analyzing aggregated data, Geary and Frencken include longitudinal data etc. For a more thorough opinion one should of course read the papers, but as a substitute for eye-balling, diagonal reading and what other euphemisms exist for a quick glance, the inspection of figure i. might do.

Looking at the plot the main distinction between the papers is that they are either exploratory, multivariate analyses of questionnaires and/or categorical data using graphical tools on the left hand side or a fairly heterogeneous lot of confirmatory analyses on the right hand side. The methodology and research methods for analyzing questionnaires clearly stems from the social sciences, while the right hand contributions have a more heterogeneous origin. The relative quality of the representation is expressed by the percentages explained inertia per axis. For this table this amounts to 70% by the first two axes.

Acknowledgements

We would like to mention in the first place Renée Moraal who gave secretarial and organisational assistance and was indispensable for the production of this reader. Furthermore we would also like to thank Stef van Buuren and Jan Radder for reviewing some of the manuscripts.

The Editors

MULTIVARIATE ANALYSIS IN PUBLIC HEALTH RESEARCH

A. Dijkstra, deputy-director NIPG-TNO*

* TNO Institute of Preventive Health Care, Leiden, The Netherlands

Abstract

Public health is at crossroads where epidemiology and the social sciences meet. Their theories, concepts and operationalizations show remarkable differences. Especially with respect to multivariate analysis, epidemiologists and social scientists seem to embrace their 'own' methods. Both disciplines, however, share the same methodology of the empirical sciences. The variety in methods has impressively increased during the last decade. This presentation illustrates, that in the same period, the application of multivariate designs in public health research has strongly increased too. In addition, data will be presented about the increasing number of different multivariate methods applied in public health research. It is of strategic importance for public health that, between disciplinary groups, methodological innovations are exchanged. Differentiation and specialization per discipline can hamper fruitful application of epidemiological and social scientific research results in public health. TNO Institute of Preventive Health Care is establishing an interdisciplinary body of methodology, based in both epidemiology and in social sciences, to be implemented in public health, disease prevention and health promotion. This interdisciplinary workshop aims to promote cross fertilization by mutual understanding of each other's research methods.

Keywords: *multivariate analysis, multidisciplinary research, epidemiology, public health research.*

1.1 PUBLIC HEALTH, EPIDEMIOLOGY AND SOCIAL SCIENCE

About seventy years ago Winslow (1920) defined public health as "...the art and the science of preventing disease, prolonging life, and promoting physical health and efficiency through organised community effort...".
According to Clark (1981), public health is a social movement concerned with protecting and promoting the collective health of a community. Its goals are not absolute or unchanging. They represent the dominant values of a society such as religion, humanitarianism, or scientific interest, as well as economic and social utility (Rosen, 1972). The shape of public health is a reflection of the society whereto it belongs. It is an art and a science, as well as an agent of social action and of social control. From this specific field concerning the health of a community, the research questions arise. Public health can claim a key role in pioneering data-based approaches to the solution of health and community problems. It is at crossroads where epidemiology and social science meet. In the field of health and health care, epidemiologists and social scientists both share interest in factors influencing disease and physiological conditions (as epidemiologists put it) or health and illness (as social scientists put it). This distinction has been convincingly presented by Spruit (in press), based on an examination of a definition of epidemiology (Lilienfeld and Lilienfeld, 1980) and a definition of medical sociology (Armstrong, 1980) as applied to medicine. According to Spruit epidemiology and medical sociology for instance share:

- the determination of risk factors for ill-health, the description and explanation of the distribution of ill-health in populations and the identification of high risk groups, aspects of health care in general and of medical care in particular.
- a common interest in themes like demographic trends, comparative studies of societies, social and economic groups, age-related illnesses, the result of medical care or therapeutical regimes, the consequences of chronic disease and the need for health care
- the recognition of the multicausality of disease origins and
- a common understanding and approach to scientific research in populations and the use of statistics as a method to reach this understanding.

Both disciplines, however, apply their own theories, concepts and measurements. The epidemiologist prefers to study individual health conditions, which can be defined clearly and unambiguously as "disease" or no "disease". In fact however, not all diseases can be defined and measured clearly and unambiguously according sound biomedical standards. In addition not all diseases are experienced as illness or as ill-health. Further it has to be kept in mind that many illnesses can not be identified with a clearly defined disease. The concept of illness includes a disease and the individual's experience as well. On this borderline of disease and illness, with regard to the prevalence as well as to the etiology and to the social consequences (i.e. medical consumption) the epidemiologist and the social scientist can closely work together. This cooperation is especially important when behaviourial -, social - and environmental risk factors of disease and illness have to be studied. The social scientist is trained to develop methods - i.e questionnaires concerning subjective health, loneliness, stressful live events, social support network and so on - which fit the specific aim of the research questions involved. The fact that the biomedical difference between disease and physiological condition is so much more marked (disease: "no", "yes") than the social scientific difference between degrees of illness, ill-health or health (3- or more point rating scale, sum of wide range of health beliefs and so on) is also reflected in the application of statistical methods. Also with respect to multivariate analysis, the simultaneous analysis of more than two variables, epidemiologists and social scientists seem to embrace their 'own' methods. Such a phenomenon is perhaps functional with respect to the development of the scientific discipline. As soon as scientific separatism hampers the mutual understanding of each other's research results, the improvement of public health will be endangered. The origins of diseases are multicausal and thus public health problems cross disciplinary borders. Consequently, they can only be solved in a multidisciplinary way. Multidisciplinary cooperation between epidemiology and the social sciences requires at least a certain degree of mutual understanding of each other's research methods. In practice however, it is hardly possible to master all different research methods. This becomes even more difficult since multivariate analysis itself becomes increasingly differentiated and specialized. This presentation aims to illustrate the growing importance of multivariate analysis in public health research and underlines the importance of mutual exchange of new developments in a multidisciplinary workshop like this.

1.2 MULTIVARIATE ANALYSIS IN PUBLIC HEALTH JOURNALS

To illustrate whether the importance of multivariate analysis in Public Health Research is increasing, a sample of empirical research articles has been drawn from all articles published in four public health journals during the period of 1978 to 1989. Multivariate methods analyze three or more variables simultaneously. In bivariate analyses two variables are involved and a univariate model contains one variable, for instance a frequency distribution with respect to different measurements in time. Examples of bivariate methods are one way analysis of variance, regression analysis, odds ratio's, t-test, non parametric tests. Multivariate methods are by example: multiple regression analysis, odds ratio's, relative risks "corrected" for confounding, loglinear models with three or more variables, multiple correspondence analysis (MCA), principal component analysis, multivariate analysis of variance.
The journals which have been selected are: the American Journal of Public Health (APH), the Journal of Epidemiology and Community Medicine (ECM), the Scandinavian Journal of Social Medicine (SSM) and the Dutch Journal of Public Health / Tijdschrift voor Sociale Gezondheidszorg (TSG). Together, these four journals reflect a considerable amount of American, British and European public health research. It was not the intention to generalize the results to the whole field concerned. It is mainly aimed to get an empirical based impression of developments with regard to the kind of methods applied in public health research. Three periods of measurement have been chosen: 1978-1979; 1984-1985 and 1988-1989. To avoid too small frequencies per period of measurement, each period counts two calendar years.

In the first period of measurement (1978-1979) the number of articles per journal varies from 40 to 143, in the second period from 50 to 207 and in the last period from 83 to 268 (table 1.1).

Table 1.1 Multivariate methods applied in articles published in four Public Health Journals; periods of measurement: 1978-1979; 1984-1985; 1988-1989

	APH			ECM			SSM			TSG		
	t1 2 n=143	t2 n=207	t3 n=268	t1 n=106	t2 n=130	t3 n=127	t1 n=40	t2 n=50	t3 n=83	t1 n=131	t2 n=149	t3 n=115
Research Articles (% RA)	81	93	96	79	91	95	88	88	82	45	54	69
Research Articles with multi-variate methods (% MVA)•	26	37	46	32	29	53	9	32	28	24	16	39
Number of different statistical methods (abs.)	15	41	48	14	25	33	3	9	13	8	11	20

APH = American Journal of Public Health, ECM = Journal of epidemiology and Community Medicine, SSM = Scandinavian Journal of Social Medicine, TSG = Tijdschrift voor Sociale Gezondheidszorg (Dutch Journal of Public Health).
t1 = 1978-1979, t2 = 1984-1985, t3 = 1988-1989.
• Denominator = number of research articles.

Three journals show an increase in the number of articles published in this period from 1978 to 1989. The decrease in TSG, in the period of 1988-1989, has been caused by the reduction from 24 to 12 issues per year. In behalf of our analysis all articles based on empirical research data were selected. The relative proportion of research articles (% RA) can be found in table 1.1. It is shown clearly, that the absolute and relative number of research articles have been increasing during the last decade.

Table 1.2 Overview of all multivariate methods applied in research articles published in four public health journals (APH, ECM, SSM, TSG) during the period 1978-1989; alphabetical order

Period	Journal	Multivariate method
1978-1979	American Journal of Public Health	canonic correlation, covariance analysis, discriminance analysis, factor analysis, loglinear + multiple regression, multiple classification analysis MCA + multiple regression, multiple discriminant analysis, multiple regression, multiple regression + partial correlation, multiple variance analysis, multivariate analysis, multivariate stratification with confounders, partial correlation, path models + special regression models, unknown, variance + factor + discriminance analysis
	Journal of Epidemiology and Community Health	covariance analysis, discriminance analysis, linear modelling, multiple logistic equation with confounders, multiple regression, multiple variance (ANOVA), multivariate discriminance analysis + covariance analysis + multiple logistic regression, multivariate modelling, multivariate variance analysis, odds ratio with confounders, partial correlation, principal component analysis, relative risk with confounders, relative risk with confounders + loglinear >2
	Scandinavian Journal of Social Medicine	covariance analysis, multiple regression, stepwise multiple regression analysis + automatic interaction detector AID
	Tijdschrift voor Sociale Gezondheidszorg	discriminance analysis, factor analysis, multifactor analysis, multiple correlation, multiple variance analysis, partial correlation, partial correlation + orthogonal factor analysis, stepwise discriminance analysis variance analysis + discriminance analysis
1984-1985	American Journal of Public Health	covariance analysis, covariance analysis (ANCOVA), covariance analysis + logistic regression model, covariance analysis + multiple regression, covariance analysis + multivariate regression, discriminance analysis, linear logistic models, logistic model, logistic regression model, logistic regression model (multivariate), logistic regression model (odds), loglinear analysis (odds ratio), loglinear model (Bishop), multiple classification analysis, multiple correspondence analysis, multiple covariance analysis (MANOVA), multiple logistic analysis, multiple logistic regression, multiple regression, multiple regression + (ANOVA - MCA), multiple regression + factor analysis, multivariate analysis, multivariate analysis: variance-/covariance analysis, multivariate logistic regression analysis, multivariate modelling, odds ratio with confounders, odds ratio with confounders + logistic regression analysis, odds ratio with confounders + multiple logistic regression, odds ratio + relative risk with confounders, partial correlation, Polliser and Diehr + separate regression variables analysis, principal components analysis, principal components factor analysis with varimax rotation yielded four factors, relative risk + logistic regression model, relative risk with confounders, relative risk with confounders + controlled by logistic regression, relative risk with multiple confounders + multiple logistic model, stepwise regression analysis, variance analysis, variance analysis + covariance analysis
	Journal of Epidemiology and Community Health	cluster analysis, covariance analysis, Cox proportional hazards general linear model (regression), discriminance analysis, linear logistic regression models + effects + confounders, linear multivariate regression, logistic regression with confounders, loglinear analysis, multiple correlation, multiple linear regression + covariance analysis, multiple logistic regression, multiple regression, multiple variance analysis, multivariate analysis, multivariate analysis with confounders, multivariate analysis: conditional logistic regression procedure, odds ratio + relative risk with confounders, odds ratio with confounders, odds ratio with confounders + logistic regression, odds ratio with confounders + regression analysis, principal components analysis, relative risk + adjusted, relative risk with confounders, relative risk with confounders + multivariate logistic regression, temporal clustering
	Scandinavian Journal of Social Medicine	covariance analysis (ANCOVA), discriminance analysis, multiple regression, multiple least square regression analysis, multiple regression analysis SRA, multivariate analysis + likelihood test,

Table 1.2 (continued)

Period	Journal	Multivariate method
		multivariate regression analysis using odds ratio, MVA=automatic interaction detector, MVA=multivariate confounder score + multiple logistic model + likelihood ratio statistic + factor analysis
	Tijdschrift voor Sociale Gezondheidszorg	anoto analysis >2 variables, factor analysis, factor analysis + partial correlation, LISREL technique, logistic regression + odds ratio + corrected, multiple correlation, multiple regression, multiple regression model, multivariate analysis, partial correlation + variance analysis (ANOVA) + covariance, variance analysis >2
1988-1989	American Journal of Public Health	cluster analysis, covariance analysis, covariance technique, Cox proportional hazards regression model, Cox hazards regression model, least square analysis of multivariate data, logistic models, logistic model with confounding, logistic regression analysis, logistic regression >2, logistic regression (confounding effect), logistic regression with confounders, logistic regression model + controlling variables, logistic regression model for multivariate analysis, logistic regression + odds ratio with confounders, loglinear modelling, MANOVA, multiple correlational analysis, multiple linear regression, multiple linear logistic regression, multiple logistic model, multiple logistic regression, multiple logistic regression + factor analysis, multiple regression, multiple regression equation, multivariate + bias regression analysis, multivariate analysis + multiple logistic models, multivariate analysis (construction new variables), multivariate analysis using OLS regression procedures, multivariate logistic regression, multivariate model + multiple logistic regression, multivariate: single regression + adjusted, odds ratio adjusted for confounders using multiple logistic regression, odds ratio adjusting for confounding bias via multiple logistic regression analysis, odds ratio controlling logistic regression >2 + logistic regression + control for confounding, odds ratio + controlling + final logistic model, odds ratio + loglinear model, odds ratio + multiple logistic regression, odds ratio with confounders, odds ratio with confounders + multiple logistic regression, principal factor analysis + varimax rotation + logistic analysis, regression analysis, regression analysis + controlling variables, relative risk with confounders, relative risk with confounders + hazard regression, relative risk with confounders + regression modelling, relative risk with confounders + logistic regression, stepwise regression model
	Journal of Epidemiology and Community Health	Boze Jenkins technique, cluster analysis, covariance analysis (MANOVA), covariance analysis + Cox linear logistic regression, Cox proportional hazards regression, grouped logistic regression, linear logistic regression, logistic regression analysis + adjusted, logistic modelling, logistic regression analysis, logistic regression model, loglinear modelling, loglinear modelling + relative risk + adjusted, multiple logistic regression, multiple logistic regression (adjusted) (Mantel Haenszel), multiple regression analysis, multiple variance analysis (ANOVA), multivariate analysis (MANOVA), multivariate analysis: loglinear model: logistic regression model, multivariate discriminance analysis + multiple logistic regression model, multivariate logistic analysis, odds ratio + adjustment (Mantel Haenszel), odds ratio + logistic regression, odds ratio controlled + logistic modelling, odds ratio with confounders (Mantel Haenszel), odds ratio estimate from the logistic regression analysis was used to adjust for confounding variables, regression analysis, relative risk + logistic regression, relative risk + logistic regression + adjustment, relative risk + regression modelling, relative risk with confounders, stepwise logistic analysis (Breslow) + relative risk adjusted, stepwise logistic regression
	Scandinavian Journal of Social Medicine	automatic interaction detection, covariance analysis, factor analysis + regression analysis, logistic regression with confounders, loglinear modelling, loglinear regression GLIM package, multiple regression, multivariate analysis (Mantels test), odds ratio + logistic regression, odds ratio with confounders + multiple regression (logistic), regression analysis, relative

Table 1.2 (continued)

Period	Journal	Multivariate method
		risk controlled for confounders, relative risk with confounders + logistic regression analysis
	Tijdschrift voor Sociale Gezondheidszorg	cluster analysis, components analysis + factor analysis (Promax Varimax), correctional correlations, covariance analysis, factor analysis, factor analysis (principal components analysis), factor analysis + multiple regression, multiple correlation, multiple logistic regression, multiple regression, multiple variance analysis, multivariate analysis, multivariate analysis (MANOVA), odds ratio + corrected, odds ratio with confounders, partial correlation, principal components analysis, principal components analysis + principal factor analysis + logistic regression with controlling variables, principal factor analysis, relative risk with confounders (likelihood ratio test)

Per research article it was established whether one or more multivariate methods had been applied, or just univariate or bivariate analyses.
Each multivariate method was described in order to compare the three periods of measurement with regard to the kind of methods applied. There is no doubt about the increasing importance of multivariate analysis in the research articles, which have been published in our sample of four public health journals (table 1.2).

In the American Journal of Public Health and the Journal of Epidemiology and Community Medicine about 50% of all research articles, nowadays, describe results based on multivariately analyzed data (% MVA). In the Scandinavian Journal of Social Medicine and in TSG these percentages are considerably lower. Compared with the late seventies, however, multivariate analysis has become more prevalent in these European journals too.
In the period of 1978 - 1979 multivariate methods like factor analysis, multiple regression and multivariate models of analysis of variance were dominating. In this period the use of multivariate methods in epidemiology was relatively underdeveloped. Now more than a decade later, the epidemiological research designs are much more present in public health research and by consequence, a lot of new multivariate methods entered the research scene. Logistic models, in a broad variety of applications, have reached an unmistakable level of popularity. Other examples of a relatively new application in public health research are Multiple Correspondence Analysis (MCA), Generalized Additive Models and Survival Analysis. In the period of 1988 - 1989 the American Journal of Public Health about 50 different methods of multivariate analysis could be registered,

35 more than a decade ago (see table 1.1). The increases, shown in the other journals reflect the same development of differentiation and specialization in the field of multivariate research methods. In table 1.2 all different methods from our sample have been described, broken down by period of measurement and journal.

1.3 CONCLUDING REMARKS

All in all, the data illustrate a growing importance of multivariate analysis in public health research. The changes in statistical methods being used in public health journals in the 1980s are in fact dramatically, because it is almost impossible for a researcher to master all different statistical methods which are now available. This phenomenon of increasing complexity in the field of statistics has not been restricted to the four Public Health journals studied in this paper. Altman (1991) concludes the same with respect to the use of statistical methods in the, highly respected, New England Journal of Medicine: "The use of simple methods remained fairly constant, but there was a doubling of the use of linear regression and non-parametric methods and a dramatic surge in the use of more complex methods. The most notably change was in the use of survival analysis (including logistic regression), multiple logistic regression and Cox regression". In addition he argued that this development can be considered as a general tendency. New methodology, more statisticians, cheap computers and software have led to more statistical analyses and more complex analyses. It may be hoped that the improvement and at the same time the increasing complexity of the new statistical methods will improve the level of empirical research. Anyhow, one may expect an exponential increase in misuse of the highly sophisticated methods, if applied without support of a statistician. This support can only be efficient if the researcher and the statistician jointly prepare the research design. In doing so, questions like "This is my data - what shall I do with it" or "Please check my computations" can be avoided. Unfortunately many statistical courses are predominantly restricted to learning of computational activities rather than to train people in the underlying skills and understanding of statistics, or to equip potential clients to collaborate effectively with a statistician (Hawkins, 1992). This author

recalls a plea for parastatisticians: professionals in their own right in other disciplines - not statistical assistants - who also have a substantial statistical understanding.
There is no reason to believe that the development of statistical methods will stop. In epidemiology as well as in the social sciences the tendency to use new, highly sophisticated and complex, statistical methods is irreversible. Whether between these bodies of knowledge, which are both necessary for the improvement of public health, will be a sufficient degree of cross fertilization depends partly on mutual knowledge and understanding of each other's concepts, measurements and last but not least the statistical methods. In fact they all share the same methodological principles of the empirical sciences. At last, only a coordinated effort from public health professionals in the field and researchers, regardless their discipline can solve the health problems in the community.

In TNO Institute of Preventive Health Care, with a substantial amount of multidisciplinary research in prevention and public health, a centrally organized body of methodological and statistical knowledge has been established to improve the quality of research and to prevent misuse of statistical methods: all research designs and all research reports have to be reviewed by a statistician. In addition, this body of multidisciplinary knowledge contributes to cross-fertilization between TNO medical and social scientific applied research in order to bridge the conceptual and methodological gap between epidemiology and social science in public health and prevention.

Acknowledgement
The author wishes to thank Patricia Staats for her assistance, especially with respect to the data collection.

REFERENCES

ALTMAN, D.G. Statistics in medical journals: developments in the 1980s. Stat. Med. 10 (1991) 1897-1913.

ARMSTRONG, D. An outline of sociology as applied to medicine. Wright, Bristol, 1980.

CLARK, D.W. A Vocabulary for preventive and community medicine. In: Clark, D.W. & B. MacMahon, Preventive and Community Medicine. Boston, Little, Brown and Co., 1981. pp. 3-15.

HAWKINS, A. Healthy Parastatisticians ? News & Notes, The Royal Statistical Society 18 (1992) 1-2.

LILIENFELD, A.M. & D.E. LILIENFELD. Foundations of epidemiology. New York, Oxford University Press, 1980.

ROSEN, G. Public Health. International encyclopedia of the social sciences. New York, London, Vol 13. (1972) pp. 164-170.

SPRUIT, I.P. Social factors and disease in epidemiology and medical sociology: some foundations of measurements. In: Tengvald K. (ed.). Problem of Applied Epidemiology: an inner and outer perspective. F.R.N., Stockholm, in press.

WINSLOW, C.E.A. The untilled fields of public health. Science 51 (1920) 23.

WORKING CONDITIONS, STRAIN AND STRESS OF AUTOMATION PERSONNEL: ANALYSIS OF COMPLAINTS AND CONTENTEDNESS

F. Andries*, C.C.J.H. Bijleveld**

* Department of Work Research, TNO Institute of Preventive Health Care, Leiden, The Netherlands

** Department of Statistics and Computer Science, TNO Institute of Preventive Health Care, Leiden, The Netherlands (Presently at Department of Psychometrics, Faculty of Social Sciences, Leiden University, The Netherlands)

Abstract

In 1989 a questionnaire was administered to approximately 3000 persons employed in the automation sector in the Netherlands. The aim of this enquiry was to make an inventory of possible stressful aspects in the career and working conditions of automation personnel. Analyzing the sample's complaints scores with multiple correspondence analysis, occupations-at-risk were identified. It appeared that middle-management occupations are at risk concerning stress; occupations on the level of the lower-management and teachers seem to be at risk concerning career-prospects. In an additional analysis of contentedness scores, it appeared that computer specialists show favourable scores on stress and hectic working conditions; combined with their complaints on autonomy and challenge, the risk of underburdening seems present for this group.

Keywords: *automation personnel, working conditions, mental strain and stress, multiple correspondence analysis.*

2.1 INTRODUCTION

Within the scientific world quite a bit of attention has been paid to the success and, especially, the failure of computerisation projects (Riesewijk & Warnerdam, 1988). With the exception of the working conditions of VDU-operators (Pot et al., 1986; Evans, 1987), the working situation of those who are otherwise employed in the automation sector has received far less attention. Research focused on job preferences of computer specialists, who appeared to show a marked preference for career possibilities and relatively little interest in social aspects of the job (Couger & Zawacki, 1980; Oborne, 1985).

While attention within the automation sector is still mainly focused on the problem of how to get and keep qualified personnel and what to offer them in reward, interviews held in the preparatory phase of the study described here (Andries, 1990) did show a growing concern for the risk of overburdening: the chronic shortage of qualified personnel, high staff-turnover, and the difficulties in planning computerisation projects, create a situation in which the danger of overburdening seems real. The indispensability of certain persons for the progress of projects might in addition induce them to ignore symptoms of physical and mental strain.

Persistent indications of this kind, together with signals of unease about the manner in which the work of automation personnel within the Computer Service Industry in the Netherlands is organised socially (Werkgroep Sociaal Beleid C.S.I., 1988), had led to the initiation of this study into the working situation of automation personnel. Given the lack of data for specific subgroups of workers within the automation sector, a choice was made for an exploratory survey among a broad range of occupations, that are, directly or indirectly, connected with automation processes. In this exploratory set up, the aim was to identify risk factors for stress and strain, for comparing occupations within the group of automation personnel.

2.2 QUESTIONNAIRE AND SAMPLE

In early 1989, a postal questionnaire was distributed among a sample of automation personnel. From the subscribers of the widely read Dutch 'computer-trade' journal "Computable" (read by approximately 80% of the target population), a stratified 10% sample was drawn, providing for, approximately, equal numbers over the 32 classes of occupations used for the registration of subscribers. The same classification of occupations will be used below. Because of the stratified sampling over occupations, the response-group, though a considerable proportion of the target population, can not be considered to be representative for Dutch automation personnel in general. The occupations that were included ranged from those in the top management (directors), middle management (heads of department) and lower management (system manager, applications manager etc.) to occupations such as computer specialists (in design, analysis and/or applications), marketing/sales personnel, administrators, teachers (in the civil service as well as in commerce), consultants and support personnel (workplanners, operators etc.).

The questionnaire contained, amongst others, questions concerning the evaluation of the job on different aspects. Aspects that were included refer to salary and prospects, the challenge of the job (engaging, offering pleasure), workload (working hours, physical strain), hectic working conditions (time pressure, unexpected events), qualifications (education, experience), quality of leadership and contacts with colleagues. The questionnaire also contained questions on work-related physical and mental strain (feeling tense, nervous, agitated etc.); a number of these items concerning work-related physical and mental strain had been taken from the "NIPG-questionnaire on Work and Health", which has been developed and validated within the TNO Institute for Preventive Health Care (NIPG-TNO) (Dijkstra et al., 1986). Using information on the specific working conditions of automation personnel obtained from interviews held in the preparatory phase of this study, another 13 items were added to the 27 items from the NIPG-questio-nnaire.

A comparison was made between the sample under investigation and a reference group.

This reference group was composed of the sample of subjects who had participated in research projects using the "NIPG-questionnaire on Work and Health", a sample which may be considered representative for the Dutch working population. As the automation group under study was known to be on average better educated than the reference group, we selected from the latter group all subjects with education-level at or higher than secondary vocational education (2000 subjects).

From the administration of the questionnaire on the sample, we obtained a datamatrix of the subjects' scores on the job aspects. Because we were not so much interested in the scores of individuals, but rather in the average scores of the occupations on the job aspects, we chose to aggregate the scores over the 32 occupations. Secondly, because we wanted to explore the relationship between occupations and aspects at category level, rather than assume linearity, we recoded the average scores of the occupations. If an occupation had an average score on a job aspect that was higher (more unfavourable) than the mean plus 0.2 times the standard deviation of that item within that occupation, it received a score on that aspect of 3; if it had an average score lower (more favourable) than the mean minus 0.2 times the standard deviation, it received a score of 1; otherwise, it was given a score of 2. Thus, the higher the score of an occupation on an item, the more unfavourable it scored on that item. Because we were interested in the exploration of global relationships between risk factors and occupations rather than in the testing of exact hypotheses, the arbitrariness of this recording was not considered to constitute a significant reduction of the information content of the data.

2.3 METHOD

The 32 by 40 datamatrix of occupations by recoded scores on job aspects was analyzed using homogeneity or multiple correspondence analysis (MCA) (Greenacre, 1984). MCA is a useful technique for exploring at category level the relationship between several categorical variables. It provides a configuration of both occupations and aspects in such a way, that a cloud of aspects is centred around each occupation, aspects that 'belong' to that occupation. Similarly, around each aspect a cloud of occupations is placed that all score high on that particular aspect. Clusters of occupations and/or aspects indicate homogeneous subgroups of occupations, that exhibit a similar scoring pattern. The more specific or homogeneous groups are, the more peripherally they are placed in the configuration or structure. Conversely, centrally placed occupations and aspects play a more diffuse and heterogeneous role. By inspecting the resulting configuration - looking for clusters of occupations and job aspects - one gets an impression of the relations between occupations and the aspects.

The placing of occupations and aspects is done in such a way that occupations that score high on a certain aspect, are situated close to that aspect. This implies that for the datamatrix described above, in which a high score implies a lot of complaints, multiple correspondence analysis produces a configuration that stresses the unhealthiness or stressfulness of occupations, in other words, a 'complaints picture'. As the absence of complaints may not always be equalled to contentedness, we also performed multiple correspondence analysis on the same datamatrix after it had been recoded in the opposite way (so that high scores imply below average complaints scores, low scores imply above average complaints scores), thus producing a 'contentedness picture'.

2.4 RESULTS

2.4.1 Demographics

From the total of 5,461 persons in the sample, 60% had returned a completed questionnaire (3233 respondents). The mean age of the sample was 36 years; on average 5% were women. The 32 different occupations that were investigated may be divided into 9 subgroups. The occupations, their assignment to the 9 groups, mean age, percentage of women, and the corresponding labels used below, are shown in Table 2.1.

Table 2.1 Occupations with mean age, percentage of women and labels

occupation	mean age	percentage women	label
directors			
financial director	43	2	(dir 1)
managing director	42	3	(dir 2)
middle management			
head of computing department	40	2	(mm 1)
head of information-processing department	40	3	(mm 2)
head of programming department	38	0	(mm 3)
head of system-analysis department	41	0	(mm 4)
lower management			
system manager	34	5	(lm 1)
applications manager	34	9	(lm 2)
network manager	33	0	(lm 3)
project coordinator	36	4	(lm 4)
computer specialists			
system-analyst	32	14	(cs 1)
system-analyst/programmer	31	8	(cs 2)
system-designer	32	9	(cs 3)
application-programmer/analyst	30	9	(cs 4)
network analyst/programmer	31	0	(cs 5)
data-/telecommunications-specialist	35	0	(cs 6)
information-analyst	35	14	(cs 7)
micro-computer programmer	30	3	(cs 8)
marketing/sales personnel			
marketing personnel hard- & software	37	7	(mark 1)
sales-personnel hardware	33	8	(mark 2)
sales-personnel software	35	7	(mark 3)
administrators			
accountant-administrative advisor	42	2	(adm 1)
internal accountant	38	2	(adm 2)
external accountant	44	0	(adm 3)
head of administration/controller	39	2	(adm 4)
teachers			
business-applications course leader	35	9	(teach 1)
informatics teacher	37	12	(teach 2)
consultants			
consultant/advisor	37	5	(cons 1)
management-consultant	38	3	(cons 2)
support personnel			
work-planner	34	2	(sup 1)
operator	30	5	(sup 2)
service-engineer/customer-support	34	1	(sup 3)

Numbers per occupation were such that further analysis on this level was considered justified. Approximately 50% of respondents had completed specific computerisation education at least higher vocational level; 30% indicated that they had not received an education specific for computerisation. Almost half (48%) of the respondents indicated that they were not (yet) properly trained for the particular computerisation job they held at present.

All 40 questionnaire items, their labels, percentages affirmative answers in the sample and percentages affirmative answers in the reference group on the appropriate items, are shown in Table 2.2

Table 2.2 Questionnaire items, labels* and percentages affirmative answers in sample and reference group

	labels	percentages affirmative answers sample	percentages affirmative answers reference group
More than 5 hours a week spent on study	(study)	27%	-
Working more than 45 hours a week	(wl 1)	36%	-
Not being able to use up all one's days off	(wl 2)	30%	-
Not being able to use up days off in desired period	(wl 3)	22%	-
Work is physically strenuous	(wl 4)	8%	26%
Salary dependent upon performance	(perf)	10%	-
Workload unequally divided over time	(hect 1)	59%	-
Working under time-pressure	(hect 2)	69%	59%
Progress of work delayed in case of one's absence	(hect 3)	41%	-
Often hampered by unexpected events	(hect 4)	66%	61%
Often bothered by poor work of others	(hect 5)	47%	41%
Often hampered by the absence of others	(hect 6)	26%	27%
Not the right education for the job	(qual 1)	48%	-
Not enough experience for the job	(qual 2)	38%	-
Doubts about keeping up with innovations	(qual 3)	27%	-
Recent health complaints	(health)	10%	27%
Work is not engaging	(chal 1)	5%	31%
Work offers no pleasure	(chal 2)	3%	13%
Work is not mentally demanding	(chal 3)	12%	39%
Work is often too tiring	(str 1)	21%	26%
Regular problems with pace of work	(str 2)	13%	17%
Feeling it is better to slow down	(str 3)	28%	28%
Often feeling tense	(str 4)	23%	33%
Often feeling nervous	(str 5)	9%	21%
Often feeling agitated	(str 6)	28%	31%
Not sleeping well	(sleep)	4%	11%
Regular headaches	(head)	10%	17%
Work is badly organised	(lead 1)	38%	38%
Insufficient consultation	(lead 2)	24%	23%
Bad day-to-day management	(lead 3)	48%	34%
Management has no clear picture of one in one's work	(lead 4)	35%	35%
Management doesn't take one's opinion into account	(lead 5)	27%	34%
Not able to decide on the organisation of work	(aut 1)	8%	-
Not able to decide on the method of working	(aut 2)	12%	-
Not able to decide on the pace of the work	(aut 3)	21%	-
Relations among colleagues not pleasant	(relat)	11%	22%
Irritated by colleagues	(irrit)	18%	41%
Work influences private life in a negative way	(privat)	31%	31%
Salary doesn't concur with performance	(salary)	34%	52%
Prospects for career are dim	(prospects)	23%	45%

*cluster labels: wl=workload, hect=hectic working conditions, qual=qualification, chal=challenge, str=stress, lead=leadership/coordination, aut=autonomy

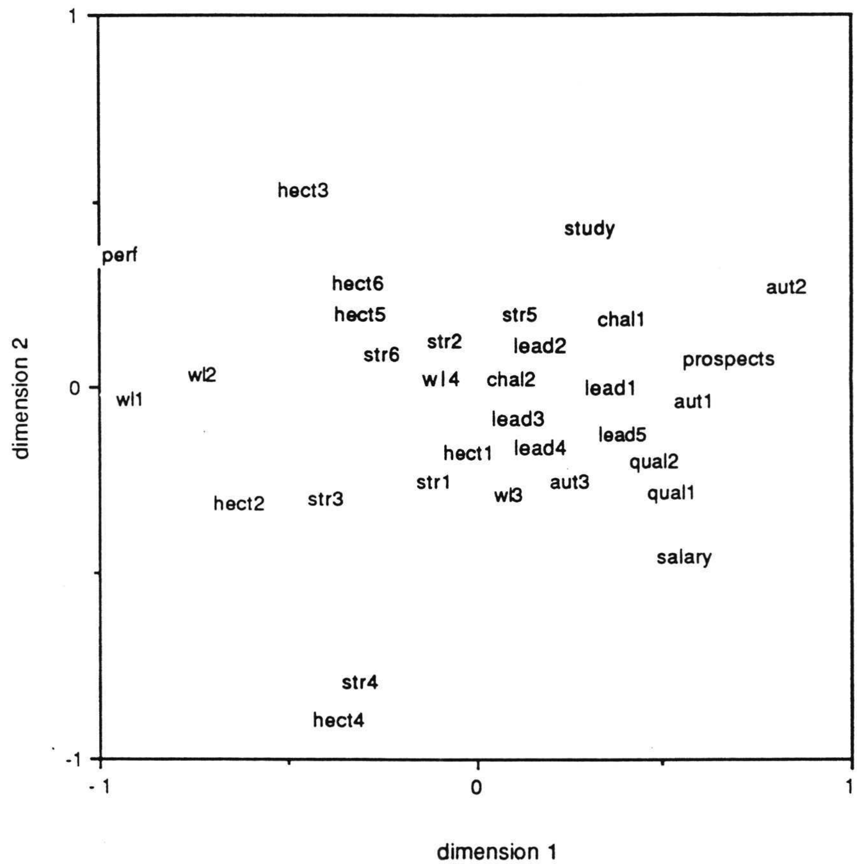
2.4.2 Comparison with reference group

The comparison of the sample under investigation with the reference group on the appropriate 27 items of the questionnaire, shows that the automation personnel under study have a more unfavourable judgement on the quality of leadership and indicate a greater amount of hectic elements in their working situation than the reference group. The judgement on the challenge offered by the job, the kind of contacts with colleagues, salary and future prospects, however, is more favourable among automation personnel than in the referencegroup. Scores on indicators of physical and mental strain are somewhat lower among automation-personnel than in the reference group, although, especially on indicators of mental strain, differences between the various occupations within the automation group under study are considerable. Comparing the two groups on such a large number of aspects, there are bound to be some differences, therefore we have chosen not to test the significance of these differences but rather present the numbers for illustrative purposes.

2.4.3 Multiple correspondence analysis of complaints

Next, we performed multiple correspondence analysis on the 32 by 40 recoded 'complaints' data matrix using the procedure CORRESP in SAS version 5. We chose a two-dimensional solution. The first and second dimension explained respectively 30.5% and 10.2% of the variance of the rescaled variables. A representation of the scores of the job aspects is in Figure 2.1a, a similar representation of the scores of the occupations is in Figure 2.1b; for the labels see Tables 2.1 and 2.2. Both figures in fact have been extracted from the solution that contains scores of job aspects and occupations; for reasons of clarity they are presented separately. In Figure 2.1a, we have only indicated those work aspects that are placed in the periphery of the picture.

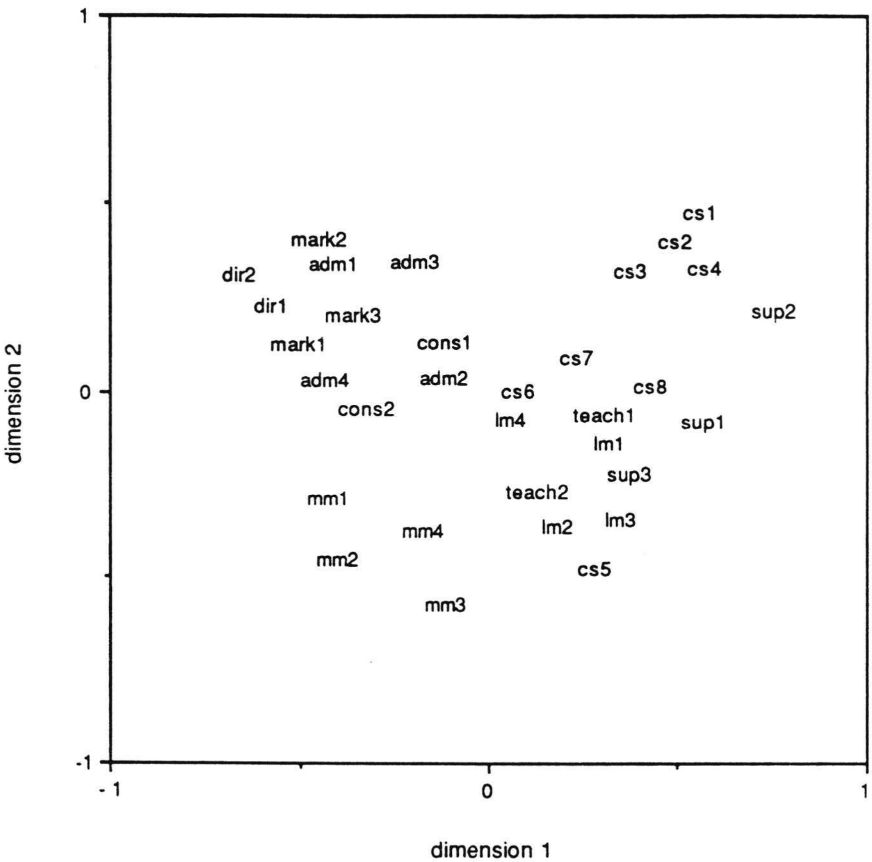
Figure 2.1a Workaspects in 'complaints' solution



Looking at the positioning of occupations in Figure 2.1a, four clusters of occupations can be observed. The first cluster is situated in the upper left side of the configuration: it is composed of directors, consultants, administrative- and marketing/sales-personnel.

Combining Figure 2.1a with Figure 2.1b, it appears that the occupations in the first cluster share a demanding and responsible character. Typical aspects of these occupations are the delay in progress of in case of absence (hect 3), working more than 45 hours a week (wl 1), and not being able to use up all days off (wl 2). It is often mentioned that salaries are dependent upon one's performance (perf). Poor work by others (hect 5) and being hampered by the absence of colleagues (hect 6) are symptoms of hectic working conditions. A feeling of agitation (str 6) is also mentioned.

Figure 2.1b Occupations in 'complaints' solution



The second cluster of occupations is found in the lower left side of the configuration. This cluster contains all four middle management occupations. These occupations are typified by impediments due to unexpected events (hect 4), tense feelings (str 4), frequently working under time pressure (hect 2), work that is experienced as being too tiring (str 1) and the notion that one had better slow down (str 3). They share high scores on aspects of 'workload' (wl 1 and wl 2) with the first cluster of occupations.

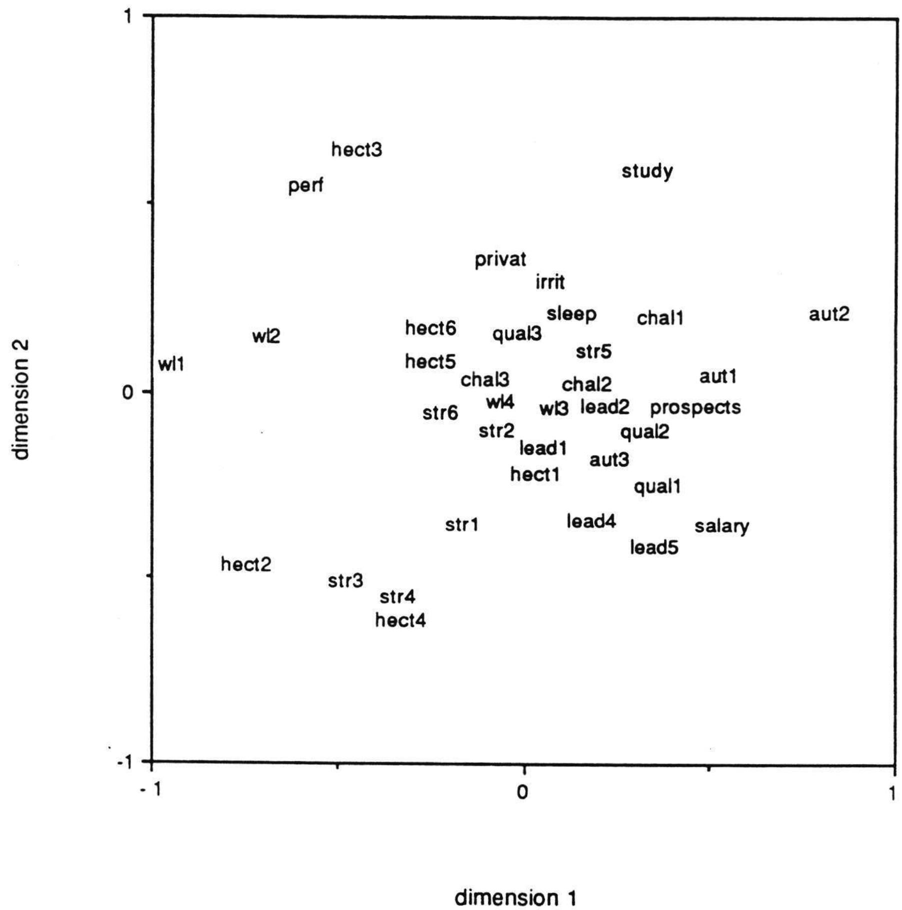
A cluster of computer specialists (cs 1 to cs 4) is placed in the upper right side of the picture; close to this cluster is also situated the operator (sup 2). In this cluster, occupations are typified by an unfavourable judgement on the degree of autonomy (aut 2), the manner in which work is organised (lead 1), the degree of consultation (lead 2) and the degree to which work is engaging (chal 1). Characteristic is the large number of hours spent on study.

The fourth cluster is situated in the lower right side of the picture. It consists of lower management personnel, teachers and support personnel (without the operator). This cluster is less marked and more heterogeneous than the previous ones. Occupations in this cluster are characterised by an unfavourable judgement on salary, a lack of education and experience (qual 1 and qual 2), an unfavourable judgement on day to day management (lead 3), and the degree of autonomy in the management and pace of work (aut 1 and aut 3). Prospects for a further career are regarded as dim.

2.4.4 Multiple correspondence analysis of contentedness

The same analysis was performed on the recoded 'contentedness' data matrix.
Again, a two-dimensional solution had been chosen; the first and second dimension explained respectively 30.0% and 11.3% of the variance of the rescaled variables. Representations of the scores of the job aspects and the occupations are in Figures 2.2a and 2.2b; for the labels see Tables 2.1 and 2.2. As in Figures 2.1a and 2.1b, only those work aspects and occupations have been indicated that were placed in the periphery of the picture.

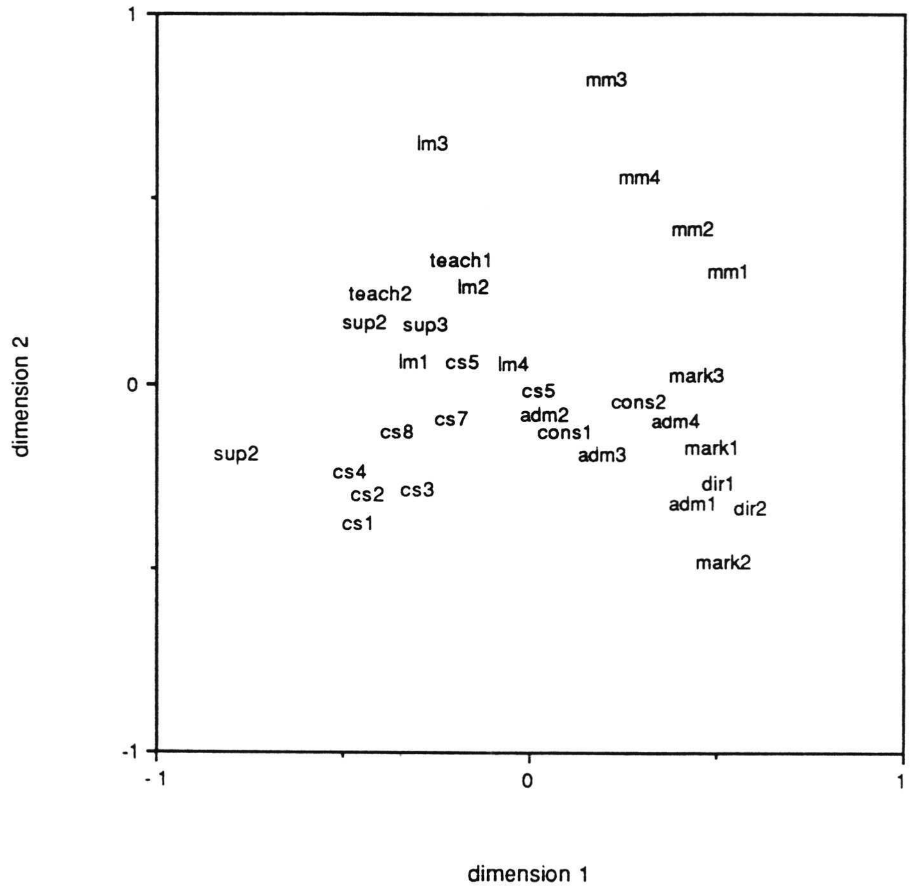
Figure 2.2a Work aspects in 'contentedness' solution



The structures of aspects and occupations in the two pictures closely match those obtained in the complaints analysis. The aspects picture has remained virtually identical; the occupations picture has been mirrored on both dimensions, as should be expected. So, in general, there is for this sample a consistent pattern of complaints and contentedness. There is a number changes, however.

The network manager (lm 3) and network analyst (cs 5) have changed their position to ones that are more compatible with their pattern of scores. The network analyst moved from its odd-one-out position in the group of teachers, lower management and support

Figure 2.2b Occupation in 'complaints' solution



personnel, to a position closer to the other computer specialist occupations; its new position is more central which matches better the fairly average scores and generally low level of complaints in this group. The network manager moved to a more peripheral position which is more consistent with the high level of complaints on salary, leadership and the qualifications in this group. Software sales personnel has moved to a position closer to the cluster of middle management personnel, which matches better their scores on nervousness; their position is now also closer to prospects, and to 'not being able to decide on the organisation of the work' (aut 1). Apart from these changed positions of software sales personnel, the network manager and network analyst, the group of computer specialists has become a more compact cluster, as has the cluster of directors, consultants and marketing personnel.The aspect 'management doesn't reckon with one's opinion' (lead 5) has moved to a more prominent place: it is now closer to the top management, and hardware sales personnel. The aspect 'salary dependent upon performance' (perf) has moved away from the periphery, closer to 'progress of work delayed in case of absence' (hect 3) and to the position of the network manager. This is also more in line with the scores of the occupations on the aspects.

Of special interest is the position of the cluster of computer specialists, together with the operator (sup 2). In the complaints analysis, these occupations had been marked by unfavourable scores on autonomy, challenge, study and prospects. Combined with their position in the contentedness analysis close to stressful and hectic aspects, it appears that for this group the risk of 'under-burdening' seems to be present (Cox, 1985).

2.5 DISCUSSION

We have investigated the general relations between job aspects and occupations in a sample of automation personnel. For this, we used multiple correspondence analyses of the automation personnel's complaints and contentedness scores. The two analyses sketched in broad lines the same picture of risk factors and occupations at risk, while, compared to the structure in the complaints analysis, the structure of the occupations in the contentedness analysis had become more compact, and a few aspects and occupations had been positioned in places that were more consistent with prior information. Apart from providing stability information, the two analyses also produced complementary results.

Combining the results from the complaints and contentedness analyses, it appears that the group of computer specialist, together with the operator, can be identified as a special risk group that shows a pattern combining little autonomy, prospects, challenge and few demands.

The stressfulness of the position of middle management personnel seems to be of specific interest. The responsibility carried for the work of one's own and that of others, combined with a great workload and hectic working conditions, seems to harbour the risk of overburdening. Added to this, middle management personnel usually has to operate between 'hammer' and 'anvil' (Mintzberg, 1973), in this case the managing board and subordinate computer specialists. The difficulties met in planning computerisation projects add further to the stressfullness of this position, especially when taking into account that until recently the turnover of personnel within project-teams was considerable. Data obtained in the same research (Andries, 1990) show that most middle management personnel stem from the ranks of computer specialists (system-analysts, programmers etc.). Earlier research (Couger & Zawacki, 1980) into work preferences of computer specialists demonstrated a marked preference for career prospects and relatively little interest in social aspects of a job, reinforced by a process of self-selection during education. One could tentatively argue that, once arrived at the managerial level, lack of social skills may consequently cause problems in one's own functioning and the functioning of others. Perhaps the relatively unfavourable judgement on day-to-day-management in the overall population should also be viewed in this context.

In other similarly hectic and demanding occupations, for instance those of director and consultant, the risk of overburdening as reflected in scores on indicators of work related mental strain are, surprisingly, not above the sample average. This may be related to the fact that in these occupations there is only indirect responsibility or the possibility to delegate responsibility for the outcome of projects.

The manner in which programmers, system-analysts and system-designers evaluate their job, is almost exactly opposite to the manner in which those employed in the middle management evaluate their jobs. The workload, hectic elements in the working situations and elements of work-related stress seem to play a much smaller role in their evaluation of the job. On the other hand complaints on the challenge offered by the job and the degree of autonomy, especially concerning the method of working, are typical for this group. These computer specialists spend a lot of time on study. They might be typified as 'starters'; the average age is just over 30, and for most this is their first job. In broad outline this description applies also to the position and opinions of operators, especially regarding the lack of autonomy in their method of working.

Lower management, teachers and support personnel (the operators excluded) share a number of common characteristics that sketch a relatively pessimistic outlook on their career-prospects. This is not only substantiated by their own unfavourable judgement on their degree of qualification and future prospects, but also by other data regarding their former career and actual level and type of education. This applies especially to lower management regarding the degree of education and to teachers regarding the type of education needed for the job. These occupations might tentatively be labelled as 'blind-alley' occupations.

Summarizing, the sample shows a great diversity in the evaluation of jobs and in scores on mental strain. There are strong signals that in some occupations working conditions such as excessive workload and constant time pressure may, with increasing age, lead to increased mental strain and in the end possibly the inability to cope with the workplace. Also, it appears that in other occupations in the automation sector, employees will probably in the near future encounter problems to find permanent employment, due to inappropriate education and the recent phenomenon of a declining growth of the automation sector in the Netherlands (CBS, 1989). The way in which employees will be affected is also dependent on general trends within the automation sector (Regtering & Riesewijk, 1987). Until recently, not only the sector itself knew rapid growth in volume, but this applied also to the number of new or newly titled occupational and corresponding educational domains. The uncertain manner in which these growing and changing occupational domains can be moulded within old and newly formed organisations, is exacerbated by the shift in the demarcation line between the responsibilities of automation specialists and users, caused by the decentralisation of facilities. These developments are only one of the manifold justifications for scientific interest into the rapidly changing world of automation, not in the least for the sector's impact on future working- and living-conditions.

For these reasons, research into the development of a system for monitoring the quality of working conditions and their effect on indicators of physical and mental well-being among automation personnel is recommended. At the same time, research proposals should focus on the development of a better understanding of causal mechanisms and confounders in the possible relation between for instance workload, autonomy, leadership, strain and stress.

REFERENCES

ANDRIES, F. Automatiseren is mensenwerk [Automation is the work of humans]. Leiden, NIPG-TNO, 1990.

CENTRAAL BUREAU VOOR DE STATISTIEK (CBS). Automatiseringsstatistieken. Particuliere sector [Automation statistics. Private sector]. 1989. 's-Gravenhage, CBS, 1989.

COUGER, J.D. & R.A. ZAWACKI, Motivating and managing computerpersonnel. Chichester, Wiley, 1980.

COX, T. The nature and measurement of stress. Ergonomics 28 (1985) 1155-1163.

DIJKSTRA, A., GRINTEN, VAN DER, M.P., SCHLATMANN, M.J.Th. & C.R. DE WINTER. Functioneren in de arbeidssituatie; uitgangspunten, ontwerp en handleiding voor onderzoek onder werknemers naar gezondheid, werk en werkomstandigheden. [Functioning at the workplace; starting points, design and guidelines for research amongst employees into health, work and working conditions]. Leiden, NIPG-TNO, 1986.

EVANS, J. Women, men, VDU-work and health: a questionnaire survey of British VDU-operators. Work and Stress 1 (1987) 271-283.

GREENACRE, M.J. Theory and applications of correspondence analysis. London, Academic Press, 1984.

MINTZBERG, H. The nature of managerial work. New York, Harper & Row, 1973.

OBORNE, D.J. Computers at work. Chichester, Wiley, 1985.

POT, F.D., PADMOS, P. & A.A.F. BROUWERS. Achter de schermen [Behind the screens]. 's-Gravenhage, Ministerie van Sociale Zaken en Werkgelegenheid, 1986.

REGTERING, H. & B. RIESEWIJK. ICON onderzoek Nederland [ICON research in the Netherlands] Nijmegen, International Computer Occupations Network, 1987.

RIESEWIJK, B. & J. WARMERDAM. Het slagen en falen van automatiseringsprojecten [The success and failure of automation projects]. Nijmegen, Katholieke Universiteit, Instituut voor Toegepaste Sociologie, 1988.

SAS Technical Report P-179. Cary, NC, SAS Institute Inc.

WERKGROEP Sociaal Beleid van de stuurgroep actieplan Computer Service Industrie. Sociaal beleid computer service industrie [Committee on Social Policy of the steering committee Computer Service Industry. Social policy computer service industry].'s-Hertogenbosch, NEHEM, 1988.

NONRESPONSE IMPUTATION IN PUBLIC HEALTH RESEARCH

S. van Buuren*, J.L.A. van Rijckevorsel**

* Department of Psychometrics, Faculty of Social Sciences, University of Utrecht (presently at the Department of Statistics and Computer Science, TNO Institute of Preventive Health Care)

** Department of Statistics and Computer Science, TNO Institute of Preventive Health Care

Abstract

This paper suggests a method to supplant missing categorical data by 'reasonable' replacements. These replacements will maximize the consistency of the completed data. Consistency is measured by a between-total variance ratio. The idea is that similar profiles obtain comparable imputations. The text outlines the optimization problem, describes relationships to the relevant psychometric theory and studies some properties of the method. Some examples are presented. The main application fields are in the analysis of survey data, rating scales and questionnaires.

Keywords: *missing data, imputation, categorical variables, consistency, least squares, MISTRESS, multiple imputation*

3.1 INTRODUCTION

Missing data are common and costly. Public health data with up to 30% missing are no exception, and attempts to do something about missing data are as old as missing data themselves. Popular ways to accommodate for missing data - pairwise - and listwise - deletion - may amount to wasting labour-intensively collected material. An alternative is to fill in missing entries with 'appropriate' replacements. The advantage of this is that standard multivariate techniques can be applied to the completed data.
Sometimes, external information is available that may help. For example, suppose that a subject is unwilling to tell his age, thereby producing a missing value. If we have the actual person in front of us we may nevertheless infer his age by using other clues, and subsequently fill in our prediction as if it had been observed. Another source of external information could be a previous score on same question or test. Unfortunately, such situations are more an exception than a rule.
In general, if we want to complete the data we should look for other sources of information. An obvious alternative is to consider the data that *are* available for the subject, in combination with the data that are collected on other sample units. Under the assumption that observations with a similar response pattern are likely to score identically on any remaining, unobserved variables, we may try to interpolate missing values. This type of imputation strategy is known as *hot deck* imputation. The basic 'reasonable imputation assumption' is: objects with almost similar profiles have the same distribution on any missing responses. Thus, the idea is to borrow the observed score from a closely related profile. Some arguments that are often mentioned in favour of hot deck methods are: the reduction of the response bias, the preservation of the distribution of the population, and - most important of all - the production of a complete data set. Particularly in large surveys, computational ease and speed are highly evaluated.
In this paper, we consider a hot deck imputation technique based on the within-homogeneity of all variables simultaneously. The method consists of two ingredients: a *donor variable* and an *imputation rule*. The donor variable measures how much individual data profiles differ from each other. So, a donor is not a sample unit or a data profile, but a latent variable. The imputation rule states how blank entries should be filled, given the values on the donor variable. As we will see, these two components are closely

intertwined in the method. Changing the donor variable may cause a modification of the imputations. The converse is also true; changing an imputation has an effect on the donor variable.

The donor variable is equal to a weighted average of all variables. We use the familiar between - total sum of squares - ratio to indicate how well the donor represents the total variation in the data. It follows that the 'best' donor variable is equal to the first principal component of the completed data. The position, or score, of each observation on the donor reflects how much sample units have in common. The function of the donor is thus much like the partitioning of observations into homogeneous classes employed by traditional hot deck procedures. The difference with existing hot deck procedures is that all (categorical) variables act simultaneously as donor.

Imputation is based on comparing donor scores. If an incomplete profile resides closely to a completely observed profile (in the sense that their donor scores differ little), then its missing entries can replaced by the known values of the observed unit. We measure 'closeness' by the squared Euclidean distance between donor scores. Mathematically, we look for imputations that minimize a sum-of-squared-distances function.

In practice, just one donor may not be representative. Primary reasons for using just one is that it is simple, and that it has some attractive analytical properties. If necessary, the extension to multiple, orthogonal donors is possible. We will indicate where this is appropriate.

The donor is the most homogeneous replacement for all variables simultaneously, and hence it is most homogeneous with regard to the complete and incomplete data. This satisfies the, according to Ford (1983), most important principle in the construction of any hot deck procedure: the 'imputation model' and the 'data model' must be the same. Both imputations and quantifications maximize the homogeneity of the completed data set. As such they are relevant to the observed as well as to the missing observations.

The present imputation method is similar to missing data estimation by the EM algorithm (Little and Rubin, 1987) in that both methods optimize an objective function over the imputations. Moreover, both methods consist of the two main steps: an Expectation (E) step that completes an incomplete data matrix, and a Maximization (M) step that estimates the model parameters. However, there are also substantial differences. We use Least Squares instead of Maximum Likelihood, we do not make any distributional assumptions, and we provide discrete instead of fractional imputations , which are spread out over more than one cell of the contingency table.

We like to emphasize at this point that maximizing consistency is by no means the only valid or useful criterion to find missing information. Suppose that we are interested in demonstrating that two variables are independent of each other. In that case, it will be clear that maximizing consistency is a bad idea since it moves us further away from the independence model. A more natural alternative here would be to do the opposite, that is, to minimize homogeneity. There exists yet no unbiased technique nor a general purpose strategy for dealing with missing data. Multivariate optimality for imputation of missing data is about impossible to define without violating some statistical model or another. Nonetheless, *if* we believe that the observed data tell us something about the missing data - and this is a fundamental assumption of all hot deck methods - then maximally consistent replacements will be attractive in general.

Good reviews imputation techniques for categorical data are Kalton and Kasprzyk (1982) and the three volumes edited by Madow, Olkin and Rubin (1983). The annual proceedings of the Section of the Survey Research Methods of the American Statistical Association offer a continuing story on the handling of missing data in survey research. The primary source for Maximum Likelihood models for missing categorical data is Little and Rubin (1987). For applications of multiple imputation in public health see Rubin and Schenker (1991). Missing data in experimental designs are discussed in Dodge (1985). For multiple imputation, in which not just one but many replacements are searched, see Rubin (1987). Hedges and Olkin (1983) give a selected and annotated bibliography on incomplete data. Ford (1983) summarizes many hot deck strategies. Little and Rubin (1990) provide a recent overview of missing data strategies in the social sciences.

The structure of this paper is as follows: first, we discuss a small imputation example. After this, we define the consistency measure, and we introduce the loss function. Subsequently, we relate the consistency criterion to other psychometric theory, and indicate a number of similar approaches. Practical use of the method is illustrated by some examples, one of them concerning multiple imputation. Finally, we summarize the main results and we discuss some practical implications and future work.

3.2 EXAMPLE

To be able to grasp the nature of consistent imputations, we discuss the small artificial data listed in Table 3.1. This table contains 10 observations on three categorical variables. There are three missing values, indicated by *a*, *b*, and *c*.

Table 3.1 Example Data

Person	Income	Age	Car
1	*a*	young	jpn
2	middle	middle	am
3	*b*	old	am
4	low	young	jpn
5	middle	young	am
6	high	old	am
7	low	young	jpn
8	high	middle	am
9	high	*c*	am
10	low	young	am

The problem is to find replacement values that are reasonable in some way. For *a* this is easy; the most consistent estimate is *low*, because this makes the profiles 1, 4, and 7 identical. A young owner of a Japanese car will have a low income simply because this is a recurring pattern. Moreover, the profile contains all Japanese cars in the data. Analogously, we find *high* for *b* and *old* for *c*. Both imputations make the remaining two incomplete profiles identical to row 6. So, the missing scores are interpolated from other profiles. We simply look for similar rows. This is the same as saying that variables must be as homogeneous as possible, i.e., measure the same thing. So here we end up with two homogeneous groups with three members each.

Since there are 3 missing values, each with 3 categories to choose from, the total number of different solutions is $3 \times 3 \times 3 = 27$. Table 3.2 lists the amount of consistency for each of these solutions. The exact definition of consistency, expressed as fit, can be found in the next section. Because the example is deliberately easy and somewhat trivial, the most consistent solution '*l h o*' can be derived by eye-balling alone. In more realistic situations, eye-balling is usually not enough. First, because the best solution may contain new, previously unobserved, profiles. It will be difficult to find such combinations. Second, because optimal consistency becomes hard to detect for more than three or four missing values.

For categorical data, Wilks procedure - filling in the average- boils down to selecting the modal category. The corresponding solutions in the example are '*l l y*' and '*h h y*'. These imputations have consistencies of 0.70104 and 0.68827 respectively, which illustrates the well known fact that Wilks method tends to discard between-groups variance.

Table 3.2 Consistency of all possible imputations for Table 1

a b c	Fit	*a b c*	Fit	*a b c*	*Fit*
l l y	.70104	m l y	.63594	h l y	.61671
l l m	.77590	m l m	.72943	h l m	.66458
l l o	.76956	m l o	.72636	h l o	.65907
l m y	.78043	m m y	.70106	h m y	.70106
l m m	.84394	m m m	.77839	h m m	.74342
l m o	.84394	m m o	.77839	h m o	.74342
l h y	.78321	m h y	.73319	h h y	.68827
l h m	.84907	m h m	.80643	h h m	.74193
l h o	.84964*	m h o	.80949	h h o	.74198

The obvious difficulty with categorical data is that distances between profiles cannot be easily derived; it makes little sense to substract Japanese from American cars. We deal with categorical variables by first transforming them into numerical data, by quantifying each category separately. Subsequently, the resulting numerical variables combine into the donor. Table 3.3 lists those donor scores for each observation and the scale values of the categories, both before and after imputation.

Table 3.3 Donor Scores and Scale Values for the Optimal Imputation

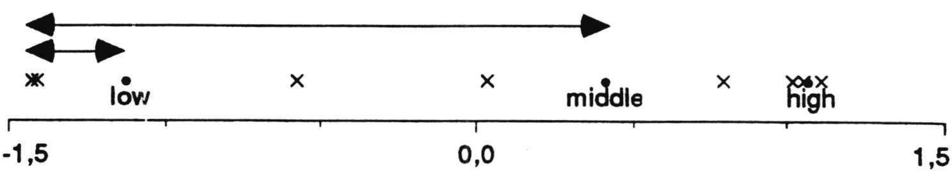
		Donor Scores		Variable		Scale Values	
		Initial	Final			Initial	Final
1	***l*** y j	-1.43	-1.33	Inc	low	-1.13	-1.15
2	m m a	0.79	0.66		middle	0.41	0.33
3	***h*** o a	1.02	1.00		high	1.06	0.98
4	l y j	-1.41	-1.33				
5	m y a	0.04	-0.01	Age	young	-0.96	-0.92
6	h o a	1.11	1.00		middle	0.92	0.79
7	l y j	-1.41	-1.33		old	1.07	1.00
8	h m a	1.05	0.92				
9	h ***o*** a	1.02	1.00	Car	jpn	-1.41	-1.33
10	l y a	-0.58	-0.59		am	0.63	0.57

The initial solution ignores missing data, an option known as 'missing passive'

(Meulman, 1982; Gifi, 1990, p. 136). In this case, donor values for subjects 1, 3, and 9 are based on two, instead of three, observed categories. As will be shown below, a scale value for a category is equal to the average of all donor scores that fall into that category. For example, the initial value of *low* is equal to –1.41 –1.41 – 0.58 / 3 = –1.13, i.e., the average donor score of observations 4, 7, and 10.

Let us now try to impute the incomplete entry in profile 1. The initial donor score of the profile is –1.43. To complete the data, we may pick any of the three income low, middle or high categories: The scale values of these categories are –1.13, 0.41, and 1.06. The joint scale of the initial donor scores and the category points is plotted in Figure 3.1.

Figure 3.1 Joint plot of objects (x) and the categories of income (•)



The most consistent imputation is that category whose scale value is closest to the donor. Here, –1.13 is closest to –1.43, so we choose *low*. Apparently, subject 1 has most in common with profiles 4, 7, and 10. So, when compared to the other two income classes, the low income group is most similar to profile 1. Consequently, we borrow the replacement value from this group. Since middle and high income groups are more distinct, imputing *middle* and *high* will increase the within-groups variance more than necessary, and so, these values should not be used as stand-ins.

We execute the same steps for the missing data in rows 3 and 9. After all missing entries have received an initial imputation, new donor scores are computed, but now using the completed data. The entire process is repeated until the consistency of the solution does not change anymore. The values of the final solution are also given in Table 3.3.

3.3 METHOD

Let the data be coded into m binary indicator vectors g_j of length k_j for $j=1, \ldots, m$, and let random vector y_j contain k_j category quantifications for the j-th variable, where k_j is the number of categories. The quantified variables $x_j = g_j'y_j$ have zero means and the random variable z contains their average, i.e., $z = 1/m \sum x_j$. The total variation of the data can be decomposed as

$$\Sigma_j (g_j'y_j)^2 = m\, z^2 + \Sigma_j (z - g_j'y_j)^2.$$

This is a between-within partitioning of the form $T = B + W$. The *correlation ratio*, denoted by η, and defined by $\eta^2 = B \,/\, T$, measures how well the average can be considered as a representative of each x_j. The ratio ranges from 0 to 1. The coefficient equals 1 if all variables are proportional.

The donor variable z tells us something about the similarity among profiles that belong to distinct replications. Let z_i for $i = 1, \ldots, n$ denote the score of the i-th profile on the donor z. The difference between z_i and $z_{i'}$ is equal to some distance norm between profile i and profile i'. The donor variable z defines a metric in which observational units can be represented, which we use to compare different data profiles. The correlation ratio η can also be interpreted as a measure of how well the entity $z_i - z_{i'}$ reflects the multivariate differences between rows i and i' over all x_j. Obviously, the larger η becomes, the better the difference $z_i - z_{i'}$ portrays the similarity between i and i'. We might say that z is a satisfactory donor variable in this case.

Procedures for finding optimal η over $y_1, \ldots, y_m$ are known as homogeneity analysis, multiple correspondence analysis, dual scaling and others (see Gifi, 1990). These techniques usually consider several orthogonal sets of z's, with corresponding η's. We define the donor variable z as the set of numbers that maximizes η.

We now discuss missing data. Let Ω denote the set of all *observed* variables and let the $x_j^* = g_j^{*\prime}y_j$ stand for an imputed value. Obviously,

$$x_j = \begin{cases} x_j, & if\ j \in \Omega \\ x_j^*, & if\ j \notin \Omega \end{cases}$$

It is possible to partition the variation into three independent quadratic components:

$$\Sigma_j (g_j'y_j)^2 = m\, z^2 + \Sigma_{j\in\Omega} (z - g_j'y_j)^2 + \Sigma_{j\notin\Omega} (z - g_j^{*}{'}y_j)^2.$$

Since $T = B + W$ the maximum of η^2 coincides with the minimum of $W / T = 1 - \eta^2$. Maximal homogeneity among the imputed variables can be found by minimizing this W / T-ratio over $z, y_1, \ldots, y_m$ *and* over the imputations $g_1^*, \ldots, g_m^*$. The corresponding loss function can be written as

$$\sigma(z; y_1, \ldots, y_m; g_1^*, \ldots, g_m^*) = \Sigma_{j\in\Omega} (z - g_j'y_j)^2 + \Sigma_{j\notin\Omega} (z - g_j^{*}{'}y_j)^2.$$

Let $\sigma(\cdot)$ stand for this loss function. The imputation problem is where to impute the '1' in the missing vector g_j^*. This is a combinatorial optimization problem.
Since larger η lead to more consistent imputations, it seems logical to look for imputations that will maximize η. We thus strike two flies at one blow: imputations will not only amplify the structure of multivariate row differences as summarized by z, but, at the same time they cause z to be a more adequate composite of those differences. This principle induces imputations such that similar looking units become even more alike, while plainly different units grow even more distinct under imputation. Dependencies in the data are thus extrapolated to the missing entries.
The solution can be computed by an iterative algorithm based on a combination of homogeneity analysis and the k-means algorithm. It works by comparing donor scores and category quantifications. The exact procedure can be found in Van Buuren and Van Rijckevorsel (1991, 1992a). See Van Buuren and Heiser (1989) for a related optimization problem.

3.4 MAXIMIZING CONSISTENCY BY IMPUTATION

The search for scores that maximize consistency is deeply rooted in psychometrics, and the following results are mainly due to this development.
It is known that η^2 is proportional to the largest eigenvalue of the correlation matrix R

of (quantified) variables. We also know that η^2 is equal to the averaged squared correlations between the quantified variables and donor variable. The average correlation among variables is another well-known measure for internal consistency. This measure is also used for categorical data, if computed from the optimal scores instead of the raw data. The average correlation is proportional with Cronbach's α (Cronbach, 1951). This is a very popular statistic in item analysis and questionnaire research. Cronbach (1951) showed that α is very similar to the average correlation. Lord (1958) showed that η^2 can be written as a function of α, so maximizing η over the missing data also maximizes Cronbach's α, the average correlation, the largest eigenvalue of the correlation matrix, and related measures.

We finish this section with the following. The idea to maximize consistency by imputation is not entirely new. Gleason and Staelin (1975) replace correlations between numerical variables by estimates that maximize the consistency of the completed data. This method is a modification of the imputation techniques proposed earlier by Dear (1959) and Buck (1960). Gleason and Staelin treat categorical data by an ad hoc rounding procedure (p. 244). Unlike the numerical case, they do not present any simulation results for their discrete imputation method. In an analysis of variance context, Hartley and Hocking (1971) identify the so-called (X, m, d) model in which one tries to find estimates for missing classifications on the experimental variables. This is a combined estimation and classification problem. They note some difficulties with the model, but they do not pursue the matter any further. Nishisato (1980) wants to impute and quantify categorical data, just like in this paper, but does not present a practical solution to the problem of selecting the optimal category to be imputed.

A difficulty with imputing categorical data in general is that one has a limited set of donor categories to choose from and no distance measure between them. One can quantify categories and use the Euclidean distances - as we do - or try to find margins that optimize consistency. Greenacre (1984, p. 237) does the latter by imputing 'consistency optimizing' rounded estimates of marginal frequencies.

3.5 MULTIPLE IMPUTATION

A drawback of any imputation method that imputes a single value is that the precision of the imputations is unknown, i.e., the variance is not estimated. In MISTRESS, one could say that the imputation variance is equal to zero, since there is only one imputation that maximizes consistency. This shows much confidence in the appropriateness of consistency as a criterion, and in the reliability of the data. According to Rubin: "It is of no use looking for the `best' or `most appropriate' imputation. Such a thing simply doesn't exist." (Rubin, 1987). What is best for one model doesn't work for another. So one has to make a distinction between the optimal value in terms of the one closest to the real, but unobserved, value and an imputation that is best in some model sense. Such values coincide if we succeed in finding that only model that generated the data; a desirable but rarely attained state of affairs, as every data analyst knows. Only in simulation studies, where indeed reality is artificially simulated and thus grossly simplified, one can hope for and achieve the coincidence of such imputations.

A different approach is to estimate the variance of imputation by generating not one, but several, say 3 to 5, completed matrices. Imputations are to be drawn from a posterior predictive distribution, or from decent approximations thereof. The spread of the imputations then conveys roughly how imputations vary. Rubin (1987) shows for a large class of statistical models that, after a model is separately fitted on each completed data matrix, simple pooling procedures can be used to obtain unbiased estimates of model parameters and the associated variances. The individual imputations do not have to be very precise, as long as together they estimate the variance. Rubin and Schenker (1991) discuss various applications of multiple imputation in health-care databases. Because multiple imputation involves a lot of work, it is worth the effort if it concerns a large body of data that is to be used by several researchers applying different models and different subsets of the data on various occasions. See also Schnell (1986, p. 227). In psychometrics, the combination of multiple imputation and sampling in various combinations is discussed by Rubin (1991).

For categorical data multiple imputations are to be drawn from a predictive distribution of categories. One can define such a distribution in several ways. The dominant distinction lies between *implicit* and *explicit* models. If we use a specified distribution to this purpose like the normal, we use an explicit model. Often there exists no proper argument to select an explicit model, and thus an implicit model, or implicit distribution is used. The most implicit model is the traditional hot deck method, where the value of the preceding observation is imputed. Multivariate simultaneous consistency is a less implicit model.

Because there is only one optimal imputation per missing value, it is impossible to generate multiple imputations by just maximizing consistency. MISTRESS yields a crisp 'all-or-none' predictive distribution for each incomplete response pattern, which is not very useful in the context of multiple imputation. For multiple imputation, we must have some way to even out the predictive category distribution so that all categories are candidates for imputation, though with varying probabilities. It would require another paper to discuss MISTRESS as a way to create posterior predictive distributions of missing data. Here we only mention some possibilities of doing so as a way to apply the method.

Let p_{ijk} denote the probability that category k of variable j is the imputation for object i, then we can specify the following:

a) a density distribution based on inverse distances

We assume that p_{ijk} is inversely related to the squared distance between the scale value y_{jk} and the donor score x_i.

b) a density distribution based on multiple donors

We mentioned that multiple orthogonal z's with corresponding y's can be considered as well. We can use each column of Z to generate a successive, separate imputation instead of using the first column of Z with the largest consistency only. The columns of Z are ordered by their respective contribution to the overall consistency, denoted by η^2. The inverse of the relative contribution defines the probability that an imputation is sampled from the imputations corresponding to the s-th column of Z. These probabilities are independent of i and j. The range of potential categories to be chosen is then restricted by their occurrence in one of the k_j imputation values. If the same category value occurs more than once as an imputation the probabilities for different s add up. One could sample as many times as one likes, but just as many draws as the number of columns of Z seems reasonable.

c) a density distribution based on conditional frequencies

One of the oldest methods assumes that p_{ijk} is proportional to the observed frequency of category k of variable j. This is a very simple way to define a predictive distribution, but it uses only univariate information. If we crosstabulate the data, each cell corresponds to a possible response pattern, and one may use the conditional frequencies instead. Note that this way of deriving the distribution will only be effective if the cells in the multidimensional crosstabulation contain a sufficient number of observations. In practice, this implies that the number of variables is limited.

These alternatives yield different predictive distributions. We do not know how this effects the results. We expect that differences will be relatively small, but more research is needed to confirm this idea. In the next section, we apply the option of inverse donor distances, with in this case quite satisfactory results.

3.6 DUTCH LIFE STYLE SURVEY

This example is taken from the Dutch Life Style Survey (Leef Situatie Onderzoek) conducted by the Netherlands Bureau of Census. The data were collected at different time points during the years 1977-1986. The data are compiled and made available to us by Anneke Bloemhoff of NIPG-TNO. As is often the case in large surveys, not all questions were posed at each occasion. Consequently, when taken together, the data contains many systematic missing entries. This example illustrates how MISTRESS can be used to find imputations for those unknown values.

The analysis sample consists of 7332 individuals. For each person, we have scores on five labour conditions. These are labelled *dirty* (**D**), *heavy* (**H**), *risky* (**R**), *stench* (**S**) and *noise* (**N**). Each subject responded whether the attribute was applicable to his, or her, job. For a subgroup of 5750 people we also know the type of job, classified into 7 categories: *management* (**MAN**), *administrative* (**ADM**), *commercial* (**COM**), *scientific* (**SCI**), *service* (**SER**) *agrarian* (**AGR**) and *industrial* (**IND**). The classification by profession is missing for 7332 – 5750 = 1582 observations. The results for single imputation, ordered by donor scores, are presented in Table 3.4.

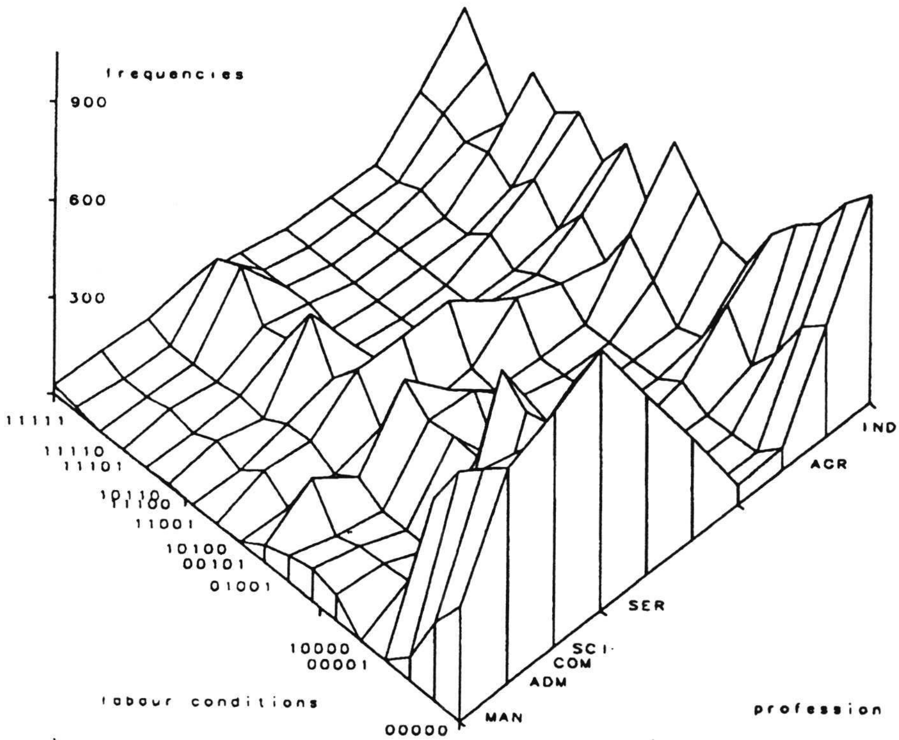
Table 3.4 Single Imputation LSO Table (**FAT** = Imputation)

Labour Conditions DHRSM	Professional Category MAN		ADM	COM	SCI	SER		AGR		IND		Donor Score
1 1 1 1 1	1		1	1	6	2		5		64	**19**	3.18
1 1 1 1 0	0		0	0	6	3		7		11	**10**	2.68
1 0 1 1 1	1		1	0	3	3		1		21	**6**	2.58
0 1 1 1 1	0		0	0	1	1		1		3		2.48
1 1 1 0 1	0		1	1	1	1		3		61	**23**	2.47
1 1 0 1 1	0		1	1	4	6		4		50	**15**	2.41
1 0 1 1 0	0		1	2	1	0		2		8	**5**	2.08
0 1 1 1 0	0		1	0	0	0		1		2	**1**	1.97
1 1 1 0 0	1		1	0	9	2		9		51	**22**	1.96
1 1 0 1 0	0		1	1	9	2		20		13	**20**	1.90
0 0 1 1 1	0		3	2	2	0		0		12	**4**	1.88
1 0 1 0 1	0		0	2	1	1		1		20	**12**	1.87
1 0 0 1 1	4		3	2	6	3		2		46	**32**	1.81
0 1 1 0 1	0		0	1	2	1		0		8	**7**	1.76
1 1 0 0 1	2		6	4	6	9		12		88	**32**	1.70
0 1 0 1 1	0		0	0	1	0		0		5	**2**	1.70
0 0 1 1 0	0		0	0	1	1		0		2	**3**	1.38
1 0 1 0 0	0		1	0	1	4		3		14	**10**	1.37
1 0 0 1 0	0		2	1	2	2		13		17	**10**	1.31
0 1 1 0 0	0		0	3	6	0		1		10	**9**	1.26
0 1 0 1 0	0		1	0	3	4		0		4	**3**	1.20
1 1 0 0 0	2		6	16	21	38		81		95	**81**	1.19
0 0 1 0 1	1		16	3	10	6		2		15	**14**	1.17
0 0 0 1 1	3		19	6	16	6		0	**29**	28		1.00
1 0 0 0 1	8		11	6	20	14		10	**48**	103		0.99
0 1 0 0 1	2		4	12	19	21		4	**16**	40		0.89
0 0 1 0 0	3		5	7	27	12	**19**	2		29		0.16
1 0 0 0 0	4		15	28	32	25	**96**	60		104		0.09
0 0 0 1 0	6		22	3	27	8	**25**	1		17		0.09
0 1 0 0 0	2		12	58	115	87	**71**	16		80		-0.02
0 0 0 0 1	21		133	40	132	54	**122**	3		125		-0.11
0 0 0 0 0	157	**816**	843	373	916	349		54		324		-0.91
	218	**816**	1100	573	1406	665	**333**	318	**93**	1470	**340**	

The majority of employees does not work in any of the disturbing circumstances. Most discomfort is experienced by blue collar workers like labourers, farmers and service personnel. All workers experiencing at least three or more adverse conditions are assigned to the group of industrial workers. So, under maximal consistency, we expect that people with many job-related harassments are labourers. Three out of 10 incomplete profiles with 2 annoyance scores are assigned to farmers. The 816 persons working in a clean environment are all assigned to the management group. This is done because this group is by far the most outspoken group.

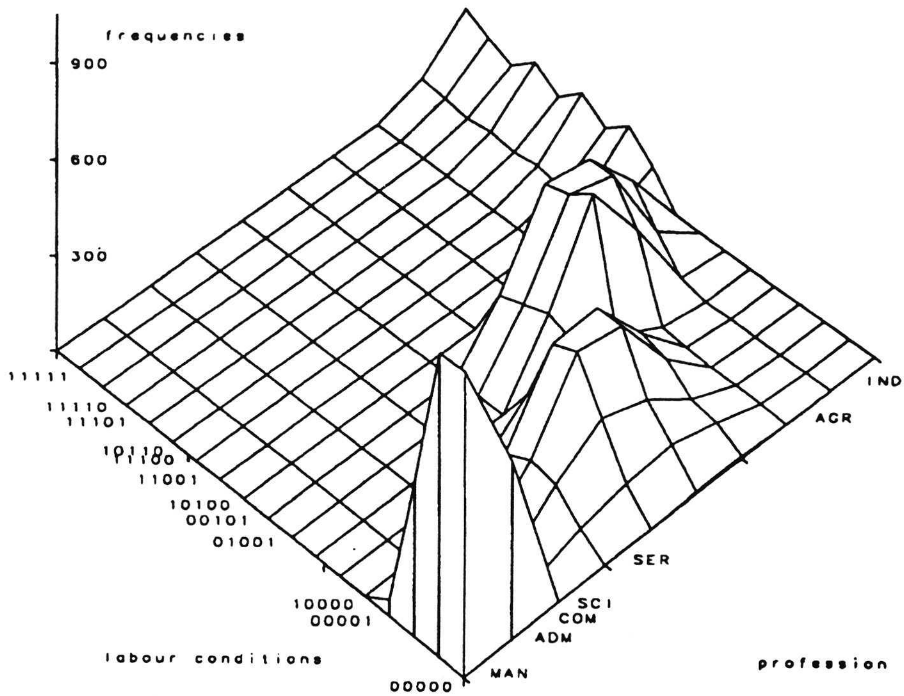
The analysis shows that it is possible to find categorical imputations such that the major trend in the data is extrapolated. Clearly, labour conditions are consistent with the type of work people do. This relationship is automatically taken into account when searching for maximally homogeneous imputation.

Figure 3.2 Observed frequencies (**Z**) versus job classes (**X**) versus labour conditions (**Y**)



The frequencies of the observed data are also pictured in Figure 3.2 by a slightly smoothed graphical analogue of the cross-tabulation in Table 3.4. The plot shows job classes on the *X*-axis, labour conditions on the *Y*-axis, and vertically on the *Z*-axis the frequencies are shown. The job classes and nuisance patterns are scaled by the consistency maximizing scores obtained by MISTRESS, with blue collar jobs relatively close together on one side and well separated from white collar jobs on the other side of the *X*-axis. The interpretation is that, based on nuisance patterns, we have two homogeneous subgroups of jobs: blue collar and white collar jobs. A similar reasoning is to be applied to labour conditions, although they do not fall apart into two groups. The conditions with few or none nuisance parameters are somewhat separated from the rest on the *Y*-axis. A consistent subset of white collar jobs experiencing hardly any nuisance in labour conditions is thus located in the lower corner pointing towards us. An intuitive interpretation of the most consistent imputation is that it should disfigure the landscape in Figure 3.2 as little as possible. Like in Figure 3.2, we can picture the frequencies of the imputed data. See Figure 3.3.

Figure 3.3 Single imputation frequencies (Z) versus job classes (X) versus labour conditions (Y)



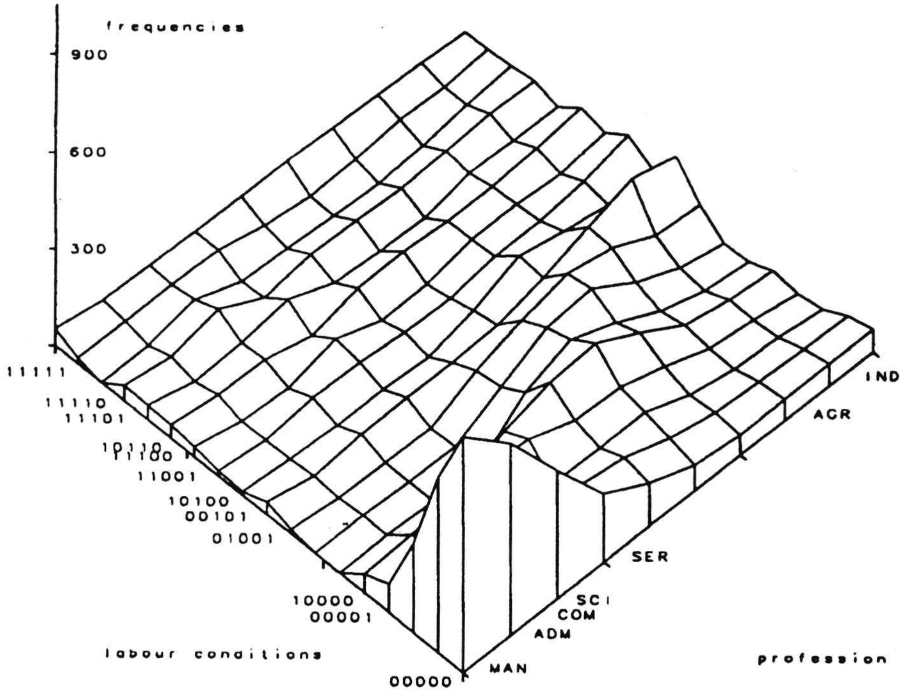
The imputations follow a curved and peaked range of frequencies from the origin {*white collar, no nuisance*} up to the far upper corner {*blue collar, maximal nuisance*}. The albeit 'reasonable' imputations are nevertheless very 'single'. All missing data with the non-nuisance pattern are singularly attributed to managers. This is a bit peculiar since other white collar workers are also reasonable candidates.

The latter observation leads automatically to the possibility of multiple imputation, where the mass of frequencies is more equally spread over other reasonable candidates. The data are completed five times by drawing imputations randomly from the predictive distributions based on inverse distances, defined in the preceding section. Although we still use the rather heuristic definition of a predictive distribution, in this example multiple imputation seems to work quite well. The multiple imputations are shown in Table 3.5 and in Figure 3.4.

Table 3.5 Multiple Imputation LSO Table (**FAT** = Imputation)

Labour Conditions	Professional Category														Donor Score
DHRSM	**MAN**		**ADM**		**COM**		**SCI**		**SER**		**AGR**		**IND**		
1 1 1 1 1	1	**2**	1	**1**	1	**2**	6	**1**	2	**2**	5	**5**	64	**5**	3.18
1 1 1 1 0	0		0	**1**	0	**1**	6		3		7	**2**	11	**5**	2.68
1 0 1 1 1	1	**1**	1		0		3		3		1	**1**	21	**3**	2.58
0 1 1 1 1	0		0		0		1		1		1		3		2.48
1 1 1 0 1	0	**2**	1	**2**	1	**3**	1	**1**	1	**2**	3	**6**	61	**7**	2.47
1 1 0 1 1	0	**2**	1		1	**2**	4	**1**	6	**1**	4	**3**	50	**6**	2.41
1 0 1 1 0	0		1		2			**1**	0		2	**2**	8	**2**	2.08
0 1 1 1 0	0		1		0		0		0		1		2		1.97
1 1 1 0 0	1	**1**	1	**1**	0	**2**	9	**1**	2	**2**	9	**5**	51	**11**	1.96
1 1 0 1 0	0		1		1	**1**	9	**1**	2	**1**	20	**5**	13	**12**	1.90
0 0 1 1 1	0		3		2		2		0		0	**1**	12	**2**	1.88
1 0 1 0 1	0		0		2		1	**1**	1	**1**	1	**4**	20	**7**	1.87
1 0 0 1 1	4	**1**	3	**2**	2	**1**	6	**1**	3	**1**	2	**7**	46	**18**	1.81
0 1 1 0 1	0		0		1		2	**1**	1		0	**2**	8	**4**	1.76
1 1 0 0 1	2	**1**	6	**1**	4	**1**	6	**1**	9	**2**	12	**9**	88	**17**	1.70
0 1 0 1 1	0		0		0		1		0		0		5	**1**	1.70
0 0 1 1 0	0		0		0		1		1		0		2	**2**	1.38
1 0 1 0 0	0		1		0		1		4		3	**3**	14	**7**	1.37
1 0 0 1 0	0		2		1		2		2		13	**2**	17	**7**	1.31
0 1 1 0 0	0		0		3		6		0		1	**3**	10	**6**	1.26
0 1 0 1 0	0		1		0		3		4		0	**1**	4	**2**	1.20
1 1 0 0 0	2	**1**	6	**1**	16		21	**1**	38	**2**	81	**39**	95	**37**	1.19
0 0 1 0 1	1		16		3		10		6		2	**7**	15	**6**	1.17
0 0 0 1 1	3		19		6		16		6	**1**	0	**14**	28	**14**	1.00
1 0 0 0 1	8	**1**	11		6	**1**	20	**1**	14	**1**	10	**22**	103	**22**	0.99
0 1 0 0 1	2		4		12		19		21	**1**	4	**10**	40	**5**	0.89
0 0 1 0 0	3		5	**1**	7	**3**	27	**2**	12	**12**	2	**1**	29		0.16
1 0 0 0 0	4	**4**	15	**7**	28	**14**	32	**13**	25	**51**	60	**4**	104	**2**	0.09
0 0 0 1 0	6	**1**	22	**3**	3	**3**	27	**4**	8	**13**	1	**1**	17		0.09
0 1 0 0 0	2	**3**	12	**6**	58	**11**	115	**15**	87	**32**	16	**2**	80	**1**	-0.02
0 0 0 0 1	21	**6**	133	**11**	40	**28**	132	**31**	54	**42**	3	**3**	125	**1**	-0.11
0 0 0 0 0	157	**314**	843	**276**	373	**98**	916	**86**	349	**32**	54	**5**	324	**5**	-0.91
	218	**341**	1100	**314**	573	**172**	1406	**164**	665	**200**	318	**171**	1470	**220**	

Figure 3.4 Average frequencies based on multiple imputation (Z) versus job classes (X) versus labour conditions (Y)



The listed imputation frequencies are the average over five multiple imputation. Because of rounding errors, not all imputations exactly add up to the marginal frequencies. Comparing Figures 3.3 and 3.4, it is obvious that the multiple imputations are more spread over jobs and nuisance patterns. Both imputations, single and multiple, follow the same gradient from {*white collar*, *no nuisance*} on the bottom to {*blue collar*, *maximal nuisance*} on top.

3.7 CONTINGENCY TABLES

This example compares some aspects of the treatment of missing data in loglinear analysis to the present method. We use the 2 × 2 × 2 table given in Little and Rubin

(1987, p. 187). The data pertain to a partly real life, partly artificial example made up by Little and Rubin. There exist three dichotomized variables: *survival* (S), *type of clinic* (C) and *amount of prenatal care* (P). Type of clinic is unknown for 255 observations (= 26%), which means that 8.8% of the observations in the three-way table is missing. Table 3.6 indicates that the preferred loglinear model for the table based on the 715 complete observations is [SC, PC], which means that type of clinic is related to survival and to the amount of prenatal care. Moreover, within the same clinic, survival and prenatal care are not related. Because deletion of the association [SP] does not alter the fit, the more parsimonious model [SC, PC] is preferred. Model [SP, SC] does not fit at all.

Table 3.6 Chi-square and p-values under EM and MISTRESS imputation

Model	Completely Classified	p	Imputation by EM	p	Imputation by MISTRESS	p
[SP, SC, PC]	0.044	0.834	0.057	0.810	4.77	0.029
[SC, PC]	0.083	0.959	0.031	0.984	9.76	0.008
[SP, SC]	169.469	0.000	0.002	0.999	355.16	0.000

The second pair of columns in Table 3.6 contains the χ^2-values that measure the difference between the expected values under the three loglinear models and the imputed contingency tables (cf. Little and Rubin, 1987, p. 190-191). These values are not statistically significant, so all models fit the data. This is caused by, amongst others, the fact that the EM algorithm finds the most favourable imputations *given* the specific loglinear model. In general, loglinear models will fit better as more missing observations are added. In most cases, this will preserve-and even emphasize-the structure among variables as described by the loglinear model.

However, things can also go less well. Observe that model [SP, SC] now fits the imputed table (p = 0.999). For the completely classified table this model does not fit at all (p= 0.000), so filling in missing data brought about some real change. But this is a hazardous aspect of EM: suppose that we really had the 255 missing observations as in Little and Rubin, and that we applied EM. Then, we would have been pleased to find a χ^2-value as low as 0.002, and we would have had little reason to question the validity our model. If we compute correlations we see what has happened: the original correlation - actually a ϕ - coefficient here - of the omitted **[PC]** effect is equal to -0.4924, which is substantial. After EM, it is -0.0130! Imputation corrupted the correlation. Rubin (personal communication) shows that in this case multiple or single imputation makes no substantial difference in estimation of the model parameters for the loglinear models. The same catch, though in the opposite direction, holds for MISTRESS. Because MISTRESS optimizes a different, almost reverse criterion, the imputed tables do not fit the loglinear model as well as the ones produces by EM. None of models fit to the imputed data. On the other hand, the χ^2-statistic clearly signals the important **[PC]** interaction.

Both the EM algorithm and our method have the same basic weakness: *if the model is wrong, imputations will be wrong*. If the model prescribes that a certain interaction does not exist, then EM will do everything to make this true. In the above case, it makes a correlation of 0.49 disappear. Analogously, MISTRESS overemphasizes tiny correlations. Generally speaking, loglinear analysis stresses absence of particular interaction, while maximizing consistency emphasizes presence of overall interaction. If we suspect that the analysis results are heavily biased by imputation, we should use these properties to our advantage.

3.8 CONCLUSION

The technique proposed in this paper is fairly simple. For categorical data, it is a way of selecting the proper category to be imputed. In the context of maximizing consistency, this seems to be new. The method optimizes a well-defined and widespread criterion. In addition, it is fast, flexible and of high practical value. Few assumptions are needed. The method stays close to the data.

It is possible to simulate various hot deck strategies. For example, by (over)weighting one of the variables we simulate a single donor variable. The method then evolves into a traditional hot deck method. In the same way, it is possible to rule out specific variables from the donor. Careful selection of variables may drastically improve the quality of the solution. Non-ignorable models, in which the pattern of nonresponse depends on the values of the data, can be evaluated by adding an indicator variable for the nonresponse distribution for each variable. Mixes of continuous and discrete data can also be analyzed. Since imputations are determined for each variable separately, mixing does not present any new problems.

There are also situations in which the method will perform less satisfactorily. The main concern is the amount of intercorrelation. If the magnitude of all correlations is below 0.20 then the method may generate imputations that overemphasize small correlations. In this case, random imputation or unconditional mean imputation often work better. It seems preferable to use MISTRESS here only in combination with a resampling method, like the bootstrap, in order to estimate the variability of consistency. Van Buuren and van Rijckevorsel (1992a, 1992b) found that the average correlation should be at least 0.50 before the method becomes practical. At that point, the technique gives reasonable results up to 10-15% missing data.

A second cautionary note concerns imputation itself. However attractive the idea may seem, we must never forget that once after we have completed the data, they are partly artificial. The main pitfall is to analyze the filled-in data as if they were real, and thus overstate precision. The sagacious researcher will set up a subconscious alert that signals any peculiarities that might result from imputation. According to Dempster and Rubin (1983), the entire idea of imputation carries one great seductive danger: "it can lull the user into the pleasurable state of believing that the data are complete after all."

We conclude with some words on applications and perspectives. The number of variables or observations hardly influences the computational efficiency of the method. Therefore, the technique can be used with large data matrices. The main application field of MISTRESS is the analysis of surveys, rating scales and questionnaires. Furthermore, the relationship with Cronbach's α makes it attractive for dealing with missing data in psychological testing. It is easy and cheap to reiterate MISTRESS in any fashion, combined with bootstrapping, multiple imputation and the like.

The most spectacular application of the method is multiple imputation. It is ambiguous to call it an application. As a matter of fact one might devote another study to MISTRESS and multiple imputation. Analytical problem number one is to decide how implicit predictive densities are to be derived within the present framework. To this purpose, we mentioned in this paper only some intuitive possibilities. This of course needs further study. It is particularly interesting to examine the shape of the predictive density function under varying levels of consistency. If the consistency equals zero, a trivial possibility, the predictive distribution should be uniform. Conversely, if the consistency approaches unity, it should have zero variance. And then, we are back at MISTRESS.

REFERENCES

BUCK, S.F. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. J. R. Statis. Soc. B22 (1960) 302-306.

BUUREN, VAN, S. & W.J. HEISER. Clustering *n* objects into *k* groups under optimal scaling of variables. Psychometrika 54 (1989) 699-706.

BUUREN, VAN, S. & J.L.A. VAN RIJCKEVORSEL. Fast linear least squares imputation of missing data. RR-01-91, Leiden, University of Leiden, Dept. of Psychology, 1991.

BUUREN, VAN, S. & J.L.A. VAN RIJCKEVORSEL. Imputation of missing categorical data by maximizing internal consistency. Psychometrika 1992A. (to appear)

BUUREN, VAN, S. & J.L.A. VAN RIJCKEVORSEL. Data augmentation and optimal scaling. To appear in: R. Steyer & K.F. Wender (eds) Proceedings of the 7th European Meeting of the Psychometric Society, July 29-31, Trier. Heidelberg, Springer-Verlag, 1992b.

CRONBACH, L.J. Coefficient alpha and the internal structure of tests. Psychometrika 16 (1951) 297-334.

DEAR, R.E. A principal component missing data method for multiple regression models, SP-86. Santa Monica Cal., System Development Corporation, 1959.

DEMPSTER, A.P. & D.B. RUBIN. Overview. In: W.G. Madow, I. Olkin & D.B. Rubin (eds) Incomplete data in sample surveys, vol. 2, Theory and annotated bibliography. New York,
Academic Press, 1983. pp. 3-10.

DODGE, Y. Analysis of experiments with missing data. New York, Wiley, 1985.

FORD, B.L. An overview of hot deck procedures. In: W.G. Madow, I. Olkin & D.B. Rubin (eds), Incomplete data in sample surveys, vol. 2, Theory and annotated bibliography. New York, Academic Press,
1983.

GIFI, A. Nonlinear multivariate analysis. Chichester, Wiley, 1990.

GLEASON, T.C. & R. STAELIN. A proposal for handling missing data. Psychometrika 40 (1975) 229-252

GREENACRE, M.J. Theory and applications of correspondence analysis. New York, Academic Press, 1984.

HARTLEY, H.O. & R.R. HOCKING. The analysis of incomplete data. Biometrics 27 (1971) 783-808.

HEDGES, B. & I. OLKIN. Selected annotated bibliography. In: W.G. Madow, I. Olkin & D.B. Rubin (eds) Incomplete data in sample surveys, Vol. 2, Theory and annotated bibliography. New York, Academic Press, 1983. pp. 3-10.

KALTON, G. & D. KASPRZYK. Imputing for missing survey responses. Proceedings of the Section of Survey Research Methods, 1982, American Statistical Association, 1982. pp. 22-33.

LITTLE, R.J.A. & D.B. RUBIN. Statistical analysis with missing data. New York, Wiley, 1987.

LITTLE, R.J.A. & D.B. RUBIN. The analysis of social science data with missing values. In: J. Fox & T. Scott Long (eds) Modern methods of data analysis. London, Sage, 1990. pp. 374-409.

LORD, F.M. Some relations between Guttman's principal components of scale analysis and other psychometric theory. Psychometrika 23 (1958) 291-296.

MADOW W.G., OLKIN I. & D.B. RUBIN (eds). Incomplete data in sample surveys, Vol. 1, 2 and 3. New York, Academic Press, 1983.

MEULMAN, J. Homogeneity analysis of incomplete data. Leiden, DSWO Press, 1982.

NISHISATO, S. Analysis of categorical data: Dual scaling and its applications. Toronto, University of Toronto Press, 1980.

RUBIN, D.B. Multiple imputation for nonresponse in surveys. New York, Wiley, 1987.

RUBIN, D.B. EM and beyond. Psychometrika 56 (1991) 241-254.

RUBIN, D.B. & N. SCHENKER. Multiple imputation in health-care databases: An overview and some applications. Stat. Med. 10 (1991) 585-589.

SCHNELL, R. Missing Data Probleme in der empirischen Sozialforschung [Missing data problems in the social sciences]. Inaugural dissertation, Bochum, Ruhr Universität, 1986.

BUILDING STATISTICAL DATABASES FOR AN AIDS SURVEY

A.G.C. Vogels*, M.M. van der Klaauw**

* Department of Children and Health, TNO Institute of Preventive Health Care, Leiden, The Netherlands

** Department of Statistics and Computer Science, TNO Institute of Preventive Health Care, Leiden, The Netherlands

Abstract

The study 'Health, Behaviour and Relations among Adolescents' was designed to provide for a strong empirical base for adequate health education with regard to sexually transmitted diseases in general and AIDS in particular; secondly, it should provide a better estimation of the risks for adolescents of HIV-contagion.
In this study, data on sexual behaviour, attitudes and knowledge were collected by means of a questionnaire that is to be completed in the classroom.

The process of data collection has been complicated by several factors. The study aimed to generalize over sex, age, type of education, religious orientation of schools involved and social economic status, for 6 distinct Dutch regions as well as for the country as a whole. Several relevant variables (e.g. homosexuality and sexual experience) were expected to have little variation; to ensure a sufficient number of observations a large sample (n=12.000) had to be approached. Data collection required cooperation of school administrators. Because of the delicate content of the questionnaire a rather high refusal rate was expected related to religious orientation of schools. The data had to be collected in cooperation with more than 45 different organizations (the Sentinel Stations for Youth Health Care); data collection had to be completed within 3 months.
To ensure a controlled collection of data on all relevant variables, several measures had to be taken. These included APRI, a Automatic Project Administration System, by means of which the representativeness of the sample was safeguarded.

These measures and their consequences will be discussed.

Keywords: *statistical databases, HIV, adolescents, sexual behaviour.*

4.1 INTRODUCTION

The study 'Health, Behaviour and Relations among Adolescents' (HBR) focuses on behaviourial, cognitive and attitudinal aspects of adolescent sexuality. For several reasons, the process of data collection, data entry and, without doubt, data analysis is rather complicated. While designing and planning the study we encountered a lot of problems. These had to be solved in order to ensure that we should indeed obtain the data that are necessary to answer the questions we had to answer. This is how we want to use the expression 'building a statistical database'.

More explicitly: we define a statistical database simply as a database containing data on which statistical analyses are to be performed in order to answer specific questions.

We shall use the term 'building' to indicate all activities undertaken to collect and record data in such a way that the questions can be answered.

First, we start with a very short introduction on the aims and the participants of the study.After that we want to concentrate on one single aspect of building the statistical database for this study: the difficulty of maintaining the integrity of the sample, given a very complicated set of sample specifications and an expected high and selective refusal rate. In doing so we will present some preliminary results of our efforts.

4.2 THE STUDY 'HEALTH, BEHAVIOUR AND RELATIONS AMONG ADOLESCENTS'

The general title 'Health, Behaviour an Relations among Adolescents' conceals a study on sexual behaviour, attitudes and knowledge of adolescents aged eleven till seventeen. Mainly inspired by the risk of HIV-infection, the study aims to provide for an adequate basis for health education and for a better estimation of the real risks confronting this age group.

The study is done in cooperation with the Dutch Institute of Social Sexological Research and the Sentinel Stations for Youth Health Care.

Questionnaires and research procedures were developed by the research institutes.

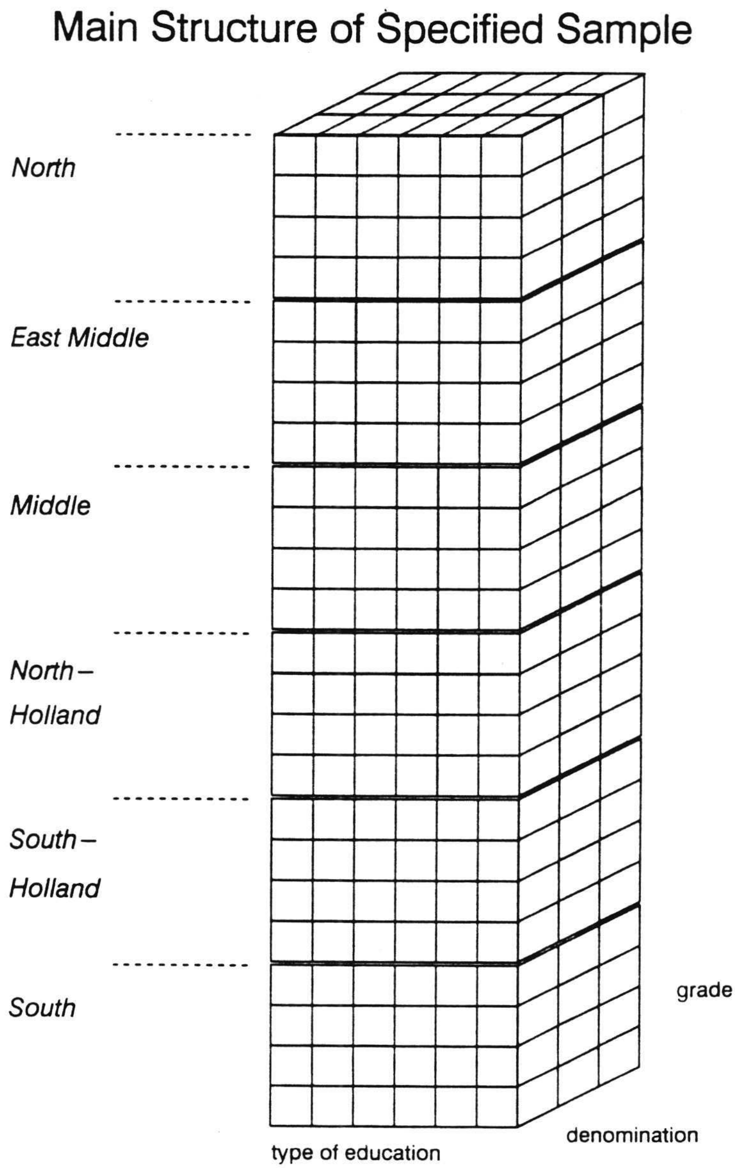
The actual data collection was carried out by the staff of the 45 participating Sentinel

Stations of Youth Health Care, all over The Netherlands.
The study started in September '88. After a period of intensive preparation, data collection started in September '89 and had to be completed within three months.

4.3 THE SPECIFICATION OF THE SAMPLE

At this moment up-to-date empirical data on the degree and nature of sexual experiences, knowledge and attitudes of young people in The Netherlands are lacking. There have been a few small-scale investigations that show a remarkable weakness in the area of generalizability.
HBR aims to generalize over sex, type of secondary education - implying different social categories -, religious orientations of schools, and grade - implying different ages -. Furthermore, the sampling technique should allow us to study differences between ethnic groups. Even with a rather crude categorization, these criteria imply a large number of categories and therefore a large number of youngsters to be approached.
However, for several reasons one cannot but assume important differences between different regions in The Netherlands. Therefore the design should also allow for representativeness on a regional level. The number of youngsters to be approached was thus getting larger and larger.

Figure 4.1 a 7 fold cube



Furthermore, it will be clear that several key variables in this study will show a limited

degree of variation.
For example; it is very unlikely that more than 3% of this age group will have developed a more or less stable homosexual identity, not to mention the very limited number of youngsters with real homosexual experiences and willing to be open about it. It is obvious that this group could not be ignored in a study like this.
Eventually, a sample size of about 12,000 was decided upon, not as the result of arithmetic but as an educated guess. For the sampling we used a database from the Ministry of Education. This database contained almost all information we needed.
Information regarding the number of immigrant pupils was however lacking. Random sampling would almost certainly result in insufficient numbers for this group. Therefore, we approached local authorities and asked them to provide us with the names of schools with a large immigrant population. In sampling these schools got a certain degree of priority.

4.4 MAINTAINING THE INTEGRITY OF THE SAMPLE DURING DATA COLLECTION

There is more to data collection, however, than specifying which data concerning which group you want to obtain. You have to get the job done. The only feasible way to collect self-report data among thousands of youngsters is to approach them in the classroom. Such a procedure does have some disadvantages.

Figure 4.2 possible sources of loss from the specified sample

Main levels of possible loss from the sample

- **Department of Youth Health Care**
  - *all schools in district*
  - *individual schools*
- **School administration**
  - *the school as a whole*
  - *individual classes*
- **Individual respondents**
  - *refusing parents*
  - *refusing, sick pupils; absenteism*

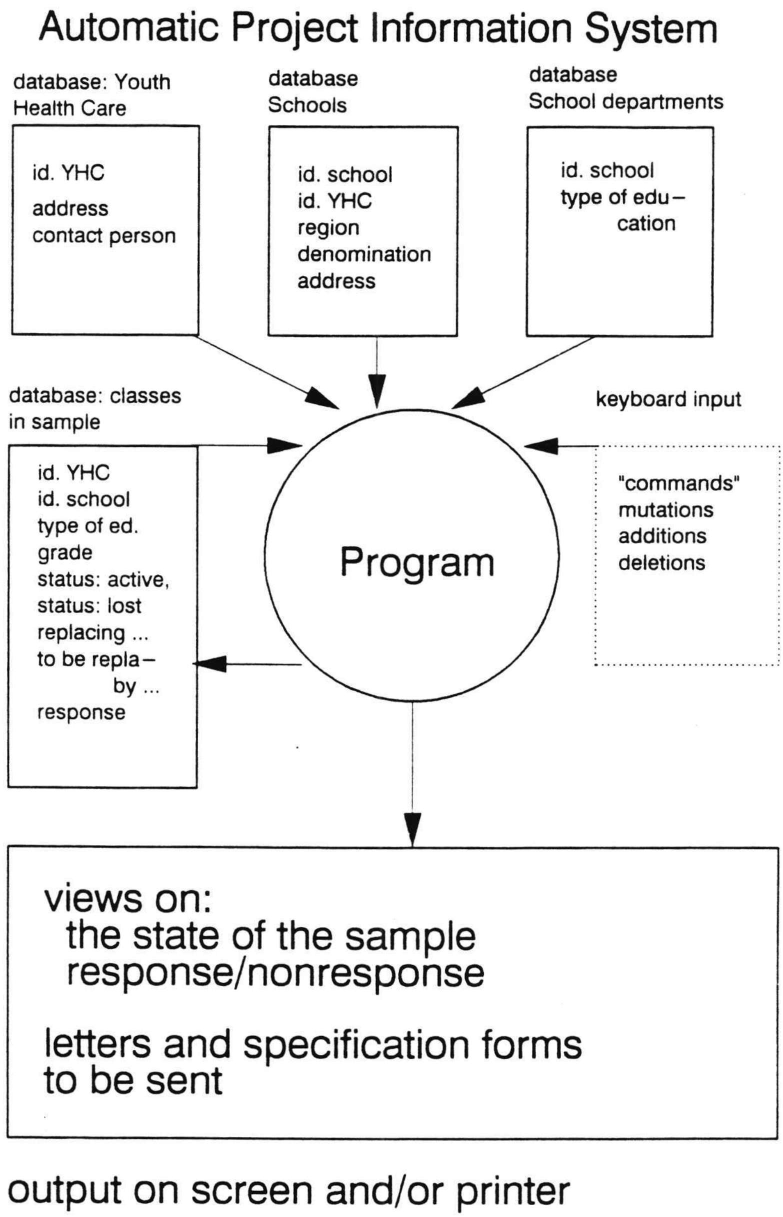
Figure 4.2 presents possible threats to the integrity of the sample during actual data collection. The most serious problems, were expected to arise at school level.
The administration of every single school would have to give permission to enter the classroom. We did not expect each administrator to be too eager to see his pupils confronted with questions on sexual intercourse, anal intercourse, homosexual desires and so on. Refusals were therefore inevitable.
Besides, we expected these refusals to be selective with regard to the sampling criteria. The organisation of the Dutch educational system is partially based on religion. Some of the organisations involved are known to have a negative attitude towards being too explicit about sex.
Such selective refusal rates posed a serious risk to the representativeness of the data collected. This risk had to be controlled, as far as possible.
For this purpose we decided to replace the classes of every refusing school with corresponding classes of a randomly chosen school in the same sampling category: same denomination, same grade and so on. Ethnicity, of course, was also controlled for.

This replacement procedure was to be continued, until a school willing to cooperate was found.

4.5 APRI

The latter implied that a lot of work had to be done in the limited period reserved for data collection. Approaching one school and getting a decision to cooperate or not to cooperate, inevitably, takes some time.
It will be clear that the processing of a refusal replacement procedure had to be done quickly - and without errors. This was one of the main reasons why we decided to build 'APRI', an acronym for 'automatic project information system'.

Figure 4.3 main functions and output of APRI



APRI is a dBASE application. It consists of several interrelated databases and a programme, integrating about 30 procedures. The purpose of building APRI, in short,

is to make the management of sampling, data collection and response administration less labour-intensive, less error-prone and more easy. Time does not allow for a detailed explanation. Figure 4.3 gives a global view on the main functions and output of the system.
With APRI the computer could take the heavy burden of the actual administrative procedures from our shoulders, though only after careful planning, intensive preparation and regular maintenance of the system.

4.6 REALISATION OF THE SAMPLE

APRI is a novelty in the management of data collection. As such it cost time, money and effort. Was this worth all the effort ? In our opinion, yes. For a short period we had to work without APRI. Every replacement took us at least 1.5 hour to process. With APRI, it took less than 10 minutes.
More important, however, is the question to what degree the integrity of the sample could indeed be maintained.
The original sample contained one hundred and seventy schools. Nearly a quarter of these were completely lost, for several reasons. This loss is clearly related to the sampling criteria.
As mentioned before, such losses were expected and it was for this very reason that we decided to implement a replacement procedure. Following this procedure and supported by APRI and by the highly motivated Sentinel Stations of Youth Health Care, we were able to maintain the representiveness of the data, in accordance with the criteria mentioned before.

Figure 4.4 composition of the original and the realised sample; in numbers of classes; by region, denomination and type of education by class.

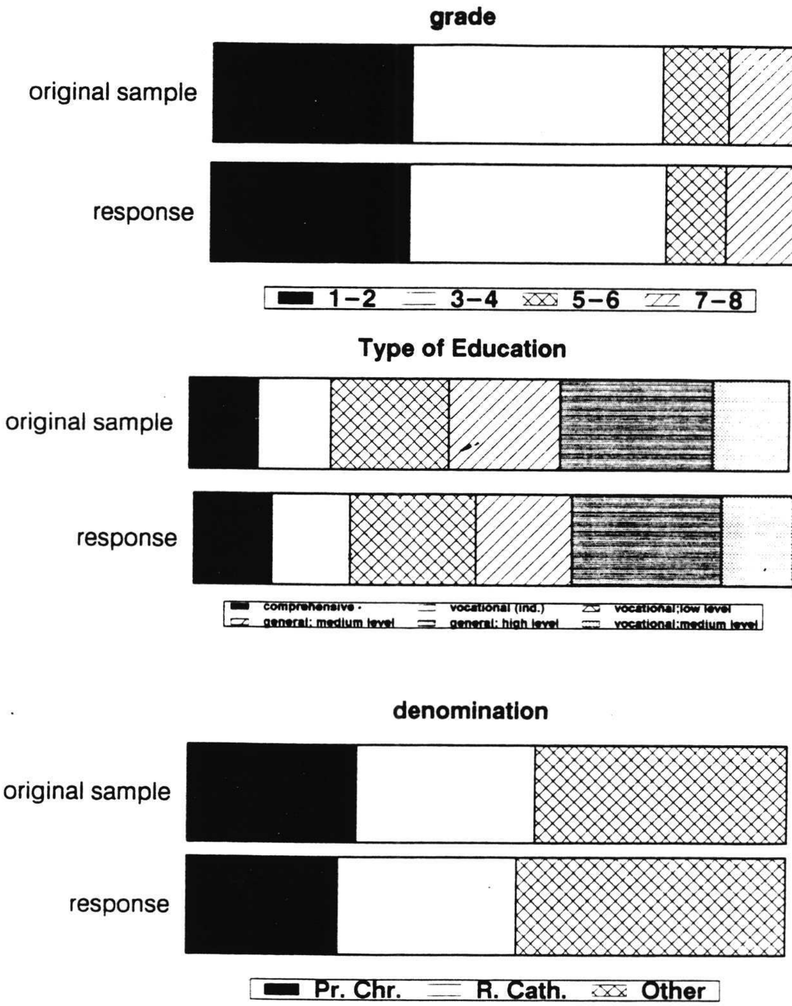


Figure 4.4 allows you to compare the composition of the original sample and that of the dataset which was ultimately collected. As you can see, they are nearly identical, in terms of grade, type of education and even denomination.

4.7 CONCLUSION

We could only comment on some of the problems we were confronted with during the process of building the statistical database for our study. Despite this, I hope it will be clear that it involved a lot of careful planning, management and sheer hard work. In this process we relied heavily on general and specific computer applications.
Did they save us time ?
Yes, all things considered, they did.
They could not prevent schools from being lost. They could not be expected to.
However, using the procedures described, we managed to avoid drowning in a bottomless lake of administrative procedures; we succeeded in getting the job done in the limited time available; and we succeeded in maintaining the representativeness of the data set, despite serious problems, due to the delicate nature of the study.
With the data now available we can indeed hope to answer the questions we have to answer.

REFERENCES

VOGELS, T. & DANZ, M. Gezondheid, gedrag en relaties: Draaiboek. [Health, behaviour and relations: Scenario.]. Leiden, NIPG-TNO, 1989.

VOGELS, T., KLAAUW VAN DER, M.M., & M. VAN LAARHOVEN. Werken met APRI: handleiding en procedures. [Using APRI: manual and procedures.] Leiden, NIPG-TNO, 1989. (internal report)

VOGELS, T., VLIET VAN DER, R.W.F., DANZ, M.J. & M. HOPMAN-ROCK. Verslag vooronderzoek Gezondheid, Gedrag en Relaties: een onderzoek van de Peilstations Jeugdgezondheidszorg [Report preliminary research Health, Behaviour and Relations: a study by the Sentinel Stations for Youth Health Care]. Leiden, NIPG-TNO/NISSO, 1990.

SOME APPLICATIONS OF THE LIFE-TABLE TECHNIQUE TO DETERMINE INDICATORS OF HEALTH STATUS

J.K.S. van Ginneken*

* Department of Public Health and Prevention, TNO Institute of Preventive Health Care, Leiden (Presently at Netherlands Interdisciplinary Demografic Institute (NIDI), The Hague, The Netherlands)

Abstract

The most common application of the life table method is in the field of mortality studies; the result of such a life table is the life expectancy at birth and at other ages. During the past 10 to 20 years a number of techniques have been developed that make the life table method also suitable for analysis of causes of death and morbidity. In this paper several of these newer applications will be discussed and examples of their use will be given. One of these applications with respect to causes of death, is the determination of gains in life expectancy due to elimination of a cause of death. More recent innovations are determination of life expectancy of those who die of a particular cause, and gains in life expectancy due to elimination of a cause for the so-called saved population. Another application of the life table method is in the field of morbidity where it has been used to construct a measure or index which combines information on both mortality and morbidity. An example of such a measure is the life expectancy free of disability.

Keywords: *life table, mortality, morbidity, life expectancy.*

5.1 INTRODUCTION

A number of years ago Sullivan proposed to utilize a summarizing measure of mortality and morbidity (Sullivan, 1971a). He used different terms to describe this measure: a single index of mortality and morbidity, a mortality-morbidity index and life expectancy free of disability. Since that time his proposals have been applied in a number of countries in particular the United States of America, Canada, France and the United Kingdom. In this paper we will describe, in the first place, the construction of such an index for the Netherlands. Next, we will present the main results of the calculations with special reference to differences in life expectancy free of disability by sex. Finally, the argument is made that life expectancy free of disability is an important health indicator to be used by countries to evaluate the long-term impact of health policies.

5.1.1 Concept, methods and data

Sullivan defined disability broadly as "any temporary or long-term reduction or restriction of a person's activity as a result of illness or injury" (Sullivan, 1971a). He distinguished three categories of disabled persons:

a. Persons with long-term institutional disability (persons confined to resident institutions for health care).
b. Persons with long-term non-institutional disability (persons living in the community but with enduring limitations of activities).
c. Persons with short-term disability (persons in the community who are temporarily disabled).

In addition, he provided operational definitions of these terms and gave an example how disability rates could be used to determine life expectancy free of disability (Sullivan, 1971b).

The methods which were developed by Sullivan have been applied by a number of investigators on data in the USA (Colvez, 1983), Canada (e.g. Wilkins and Adams, 1983), France (e.g. Robine, Brouard and Colvez, 1987) and England and Wales (Bebbington, 1988). We would like to add that the impression exists that the operational definitions used by these investigators are somewhat different from those used by

Sullivan; it is not always easy to verify to what extent this is the case due to lack of details in the above-mentioned publications on this subject.

In this paper we will describe the various steps leading to the construction of two measures of life expectancy free of disability in the Netherlands. This will be done on the basis of the approach developed by Sullivan, but, as we will see below, some changes in his methodology were necessary and, in addition, a second measure was used as well. For each of the various categories of disability we will describe in more detail the operational definitions which were adopted, calculations which needed to be made, adjustments which were needed and sources of data used.

Two approaches will be used to determine levels of disability. According to Method 1 long-term and short-term disability will be ascertained separately by means of questions from the Continuous Health Interview Survey of the Netherlands Central Bureau of Statistics (NCBS) while according to Method 2 both types of disability will be measured together with a single question on perceived health status from the same NCBS survey. More details on data used and methods of calculation are found in van Ginneken, Bannenberg and Dissevelt (1989).

Long-term institutional disability (Method 1 and 2)

Age- and sex-specific rates of prevalence of institutional disability are derived from admission and discharge figures of various types of institutions in particular mental hospitals, nursing homes, institution for the mentally deficient and homes for the aged. With respect to the number of persons in homes for the aged, a correction needed to be made because a proportion of persons staying in such homes are in reasonable or good health. Such persons were excluded from the total number of persons staying in institutions.

Long-term non-institutional disability (Method 1)

A second disability component consists of prevalence of long-term non-institutional disability leading to limitation of normal daily activities. This type of disability was measured with the OECD indicator consisting of 14 items on the ability to carry out a number of activities which are considered to be indispensable for normal daily functioning such as able/unable to dress, able/unable to get in and out of bed etc. The analysis to be described here has made use of an indicator consisting of 10 items (van Sonsbeek, 1988). Data were derived from the NCBS Continuous Health Interview Survey which was limited to the non-institutional population of the Netherlands (Nederlands Central Bureau of Statistics, 1988).

Short-term disability (Method 1)

The third dimension of disability consists of prevalence of short-term disability due to illness. Data for this component are also derived from questions of the NCBS Continuous Health Interview Survey on limitation of activity in the 14 days prior to the date of interview. It was necessary to make two corrections, because there is a certain amount of overlap of short-term and long-term disability. A number of persons who reply that they have experienced limitations of activity during the past two weeks, are those who already suffer from long-term disability. The first correction consists thus of subtraction of these persons from the total number of persons who state to have experienced short-term disability during the past two weeks.

These excluded persons do, however, suffer from different forms of short-term disability and the second correction consists of addition of the number of person-years spent with different forms of short-term disability for those persons. Exact data on this number of person-years were not available and these were, therefore, estimated.

Long-term and short-term disability (Method 2)

An alternative method to measure long-term and short-term disability was to use information from the NCBS survey on the question: "How is your health in general?" Answers were classified into 5 categories: very good, good, fair, sometimes good and sometimes bad, and bad. All persons who replied that their health was sometimes good and sometimes bad, and bad, were considered not to be in good health and such persons were in this paper defined as suffering from disability leading to restriction of activities. It was necessary to adjust the figures obtained for the following reason. If the calculations are based on the unadjusted percentages of persons who are in less good health, it is assumed that these persons experienced less good health for the duration of a whole year. Some persons were ill, however, for a short period of time while others were disabled for a whole year. Using information from the NCBS survey on the duration of short-term disability, it was possible to make this adjustment. The details are described in van Ginneken, Bannenberg and Dissevelt (1989).

Method of calculation

The point of departure of the various calculations is the abbreviated life table for men and women separately in a certain year or period. The ${}_nL_x$ column of this life table represents the number of years lived in the age interval between x and x + n years. In each age interval one has now to subtract the number of years spent in different states of disability. One obtains these figures by multiplication of the number of years lived in an age interval (${}_nL_x$) with the disability rates for this particular age interval. When figures on the number of years spent in a disabled state have been calculated for all age intervals, one can then calculate life expectancy with disability at various ages and this leads to life expectancy free of disability.

5.2 RESULTS

Table 5.1 shows that according to the first method of calculation men can expect from birth onwards to spend 13.3 years with different forms of disability and women 21.3 years (in 1981-1985).

Table 5.1 Life expectancy without disability (at birth) according to two methods of calculation, 1981-1985 (in years)

	Men		Women	
	Method 1*	Method 2*	Method1*	Method 2*
Total life expectancy	72.81	72.81	79.46	79.46
In institutions	0.99	0.99	2.46	2.46
Long-term disability	8.87	13.03	15.04	16.29
Short-term activity limitation	3.45		3.80	
Years with disability	13.31	14.02	21.30	18.75
Life expectancy without disability	59.50	58.79	58.16	60.71

* Method 1: using data on short-term and long-term activity limitation; Method 2: using data on perceived health status

Using the second method of calculation, the number of years in less good health is 14.0 for men and 18.8 years for women. An implication of this result is that although women have a higher total life expectancy than men (79.5 years for women compared to 72.8 years for men) this difference becomes much smaller or even goes in the opposite direction when we focus on life expectancy free of disability. According to Method 2 female life expectancy free of disability is 1.9 years higher than male life expectancy; according to Method 1 male life expectancy free of disability exceeds female life expectancy by 1.3 years.

Table 5.1 further shows that men can expect to spend about 1 year out of 72.8 years in institutions and women 2.5 years out of 79.5 years. The number of years with long-term and short-term disability leading to restriction of activity for men was 12.3 years and 13.0 years for Method 1 and 2 respectively. Corresponding values for women were 18.1 and 16.3 years respectively.

Data on life expectancy free of disability at various ages according to both methods are found in tables 5.2 and 5.3.

Table 5.2 Life expectancy without disability by age and sex using data on short-term and long-term activity limitation (Method 1), 1981-1985 (in years)

Age and sex	Total life exp.	In insti-tutions	Long term dis-ability	Short-term limi-tations	Years with dis-ability	With-out dis-ability	% dis-abled
MEN							
0	72.81	0.99	8.87	3.45	13.31	59.50	18.3
5	68.66	1.00	8.82	3.33	13.15	55.51	19.2
10	63.75	1.00	8.68	3.18	12.86	50.89	20.2
15	58.84	0.98	8.54	3.09	12.61	46.23	21.4
20	54.01	0.97	8.43	2.96	12.36	41.65	22.9
25	49.23	0.95	8.28	2.75	11.98	37.25	24.3
30	44.41	0.93	8.12	2.49	11.54	32.87	26.0
35	39.60	0.91	7.97	2.21	11.09	28.51	28.0
40	34.84	0.89	7.69	1.92	10.50	24.34	30.1
45	30.17	0.88	7.37	1.58	9.83	20.34	32.6
50	25.67	0.87	6.80	1.25	8.92	16.75	34.8
55	21.44	0.88	6.10	0.94	7.92	13.52	36.9
60	17.53	0.90	5.47	0.71	7.08	10.45	40.4
65	14.02	0.96	4.65	0.47	6.08	7.94	43.4
70	10.99	1.06	4.09	0.31	5.46	5.53	49.7
75	8.45	1.23	3.61	0.20	5.04	3.41	59.6
WOMEN							
0	79.46	2.46	15.04	3.80	21.30	58.16	26.8
5	75.19	2.49	15.07	3.67	21.23	53.96	28.2
10	70.26	2.48	14.96	3.51	20.95	49.31	29.8
15	65.32	2.48	14.83	3.59	20.70	44.62	31.7
20	60.41	2.46	14.54	3.14	20.14	40.27	33.3
25	55.51	2.45	14.19	2.86	19.50	36.01	35.1
30	50.61	2.44	13.82	2.53	18.79	31.82	37.1
35	45.75	2.42	13.44	2.15	18.01	27.74	39.4
40	40.93	2.42	13.03	1.87	17.32	23.61	42.3
45	36.18	2.41	12.45	1.57	16.43	19.75	45.4
50	31.55	2.42	11.64	1.27	15.33	16.22	48.6
55	27.07	2.44	10.81	1.01	14.26	12.81	52.7
60	22.73	2.48	9.64	0.77	12.89	9.84	56.8
65	18.57	2.55	8.50	0.58	11.63	6.94	62.6
70	14.67	2.65	6.91	0.40	9.96	4.71	67.9
75	11.14	2.76	5.33	0.24	8.33	2.81	74.8

Table 5.3 Life expectancy without disability by age and sex using data on perceived health status (Method 2), 1981-1985 (in years)

Age and sex	Total life exp.	In insti-tutions	In less good health	Years with dis-ability	With-out dis-ability	% dis-abled
MEN						
0	72.81	0.99	13.03	14.02	58.79	19.3
5	68.66	1.00	12.78	13.78	54.88	20.0
10	63.75	1.00	12.45	13.45	50.30	21.1
15	58.84	0.98	12.24	13.22	45.62	22.5
20	54.01	0.97	12.09	13.06	40.95	24.2
25	49.23	0.95	11.87	12.82	36.41	26.0
30	44.41	0.93	11.57	12.50	31.91	28.1
35	39.60	0.91	11.11	12.02	27.58	30.4
40	34.84	0.89	10.55	11.44	23.40	32.8
45	30.17	0.88	9.93	10.81	19.36	35.8
50	25.67	0.87	8.99	9.86	15.81	38.4
55	21.44	0.88	7.77	8.65	12.79	40.3
60	17.53	0.90	6.48	7.38	10.15	42.1
65	14.02	0.96	5.12	6.08	7.94	43.4
70	10.99	1.06	4.12	5.18	5.81	47.1
75	8.45	1.23	3.21	4.44	4.01	52.5
WOMEN						
0	79.46	2.46	16.29	18.75	60.71	23.6
5	75.19	2.49	16.23	18.72	56.47	24.9
10	70.26	2.48	15.95	18.43	51.83	26.2
15	65.32	2.48	15.76	18.24	46.99	27.9
20	60.41	2.46	15.41	17.87	42.54	29.6
25	55.51	2.45	15.01	17.46	38.05	31.5
30	50.61	2.44	14.55	16.99	33.62	33.6
35	45.75	2.42	14.02	16.44	29.31	35.9
40	40.93	2.42	13.41	15.83	25.10	38.7
45	36.18	2.41	12.59	15.00	21.18	41.4
50	31.55	2.42	11.50	13.92	17.63	44.1
55	27.07	2.44	10.17	12.61	14.46	46.6
60	22.73	2.48	8.68	11.16	11.57	49.1
65	18.57	2.55	7.11	9.66	8.91	52.0
70	14.67	2.65	5.53	8.18	6.49	55.8
75	11.14	2.76	3.97	6.73	4.41	60.4

Of interest in both tables is the increase in the percentage of time spent with various forms of disability by age in the last column of both tables. This increase is most obvious in the highest age groups (in particular at ages 65 and more) and is especially pronounced for women. It can further be seen in tables 5.2 and 5.3 that the percentage of time spent with various forms of disability for men is very similar when calculated for Method 1 and 2. This is, however, not the case for women: Method 1 shows higher percentages in all age groups than Method 2; this difference is especially large in the highest age group (at age 75).

Sex differences in total life expectancy at various ages and for life expectancy free of disability are shown in table 5.4.

Table 5.4 Sex differences in total life expectancy and life expectancy free of disability by age, 1981-1985 (Difference between life expectancy of women and men divided by male life expectancy and expressed as a percentage)

Age	Total life expectancy	life exp. free of disability Method 1	life exp. free of disability Method 2
0	+ 9.1	- 2.3	+ 3.3
5	+ 9.5	- 2.8	+ 2.9
10	+ 10.2	- 3.1	+ 3.0
15	+ 11.0	- 3.5	+ 3.0
20	+ 11.5	- 3.3	+ 3.9
25	+ 12.8	- 3.3	+ 4.5
30	+ 14.0	- 3.2	+ 5.4
35	+ 15.5	- 2.7	+ 6.3
40	+ 17.5	- 3.0	+ 7.3
45	+ 19.9	- 2.9	+ 9.4
50	+ 22.9	- 3.2	+ 11.5
55	+ 26.3	- 5.3	+ 13.1
60	+ 29.7	- 5.8	+ 14.0
65	+ 32.5	- 12.6	+ 12.2
70	+ 33.5	- 14.8	+ 11.7
75	+ 31.8	- 17.6	+ 10.0

Sex differentials were calculated here as the difference in life expectancy between women and men and divided by male life expectancy. The result is expressed in the form of percentages. This table shows once more that at all ages differences in life expectancy free of disability between the sexes are much smaller than differences in total life expectancy. It can also be seen that Method 1 leads to somewhat different results than Method 2; this is in accordance with the findings from tables 5.1-3.

A comparison of the disability-free life expectancy in the Netherlands with Canada, France and England and Wales is made in table 5.5.

Table 5.5 Life expectancy without disability (at birth) by sex in four countries

Age and sex	Total life exp.	In insti-tutions	Long term dis-ability	Short-term limi-tations	Years with dis-ability	With-out dis-ability
MEN						
Canada (1978)	70.8	0.8	9.7	1.1	11.6	59.2
France (1982)	70.7	0.6	7.3	0.9	8.8	61.9
England, Wales (1985)	71.8	-	-	-	13.1	58.7
Netherlands Meth.1 (1981-1985)	72.8	1.0	8.9	3.5	13.3	59.5
Netherlands Meth.2 (1981-1985)	72.8	1.0	13.0		14.0	58.8
WOMEN						
Canada (1978)	78.3	1.5	12.2	1.8	15.5	62.8
France (1982)	78.9	1.1	9.7	0.9	11.7	67.2
England, Wales (1985)	77.7	-	-	-	16.2	61.5
Netherlands Meth.1 (1981-1985)	79.5	2.5	15.0	3.8	21.3	58.2
Netherlands Meth.2 (1981-1985)	79.5	2.5	16.3		18.9	60.7

Source: Canada: Wilkins en Adams, 1983
France: Robine, Brouard en Colvez, 1987;
England, Wales: Bebbington, 1988

The number of years spent with different forms of disability is in the Netherlands for men somewhat higher than in the other countries and is for Dutch women considerably higher. Table 5.5 also shows that the life expectancy free of disability for men is about the same in Canada, England and Wales and the Netherlands; for women this life expectancy is about the same in Canada, England, Wales and the Netherlands according to Method 2. The largest deviations are thus to be found in France for both men and women and in the Netherlands when Method 1 is used. In addition, one can see differences in values for the different forms of disability separately in the various countries. One can see, for instance, that in the Netherlands the number of years with short-term limitation of activity is larger than in Canada and France. Furthermore, the large number of years with long-term disability for Dutch women is noticeable.

5.3 DISCUSSION

The topic of life expectancy free of disability is for the following reason important. In the Netherlands and other industrialized countries life expectancy has increased considerably during the past century. This raises the question what the evolution has been with respect to morbidity and disability. Has the number of years spent with various forms of morbidity and disability diminished during this time or has the number of years with disability actually increased in the past century? If the first of these two cases applies, we can speak of a great success in the attainment of health policy objectives. If the second of these two cases applies, however, this success has both positive and negative aspects. When interpreting these results, one has namely to take into account that the objective of health policies is not so much or not only to increase life expectancy, but also to maximize the number of years spent in good health.

In this report we cannot provide answers to the basic questions raised above. Instead we will limit ourselves to the provision of information on life expectancy free of disability in the Netherlands in 1981-1985. Two approaches have been used. According to the first method a life expectancy free of disability for men of 59.5 years was found and for women 58.2 years. According to the second method the number of years free of disability for men was 58.8 years and for women 60.7 years.

In spite of the fact that in the case of women the two methods lead to somewhat different results, the conclusion can be drawn that the difference in life expectancy free of disability between women and men is much smaller than the difference in total life expectancy (this difference between women and men was 6.7 years in 1981-1985). There are two reasons which explain this phenomenon. The first is that the prevalence of two of the three types of disability in nearly all age groups is higher for women than for men. The second reason is a consequence of the higher female then male total life expectancy. The number of years lived by women at higher ages $_nL_x$ is in the Netherlands considerably larger than by men and precisely in the higher age groups is the difference in prevalence of disability between women and men the most pronounced.

Another conclusion which can be drawn concerns the somewhat different results achieved with the two methods for women. We have seen that cross-sectional sample surveys were an important source of information to determine disability. A characteristic of such surveys is that results achieved in the measurement of concepts such as disability are very much dependent on methods of questioning, operational definitions, phrasing of questions, etc. The different questions or set of questions on morbidity or disability can very well yield different outcomes, because the various questions can measure different aspects of these concepts. Therefore, it is certainly plausible for Methods 1 and 2 to lead to different results on levels of disability.

The Dutch findings were also compared with studies undertaken in other countries. The number of years free of disability was about the same in Canada, England and Wales and the Netherlands according to Method 2 for men and women. This result was obtained in spite of the fact that the operational definitions of disability were not similar in these studies. The number of years free of disability in France and the Netherlands according to Method 1 deviated from the values obtained in the other studies (in particular for women). This is probably to a considerable extent due to differences in operational definitions in methods of questioning etc. between France and the Netherlands (Method 1) on one hand, and Canada, England and Wales and the Netherlands (Method 2) on the other hand. More research is needed to clarify these findings and more efforts need to be made to increase the comparability of measurement of key concepts.

Finally, it is important to draw attention to the contribution of the concept of life expectancy free of disability to the study of sex differentials in health and mortality. We have already referred to the relevance of this concept for evaluation of the success of health policies in industrialized countries. Its particular advantage is the combination of mortality and morbidity/disability elements in a single summarizing measure. Because this measure combines mortality and morbidity/disability elements, it is also an important concept to be applied in the study of sex differentials in health and mortality.

REFERENCES

BEBBINGTON, A.C. The expectation of life without disability in England and Wales. Soc. Sci. Med. 27 (1988) 321-326.

COLVEZ, A. & M. BlANCHET. Potential gains in life expectancy free of disability: a tool for health planning. Intern. J. Epidemiol. 12 (1983) 224-229.

Netherlands Central Bureau of Statistics. Health Interview Survey 1981-1985. The Hague, Staatsuitgeverij, 1988.

ROBINE, J.M., BROUARD, N. & A. COLVEZ. Les indicateurs d'espérance de vie sans incapacité (EVSI). Des indicateurs globaux de l'état de santé des populations. Rev. Epidemiol. Santé Publique 35 (1987) 206-224.

SONSBEEK, J.L.A. Methodological aspects and content of the OECD-long-term indicator (in Dutch). Maandbericht Gezondheidsstatistiek 6 (1988) 4-17.

SULLIVAN, D.F. Disability components for an index of health. Washington, U.S. Government Printing Office, 1971a. (PHS Publication No. 1000 Series 2, No. 42)

SULLIVAN, D.F. A single index of mortality and morbidity. HSMH Health Reports 86 (1971b) 347-354.

VAN GINNEKEN, J., BANNENBERG, A. & A. DISSEVELT. Loss of Health as Result of a Number of Important Groups of Diseases in The Netherlands in 1981-1985: Methodological Aspects and Results (in Dutch). Internal report. Leiden and Voorburg, Netherlands Institute of Preventive Health Care and Netherlands Central Bureau of Statistics, 1989.

WILKENS, R. & O.B. ADAMS. Health expectancy in Canada, late 1970s: demographic, regional and social dimensions. Am. J. Public Health 73 (1983) 1073-1080.

NONLINEAR PARTIAL CANONICAL CORRELATION ANALYSIS OF HEALTH, SOCIAL CLASS AND WORK INDICATORS; IN COMPARISON WITH CUMULATIVE INCIDENCE RATIO'S

A. Bloemhoff*

* Department of Occupational Health Research, TNO Institute of Preventive Health Care, Leiden, The Netherlands

Abstract

The impact of social class on disability is unclear. On the one hand the disabled are characterised as older labourers with little education, who have done hard physical labour for a long time. On the other hand there is the teaching-profession with a high incidence of disability; the educational level of teachers is relatively high and they are under a mental rather than physical strain. As a part of the research programme on socio-economic differences in health (financed by the Ministry of WVC), the NIPG/TNO has initiated a study to analyse the relationships between social class and disability. The objectives of this study was to analyse the relationship between social class and incidence of disability, adjusted for the confounders age and working conditions. The data pertained to a historical cohortstudy by De Winter (1991), where 2791 male employees of four companies completed a questionnaire on person- and job-characteristics, perceived working conditions and health. During a follow-up period of five years, data became available on the incidence of disability.

Two different analysis methods were used. The first method consisted of estimating the nonlinear partial canonical correlation between disability and educational level, stratified by age, while adjusting for the covariates working conditions. The second method consisted of the multivariate estimation of independent Cumulative Incidence Ratio's (RR) of education level, age and working conditions. The results of both methods will be discussed and compared. The conclusion is that social class has a limited impact on disability (partial = 0.09), with age (RR=4.6) and work-strain (RR=2.8 and RR=1.8)

as significant confounders.

Keywords: *social class, disability, working conditions, nonlinear partial correlation analysis, cumulative incidence ratio.*

6.1 INTRODUCTION

Work disability is a major social and financial problem in the Netherlands. (Aarts & De Jong, 1991). Since the introduction of the General Disablement Benefit Act in 1967, the number of workers classified as work disabled has risen to almost 900,000 in 1990. One out of eleven potential workers (aged 15-65) in the Netherlands is unfit for work.

In Dutch literature work disabled are characterised as older labourers, with little education, who have done hard physical labour for a long time. However there is the teaching profession with a high incidence of work-disability. Teachers on the average don't belong to low social classes, have a relatively high education level and are under a mental rather than a physical strain.

At the request of the Programme Committee on Socio-Economic Health Differences of the Ministry of Welfare, Health and Cultural Affairs, a study was initiated to clarify the relationships between social class and work disability taking into account age and working conditions (Bloemhoff & De Winter, 1990).

The aim of this paper is to draw a comparison between two methods of estimating the relationships between social class and incidence of work disability, adjusted for age and working conditions.

6.2 METHOD

6.2.1 Study design

The study can be characterised as a longitudinal study. The data pertain to a historical cohort study by De Winter (1991): 2759 male employees of four large companies completed a questionnaire on person- and job-characteristics, perceived working-conditions and health. During a follow-up period of five years, data came available on the incidence of work disability. The study-population can be characterised as a static cohort. After completing the questionnaire, the men either left the job because of work-disability during the follow-up period or they remained in the job during this period.

The following variables were included:

- the variable 'education level', as an indicator of social class; expressed by four levels, namely 1) primary education only (n=772), 2) elementary general secondary (LAVO) or junior vocational education (LBO) (n=1085), 3) general secondary education (MAVO, HAVO, VWO) or senior secondary vocational education (MBO) (n=758) and 4) higher vocational education (HBO) or university level (n=144);
- the variable 'incidence of work-disability during the follow-up period'; 1) work disabled (n=278) and 2) not work disabled (n=2513);
- the variable 'age', divided into two categories, namely 1) younger than 50 years of age (n=2173) and 2) 50 years and older (n=618);
- the variable 'work strain' as an indicator of working conditions; 15 questions on physical and mental strain and the effects of strain in terms of health are transformed to one indicator variable with three levels: 1) few complaints (n=765), 2) average complaints (1603) and 3) many complaints (391); four other indicators of working conditions will not be discussed in this paper.

6.2.2 Statistical analysis

6.2.2.1 Nonlinear partial canonical correlation analysis

To estimate the relationships between education level and work disability, adjusted for age and working-conditions the following method was used: nonlinear partial canonical correlation analysis (Van de Geer, 1988). In general, for estimating the relationship between two variables correlation analysis is used. By using **canonical** correlation analysis, it is possible to estimate correlations between sets of variables. Just like the ordinary type of correlation analysis, it is possible with canonical correlation analysis to adjust for some covariates by partialisation. Because all variables used in this study are categorical, at a nominal or an ordinal measurement level, the **nonlinear** partial canonical correlation analysis was used by means of the statistical programme CANALS (Van de Burg, 1985)

6.2.2.2 Cumulative incidence ratio estimation by logistic regression analysis

To estimate the independent work disability risk of education level, age and working conditions a second method was used: estimating the cumulative incidence ratio's by logistic regression analysis (Rothman, 1986). The cumulative incidence is calculated by dividing the total number of new work-disabled during the observation period by the total number of people at risk at the beginning of the observation period and it expresses the work disability risk. The cumulative incidence ratio or Risk Ratio (RR) represents the work disability risk of an exposed group compared to the work disability risk of a non-exposed group. Logistic regression analysis is a multivariate method to estimate independent Odds Ratios (OR). When the incidence is small (10% or less) OR are good estimators of Risk Ratio's (Clogg & Eliason, 1988).

6.3 RESULTS

The results presented here will be restricted, because of the aim of this paper. The extended results are written elsewhere (Bloemhoff & De Winter, 1990). The unadjusted estimated nonlinear canonical correlation-coefficient for work disability and education level is equal to 0.17. The estimated nonlinear partial canonical correlation-coefficient for work disability and education level, adjusted for age and work strain is 0.08. The relationship between work disability and education level, adjusted for age and work strain is significant, but weak. So education level seems to have a significant but limited impact on work disability. The estimated adjusted correlation-coefficient is two times smaller than the non-adjusted one. Therefore we know that age and work strain have some impact on work-disability.

The results of the **second method** are presented in table 6.1. First the estimated Risk Ratio of education level with the 95% confidence interval is presented. Only the estimated RR of education level 1 is significantly different from 1. The other results in table 6.1 are the estimated independent Risk Ratio's of education level, age and work strain with their 95% confidence-interval. These results indicate that only the Risk Ratio of education level 1 is significant higher than 1, when adjusted for age and work strain. In other words, regardless of age and number of complaints concerning work strain, the work disability risk of men with only primary school is two times that of men with university education. The work disability risk of men with more then primary school education is not significantly different of the risk of university educated men. So education level has some impact on work disability, looking at these results. Table 6.1 shows also significantly high independent work disability risks for men older then 50 years (RR=4.6) and for men with many or an average number of complaints with regard to work strain (RR=2.8 and RR=1.8).

Table 6.1 Estimated risks ratio's (RR) of work disability by education level, age and work strain, with associated 95% confidence intervals (CI) (n=2759)

	no adjustment		adjustment for 'age' and 'work strain'	
	RR	95% CI	RR	95% CI
Education level (n)				
primary school (772)	2,5	(1,6-3,9)	2,0	(1,2-3,2)
LAVO/LBO (1085)	1,6	(1,0-2,6)	1,7	(1,0-2,7)
MAVO/HAVO/VWO/MBO (758)	1,2	(0,8-2,0)	1,5	(0,9-2,5)
HBO/University (144)	1,0		1,0	
Age (n)				
50-64 year (611)			4,6	(3,8-5,4)
15-49 year (2148)			1,0	
Work strain (n)				
many complaints (765)			2,8	(2,1-3,7)
average number of complaints (1603)			1,8	(1,4-2,3)
little complaints (391)			1,0	

6.4 DISCUSSION

Both methods will be compared and discussed by examining a few important qualities of both methods, namely the type of data that are required, the quality and contents of the output-information and the results.

Type of data

Both methods are developed for categorical data, so they both are suited for the variables used in this study. The number of variables and/or categories that can be used is limited with logistic regression analysis, because empty cells in the contingency tables are not allowed and missing data cannot yet be handled with ease. With canonical correlation analysis there are no restrictions regarding the number of variables and categories.

Quality of the output-information

The calculated canonical correlations can be considered as fairly good estimators of the real correlations. Logistic regression analysis doesn't estimate Risk Ratio's but Odds Ratio's. In this study the incidence is 10%, so as indicated before in that case the estimated OR are estimators of Risk Ratio's.

Contents of the output information

The estimated partial correlation-coefficient gives an overall impression of the relationship between work disability and education level. Comparing the adjusted and the not-adjusted correlation-coefficients gives information about the overall impact of the covariates. So the output is just one number and simple to interpret, and so is the final conclusion.

Estimating independent Risk Ratio's by logistic regression gives information about the relationship between work disability and education level for each category and it gives at category-level information about the impact of the covariates. The output needs some transformations, before the results can be interpreted as Odds Ratios or Risk Ratios and the interpretations and conclusions are more detailed but not always simple.

Results

The estimated nonlinear partial canonical correlation coefficient indicates that the overall association between work disability and education level is significant, but weak. By adjusting for age and work strain the correlation coefficient for work disability and education level decreases considerably. Therefore age and work strain do have an impact on work disability.

The estimated independent Risk Ratio's indicate that the work disability risk of education level one (only primary school) is significantly, but moderately increased compared to that of education level four (university education). The estimated work disability risks of level two and three are not significantly different from that of level four. The estimated independent Risk Ratio's of age and work strain are significantly increased and they show a relatively strong association with work disability.

The comparison of both methods leads to the following *conclusion*: The results of both methods confirm each other but both give different specific information on the relationship between education level, work disability, age and work strain.

REFERENCES

AARTS, L.J.M & P.R. DE JONG. Economic aspects of disability behavior. Rotterdam, Erasmus University, 1991. (Dissertation)

BLOEMHOFF, A. & C.R. DE WINTER. De invloed van sociaal economische status op arbeidsongeschiktheid: een longitudinale analyse. Leiden, NIPG-TNO, 1990.

BURG, E. VAN DER. CANALS. Leiden, University of Leiden, Department of Data Theory, 1985.

CLOGG, C.C. & S.R. ELIASON. Some common problems in log-linear analysis. In: J. Scott Long (ed) Common problems/proper solutions: avoiding error in quantitative research. Newbury Park etc., Sage, 1988, pp. 226-257.

GEER, J.P. VAN DE. Analyse van kategorische gegevens. Deventer, Van Loghum Slaterus, 1988.

ROTHMAN, K.J. Modern epidemiology. Boston/Toronto, Little, Brown & Co., 1986.

MIXED-LONGITUDINAL STUDY DESIGN IN MONITORING THE PREVALENCE OF DENTAL CARIES

J.E.F.M. Frencken*, M.A. van 't Hof**

* Department of Children and Health, TNO Institute of Preventive Health Care, Leiden, The Netherlands

** Department of Medical Statistics, University of Nijmegen, The Netherlands

Abstract

In studying developmental processes, variables other than chronological age must often be taken into account if differences in the developmental patterns of the group under study are to be characterized. In particular, in addition to age, appropriate models for developmental studies may have to incorporate cohort and period effects (Van 't Hof et al., 1976, Palmore, 1978)

The more traditional designs keep one of these time components constant and this results in the confounding of the two factors remaining. In this way age effects are confounded with cohort effects in the cross-sectional design and with period effects in the pure longitudinal design, while cohort and period effects are confounded in the time-lag design. The mixed-longitudinal design was developed to overcome the difficulties inherent in each of the more traditional approaches. This design however, does not completely avoid the problem of confounding, although it does provide a structure in which isolation of the contribution of age, period and cohort effects can be achieved (Van 't Hof et al., 1976). The Age-Period-Cohort procedure intends to separate the effects of the three time-components age, period and cohort. However, for confirmation and/or adjustment, external (prior) information must always be taken into account (Kleinbaum et al., 1982, pp. 130-134).

The mixed-longitudinal design and A.P.C.-procedure (in this study using analysis of covariance) where used to detect a possible upward trend of dental caries experience in a child population between 1984 and 1988. It resulted in a strongly significant age

effect (which served as prior information, as caries is considered an age related disease in children), a period effect (change in the application of the diagnostic criteria over the 4 year period) and a cohort effect (decrease in the caries experience).

Key words: *development, mixed-longitudinal design, age-period-cohort analysis, dental caries.*

7.1 INTRODUCTION

The prevalence of dental caries in children in The Netherlands and in other industrialized countries has declined steeply over the last, let's say, 15 years (Kalsbeek & Verrips, 1990). Caries in children is an age related disease. The number of carious cavities and/or fillings continuously increases and it tends to do so until it decreases artificially by extraction. Because of its age dependency, caries fits into statistical models designed to study developmental processes. Such a process is the study of a possible increase or decrease in the prevalence and severity of caries over time. Monitoring the prevalence of caries is, therefore, a very important activity. It is important for the formulation of preventive dental strategies and for the formulation of dental health policies in other areas like education.
The different time effects, which have an important impact on the study of developmental processes, are discussed underneath.

7.2 AGE, PERIOD, COHORT AND TEST EFFECTS IN DEVELOPMENTAL STUDIES

In studying developmental processes it is important to incorporate in the model, besides the chronological age, also cohort, period and test effects (Van 't Hof et al, 1976).

1. **Age effect** is defined as the net contribution that age has on the phenomena studied. Age effects (or, in the case of caries, accumulation of lesions over a given period) are in fact what developmental studies are usually designed for to measure.
2. **Period effect** originates from temporary circumstances affecting the measurements or observations being made. Studies concentrating on period effects are often designed to evaluate effects of public health programmes or information campaigns. Usually, period effects occur simply due to decalibration during the progress of the study.
3. **Cohort effect** stems from basic underlying differences in the patterns of diseases between groups of individuals born at different times, having thus a different life history. Such differences over a long period of time are known as secular trend.

Cohort effects make it difficult to generalize the results and are the primary cause of differences between average disease curves derived from cross-sectional and longitudinal studies.

Cohort and period effects are often considered disturbing factors.

Another effect which should be considered when studying developmental processes is a test effect.

- **Test effect** may be due to frequent contacts with the investigators or to the repetition of the same measurements or observations. Test effects may either decrease or increase the level of measurements over time. In some cases, the measurement accuracy is improved as the subject and the investigator become familiar with one another. In other cases, repetition impairs motivation and the precision of measurements at a later stage decreases. Yet another kind of test effect (study genetic effects) may be caused if the subjects are receiving better medical attention, as a reward for their participation, or become more aware of the responsibility with respect to their own health. Such effects are a disadvantage affecting the meaning of results of a study, but they are obviously to the advantage of the participants. Test effects may be estimated using a control group consisting of individuals of the same cohort, who are measured at the same time, but differ in study experience.

7.3 STUDY DESIGNS FOR AGE-RELATED PHENOMENA

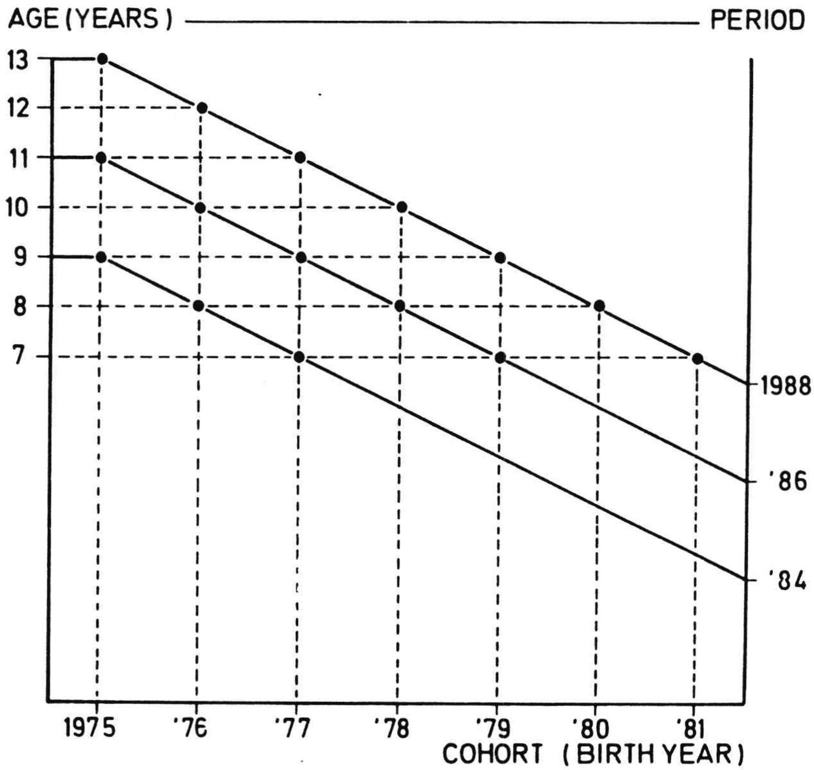
If the objective of the study is to monitor changes in developmental processes over time e.g. growth of children or dental caries, one is primarily interested in cohort effects. This then leads to the question: 'what type of study design is appropriate'? Let's consider the possible designs and discuss their appropriateness.

In fact the most appropriate design for the study of cohort effects is a time-lag design. Such a design includes the study of a specific age group during several generations. However, this bifactorial design confounds cohort with period effects. Comparable confounders are also present in both other bifactorial designs; cross-sectional and pure longitudinal design. In other words, in the cross-sectional design age effects are confounded with cohort effects and in a pure longitudinal design age effects are confounded with period effects. The difficulties encountered in these more traditional designs can be overcome if a mixed-longitudinal design is chosen. Mixed-longitudinal designs do not completely avoid the problem of confounding, but they can be used to provide a structure in which the contributions of age, cohort and period effects can be isolated. This separation of the three time-components age, period and cohort can be attained by using the so called: Age-Period-Cohort analysis. The A.P.C.-analysis is not an analysis technique, but it is an analysis procedure by which the contribution of age, period and cohort to the developmental variable under study. It is important, however, to realize that for the isolation of the time-effects external information must always be taken into account (Van 't Hof et al, 1976; Kleinbaum et al, 1982). An example of required prior information is presented when we discuss the application of the A.P.C.-analysis on the dental data of a 4-year mixed-longitudinal study.

7.4 A MIXED-LONGITUDINAL CARIES STUDY IN CHILDREN BETWEEN 1984 AND 1988

The mixed-longitudinal design of the study described here, was based on the age of the children (7-13 years) and a measurement interval of two years (Figure 7.1). There were seven birth cohorts (1975 up to 1981) of which the first three were examined repeatedly in 1984, 1986 and 1988, and the following two in 1986 and 1988. The number of children examined was 722 (1984), 833 (1986) and 989 (1988). There were three geographical areas distinguished; a fluoridated and a non-fluoridated rural, and an urban area (Frencken et al, 1989). The nul-hypothesis was: there is no increase in the prevalence of dental caries in these child generations between 1984 and 1988. In other words, there is no positive cohort effect.

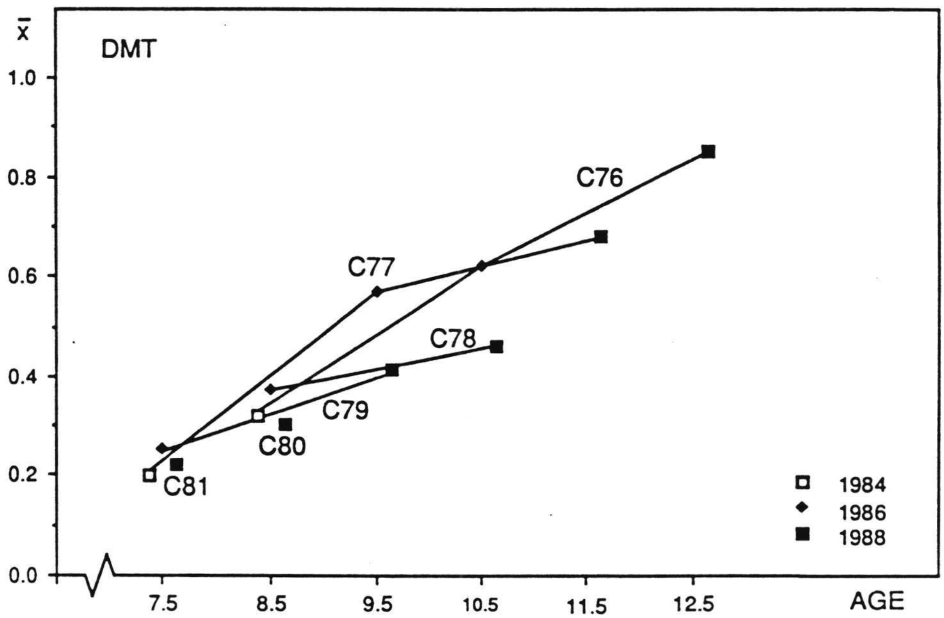
Figure 7.1 Mixed-longitudinal design of the present study



Statistical methods: The dependent caries variable was the mean DMT-score. 'D' stands for Decayed teeth, 'M' for Missing teeth due to caries and 'T' for Teeth. Fillings were not observed. The side cohort 1975 was not included in the analysis. The three different study areas were not specified separately but were combined in a weighted average DMT-score using a constant proportion for each year of measurement: 0.5 (urban), 0.3 (rural) and 0.2 (fluoridated rural). These aggregated data were sufficiently normally distributed to justify the use of analysis of co-variance, with age as co-variate, to test the effects of the three time-components on the caries variable.

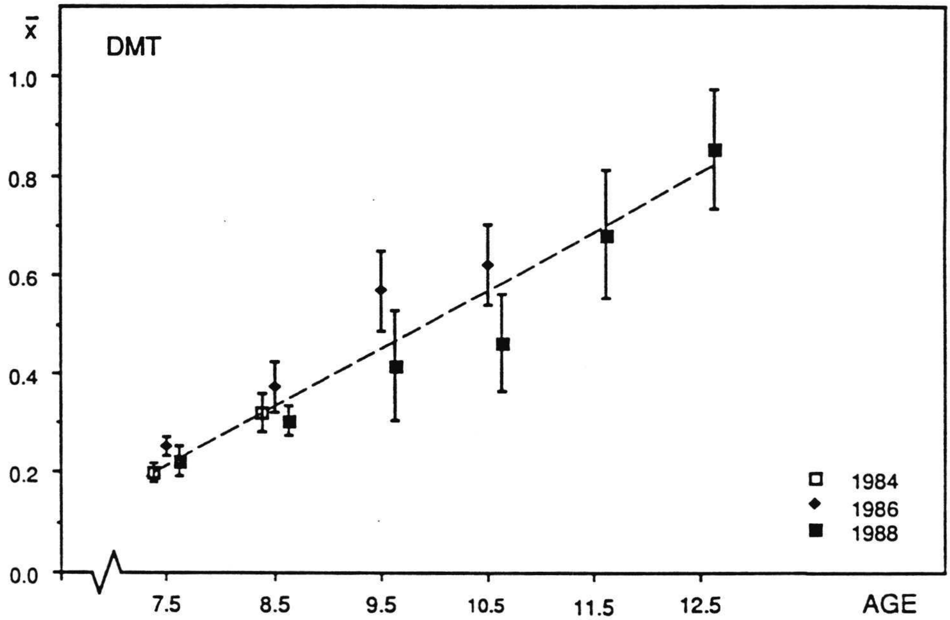
Results: The application of an A.P.C.-analysis on the caries data is only possible if additional assumptions are entered in the model; one of them being an increasing age curve. This is justified by the fact that caries in children usually increases with increasing age. Figure 7.2 suggests that a linear age relationship is not unrealistic. However, in this design the assumption of a linear age relationship was not sufficient to separate the three time effects. Also adding the assumption of a linear cohort relationship was not sufficient. Therefore, the A.P.C.-analysis was carried out using the prior information: (a) linearity with age; (b) cohort gap between cohorts 1976-1977 and 1978-1981; (c) no assumptions about period effects. Eventually, the hypothesis tested was: there is no difference in caries prevalence in the earlier and the later birth cohorts.

Figure 7.2 Age differences when absence of cohort and period effects is assumed



A suitable hypothesis testing was found in an analysis of co-variance on the data of Figure 7.2. The results of this analysis revealed a significant period ($p = 0.01$) and cohort effect ($p = 0.02$), and a strongly significant age effect ($p < 0.001$). Lower caries scores were found in 1984 than in the two other years of investigation and the magnitude of this period effect was estimated to be 0.13 DMT. Cohorts 1976-1977 had higher caries scores than cohorts 1978-1981 and the magnitude of the cohort effect was estimated to be 0.12 DMT. An impression of the cohort effects, neglecting period effects, is given in Figure 7.3. The figure shows a systematic trend of generation developments; i.e. cohorts 1976-1977 are situated above cohorts 1978-1979.

Figure 7.3 Illustration of cohort differences when absence of period effects is assumed



7.5 DISCUSSION AND CONCLUSION

Usually, changes in caries prevalence are monitored by comparing results of cross-sectional studies, carried out at different years of investigation. In doing this, generation differences (cohort effects) and measurement problems or results of intervention programmes (period effects) cannot be distinguished. Therefore, comparison of cross-sectional studies are bound to lead to invalid conclusions.

The present study has demonstrated the advantages of the mixed-longitudinal approach in combination with the A.P.C.-analysis. The small period effect, most probably indicates a disagreement within the group of observers over time. The apparent presence of a decreasing cohort effect means that the prevalence of caries had not increased in these child populations between 1984 and 1988.

REFERENCES

FRENCKEN, J.E., VAN 'T HOF, M.A., TRUIN, G.J., LEMBARITI, B.S. & K.G. KÖNIG. Cohort effects in the prevalence of caries in child populations in Tanzania. J. Dental Res. 68 (1989) 1777-1780.

KALSBEEK, H. & G.H.W. VERRIPS. Dental caries prevalence and the use of fluorides in different European countries. J. Dental Res. 69 (1990) Special issue 728-732.

KLEINBAUM, D.G., KUPPER, L.L. & H. MORGENSTEIN. Epidemiologic research, principles and quantitative methods. New York, Van Nostrand Reinhold, 1982.

HOF, M.A. VAN 'T, PRAHL-ANDERSEN, B. & C.J. KOWALSKI. A model for the study of developmental processes in dental research. J. Dental Res. 55 (1976) 359-366.

NONLINEAR MULTIVARIATE ANALYSES OF PARENTAL SAFETY BEHAVIOUR

J.J. Radder*, C.C.J.H. Bijleveld*, E. Wortel**

* Department of Statistics and Computer Science, TNO Institute of Preventive Health Care, Leiden, The Netherlands

** Department of Statistics and Computer Science, TNO Institute of Preventive Health Care (Presently at the Department of Psychometrics, Faculty of Social Sciences, Leiden University, The Netherlands)

*** Department of Public Health and Prevention, TNO Institute of Preventive Health Care, Leiden, The Netherlands

Abstract

We analyze the relation between parents' reported safety behaviour and six determinants of that behaviour (Wortel, Ooijendijk & Stompedissel, 1988). As the determinants were mixed dichotomous and nominal variables, and as the reported safety behaviour had been recorded on ordinal measurement level, and as no obvious analysis method for this type of problem stands out, several nonlinear analyses were performed to investigate the relation between safety behaviour and its determinants. Measurements had been obtained from 1129 parents with preschool children for several safety measures; we chose to analyze the safety behaviour 'removing teapot/coffeepot', in relation to the confounding variable 'education' and six determinants. The safety behaviour variable had three categories: 'unsafe behaviour', 'relatively safe behaviour' and 'safe behaviour'. A total of 754 parents had no missing measurements, these were selected for further analysis. In order to explore the supposed confounding effect of education, we performed multiple correspondence analysis (Greenacre, 1984). After adding an interactively coded combination of education and safety behaviour, the analysis showed that such an interaction between education and safety behaviour was absent. In fact, education played

little or no role at all.
Those who themselves, or whose partner thought safety measures unnecessary, exhibited unsafe behaviour; those who thought that they would not succeed, who considered the child too old for safety measures, and, to a lesser extent, who thought safety measures wouldn't help and would be an inconvenience, exhibited relatively safe behaviour.
For investigating the relative importance of the six determinants and the confounder education in predicting safety behaviour, we performed nonlinear multiple regression analysis (Gifi, 1981) on the same variables, thereby relating safety behaviour on the one hand to the behaviour determinants on the other hand. It turned out that 'succeed', 'necessary' and 'partner' were identified as most important. Again, education had no relation whatsoever with safety behaviour.
Nonlinear discriminant analysis (Van der Burg, 1985, pp. 82-85) served to identify those variables on which the safety behaviour categories could be distinguished. On the first discriminant dimension, safe behaviour was contrasted with less safe and unsafe behaviour; variables that contributed most to this contrast were again 'succeed', 'necessary' and 'partner'. On the second discriminant dimension, relatively safe behaviour was contrasted with unsafe behaviour; the variable that contributed most to this contrast was partner: only the unsafe reported that they think their partner does not find the safety measures necessary. Thus, the various nonlinear analyses supported the same global conclusions; the various analyses highlighted different interrelationships between the variables.

Key words: *parental safety behaviour, multiple correspondence analysis, nonlinear regression.*

8.1 BACKGROUND

What we like to present here is a comparison of three different non-linear analyses on the same dataset. The data were collected in a research project on behaviour of parents to prevent household accidents of their children and the determinants of such behaviour. The research project took place in the city of Nijmegen and involved mothers of children in the age of 0-4 years. Part of the research question was whether safety behaviour was related to a number of determinants behaviour. School-education of the mother was considered a possible confounder.

8.2 MATERIAL AND METHODS

Safety behaviour was defined in this study as taking precautions to make the environment of the child safe and was scored in the categories safe, relatively safe and unsafe. Education was scored as lower, middle and higher.
Considering the nature of the variables involved (mixed ordinal, nominal and dichotomous) numerical analyses that assume interval scored variables were not considered feasible. Since it was not obvious what the best method would be to analyze the data we chose to perform three different analyses and compare the results.

We chose to analyze the measure 'to keep tea/coffeepot out of reach'. The determinants considered and their categories were:

necessary : do you think it necessary to take this measure at this age of your child with the categories: necessary, not necessary because the child is too young, not necessary because the child is too old, not necessary indifferent of the child's age.

help : do you think this measure helps to avoid burning injuries.

inconvenient : how inconvenient do you think it is to take this measure.

succeed : do you think you will succeed in administering this measure

partner : do you think your partner thinks this measure has to be taken.

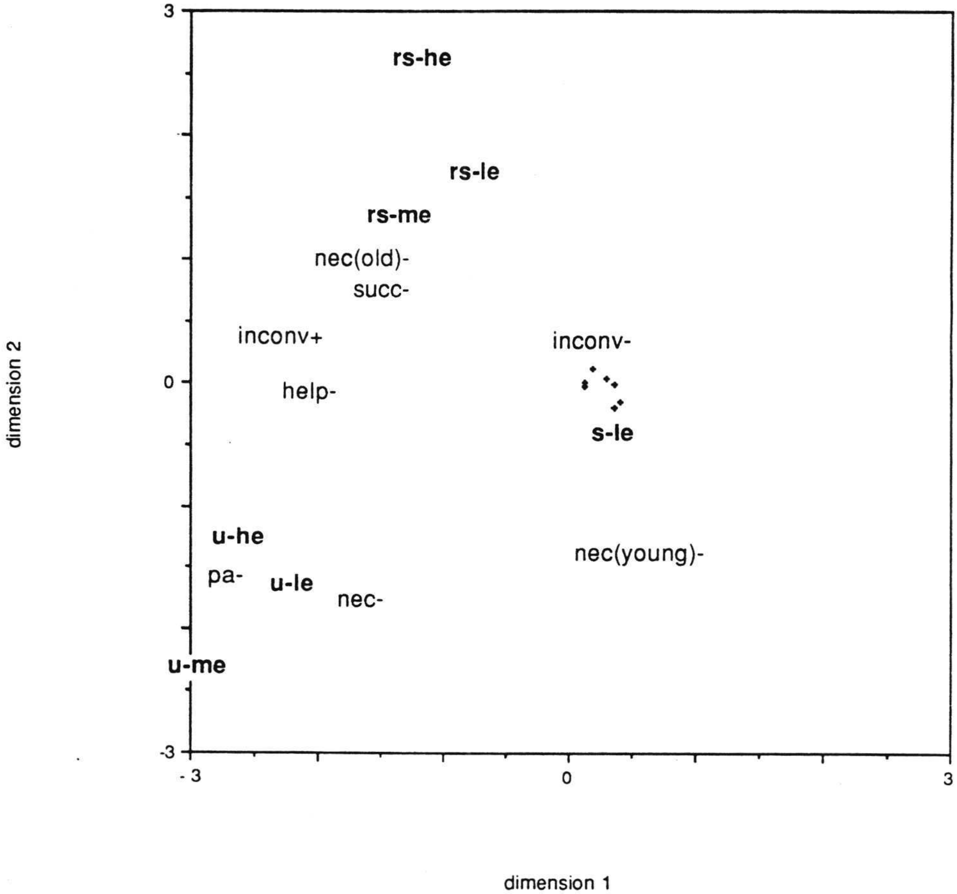
The last four with the categories no and yes.

8.3 RESULTS

We performed three different analyses, namely a Multiple Correspondence Analysis (MCA), which was earlier described by Andries et al in chapter two, and two variations of non-linear canonical analysis (CANALS), which was discussed by Bloemhoff in chapter six.

First we will discuss the results of the Multiple Correspondence Analysis. This analysis was performed on an interactive coding of behaviour and education, and the determinants. Interactive coding means that a new variable was constructed from the nine possible combinations of the categories of behaviour and education. So the categories of this new variable consisted of 'higher education-safe behaviour, higher education-relative safe behaviour, etc. down to lower education-unsafe behaviour. The results of this analysis are shown in figure 8.1.

Figure 8.1 Removing teapot/coffeepot (multiple correspondence analysis)



This figure shows the nine combinations of the interactive coding in bold face, for instance **rs-le** (Relatively Safe-Lower Education), **u-me** (Unsafe-Middle Education) and the determinants, for instance help- (does not help; the minuses denote the no- and the pluses the yes answers), succ- (will not succeed) and nec(old)- (not necessary, too old). The cluster of points in the middle are the most favourable categories of the determinants, which are very close together. What can be seen from this picture in the first place is the lack of interaction between behaviour and education: the three levels of education cluster with each category of behaviour. The interpretation is that education does not play any role when looking at different types of behaviour.

As to the determinants, the outcome is that those who think the measure not necessary, or think their partner does not think the measure needs to be taken show unsafe behaviour. Again: irrespective of education. Those who think they will not succeed in administering the measure, who think their child is already too old, and, to a lesser extent, who think this safety measure would not help and would be inconvenient, exhibited relatively safe behaviour. Please note the distinct place of not necessary-too young.

Next we analyzed the data with the CANALS computerprogramme with safety behaviour as the only variable in the first set and education and the determinants in the second set. This may be considered a non-linear multiple regression analysis. The canonical correlation was 0.79. The correlations of the variables of the second set with their canonical variate are listed below:

education	0.00
help	0.38
inconvenient	-0.35
necessary	0.69
succeed	0.89
partner	0.63

The projections of the category quantifications of necessary were:

necessary	
yes	0.22
no, too young	0.52
no, too old	-1.32
no, never	-1.30

From this data it is evident that necessary, succeed and partner were closest related to behaviour and, to a lesser extent, help and inconvenient. Education showed no relation at all. Looking at the signs of the correlations the direction of the relationships is: safe behaviour comes with thinking one will succeed, partner thinks it important etc. Necessary deserves special attention. From the category quantifications it can be seen that those who thought it necessary and, what is remarkable, those who thought it not necessary while their child was still to young, showed safe behaviour.

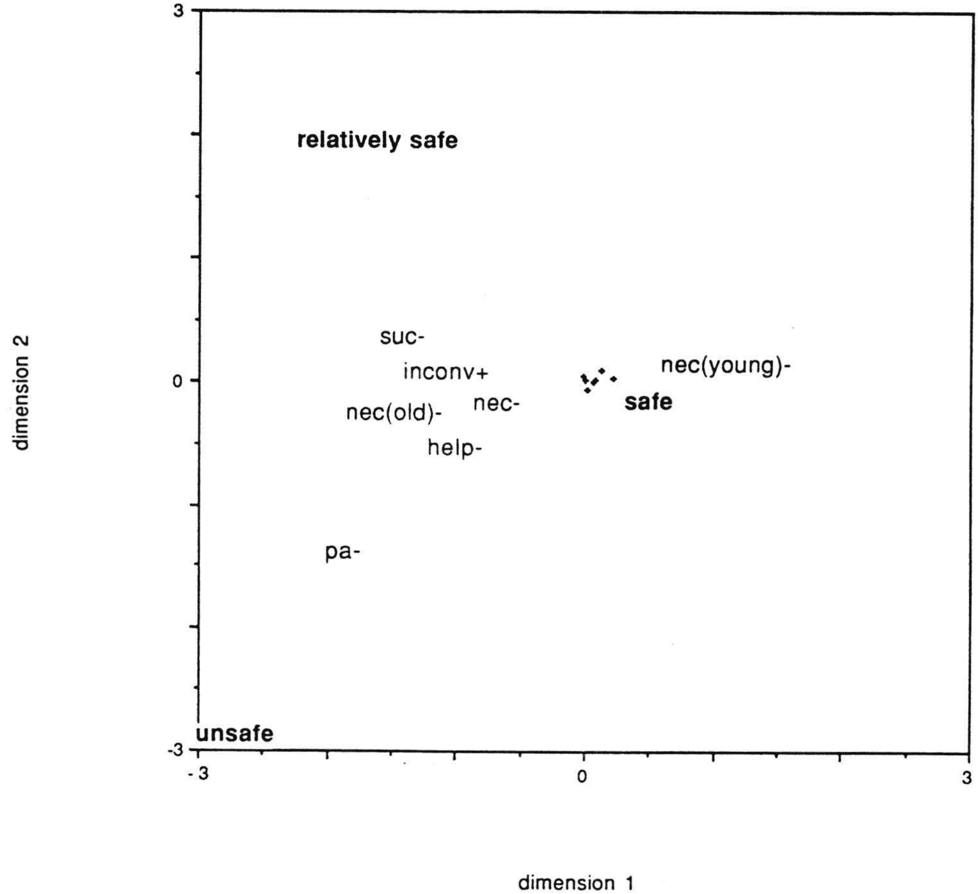
The last analysis we performed was non-linear discriminant analysis. This was done with the CANALS program by recoding behaviour into three dummy variables and then putting these in one set and again the other variables in the second set. This time we chose a solution with two dimensions.
The results of this analysis are presented in table 8.1.

Table 8.1 Results of the non-linear discriminant analysis

	correlations with canonical variate		can. correlations	
			0.79	0.49
dummy1 (unsafe)	-0.70	-0.71		
dummy2 (rel.safe)	-0.64	0.77		
dummy3 (safe)	0.97	0.23		
education	-0.01	0.10		
help	0.38	0.15		
inconvenient	-0.35	-0.02		
necessary	-0.69	-0.20		
succeed	0.89	-0.38		
partner	0.63	0.74		

As an illustration of this analysis we made a figure in witch are plotted the mean scores of the categories on the canonical variates (see figure 8.2).

Figure 8.2 Removing teapot/coffeepot (nonlinear discriminant analysis)



Nearly all 'not'-categories are on the left side, between unsafe and relatively safe. Only 'partner' lies rather close to unsafe. All positive answers (points) are close to safe. The correlations reveal that the contrast on the first dimension was between safe and less safe (relatively safe plus unsafe). This contrast is mainly due to the variables that also did have a dominant role in the previous analysis: necessary, succeed and partner. On the second dimension unsafe was contrasted with relatively safe. This is mainly determined by the variable partner: the unsafe more often think their partner doesn't think this behaviour important.

8.4 DISCUSSION

In this final part of the paper we will discuss what the contribution was of the three analyses performed to the answer of the research question.

First of all, all three analyses showed that for this safety measure education played no role. Also there obviously exists a strong relationship in the data between safe behaviour and 'positive thinking' about this measure. The unfavourable answers were somewhat in between relatively safe and unsafe. Only the opinion about how the partner thinks is more related to unsafe behaviour.

The contrast between relatively safe and unsafe behaviour appears on the second dimension. Also the distinct place of 'not necessary, child still too young' shows only in the two dimension solutions.

The multiple regression and discriminant analyses only showed that education had no relation with safety behaviour. The Multiple Correspondence Analysis provided additional information on the interaction between education and safety behaviour. From the Multiple Correspondence Analysis we suspected that 'partner does not think the measure should be taken' and 'unnecessary' pertained specifically to the unsafe; discriminant analysis showed that in fact partner was the only variable on which the unsafe distinguished themselves.

REFERENCES

GREENACRE, M.J. Theory and applications of correspondence analysis. London, Academic Press, 1984.

WORTEL, E., W.T.M. OOIJENDIJK & I. STOMPEDISSEL Preventie van privé-ongevallen bij 0-4 jarigen [Prevention of preschool children's accidents at home]. Leiden, NIPG-TNO, 1988.

BURG, VAN DER, E. CANALS. Leiden, Dept. of Datatheory FSW, 1985.

GRAPHICAL REFERENCE-PROFILES: A METHOD FOR ANALYSIS AND PRESENTATION OF QUESTIONNAIRE DATA

F.H.G. Marcelissen*, D.J. van Putten**, J.J. Radder***

* Department of Occupational Health Care, TNO Institute of Preventive Health Care (Presently at Digital Equipment Enterprise B.V., Vonderweg 11, 6500 MD, Eindhoven, The Netherlands).

** Department of Occupational Health Care, TNO Institute of Preventive Health Care, Leiden, The Netherlands

*** Department of Statistics and Computer Science, TNO Institute of Preventive Health Care, Leiden, The Netherlands

Abstract

Many Occupational Health Centres in the Netherlands organize "Periodic Medical and Workplace Surveys" (PMWS). During such a survey a self administered questionnaire is used for obtaining insight into the health problems and working conditions as perceived by the employees. Together with other available information on health and work (eg. biometric data and workplace survey) the questionnaire provides the occupational physician and industrial hygienist with basic information about health and work.

Instruments and methods for the analysis, interpretation and presentation of large numbers of items and subscales from questionnaires are not readily available for every day (occupational) practice. We developed the Graphical Reference-Profile (GRP) method, which has the following advantages:

- *it gives a comprehensive visual presentation of series of items and subscales;*
- *the difference between scores of different studygroups is shown;*
- *confidence intervals are shown to facilitate inferences;*
- *confounders can be corrected for by means of indirect or direct standardization;*

Although the questionnaire is developed for use in Occupational Health Care settings, it can also be applied to other situations using other questionnaires.

9.1 WHY GRAPHICAL REFERENCE-PROFILES ?

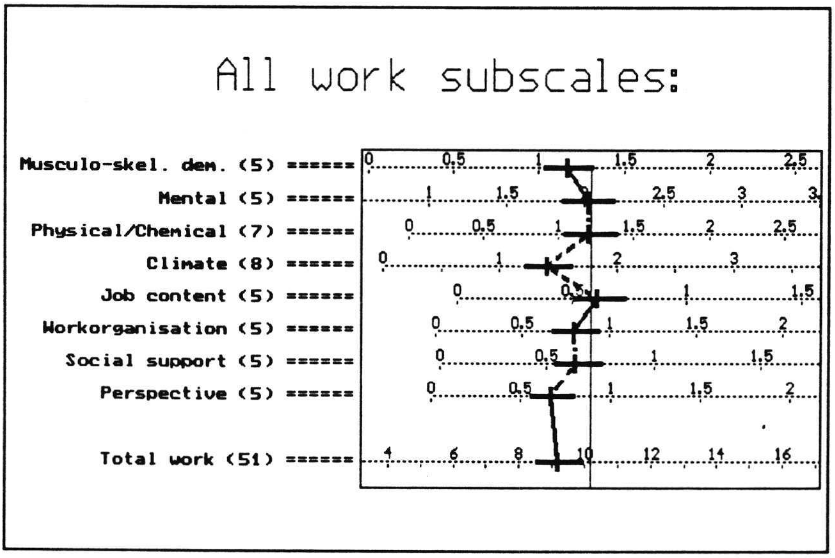
Graphical Reference-Profiles (GRP) form a method that makes it possible to analyze and present survey-results (amongst others) in a simple way.
Graphical Reference-Profiles were originally developed for the Periodic Medical and Workplace Survey (PMWS) which is administered by most Dutch occupational health service centres. Here yearly many employees fill in a questionnaire that asks about their health, and about problems they experience at their work and working conditions. Together with data from workplace-surveys and medical examinations these data enable the occupational physician to advice the employee about his personal health and to advice the employer on riskfactors at work. This last type of advice about groups of employees asks for the results of the company or department to be compared with the results of other, more or less comparable companies since absolute norms for acceptable questionnaire-scores are not available.
The method of GRP has been developed because there was a need for a method to present and interpret the results of the PMWS-examination in a simple, statistically meaningful way. Most important part of the method of reference-profiles is to compare the results from one or more groups of interest (studygroups) with a referencegroup.
For the construction of GRP's a menu-driven computer program with manual is developed. At this moment a beta-version this program (V-PROF) is available and research on the statistical and methodological merits are in progress.

9.2 WHAT IS A GRAPHICAL REFERENCE-PROFILE ?

A GRP shows in a vertical way the answers of a studygroup on a series of variables. An example of the output is given in figure 9.1. On the left side are the names of the subscales and within brackets the number of questions the subscale is constructed from. The midpoint of the scale is the mean of a referencegroup. The scales are organised one under the other in such a way that all means of the referencegroup can be connected by a straight vertical line. The left- and right margin of each scale represent one standard deviation from the mean of the referencegroup of that scale. Thus, when drawn into the

figure, the scores of a studygroup on different variables can be compared directly with each other. The more extreme a studygroup scores on certain variables, when compared with the referencegroup, the greater the distance of that score from the midpoint. The most extreme scores of a group will be near the margins.
Around the means of the examined group a 95% confidence interval is drawn. The width of this confidence interval is a measure of the precision of the score of the examined group. When a group consists of many employees the confidence interval will be very small. The score found in the group will be very close to the "true" score for the population the group was drawn from. When the studygroup is small, the confidence bands will be wide which means that the measured score may be far away from the "true" score. In this case it is not possible to make meaningful inferences about the means of the studygroup.

Figure 9.1 Subscale-profile of company X (curved line) and a referencegroup with the 95% confidence interval for the studygroup



Whether or not confidence intervals for the referencegroup are of interest depends on the size of the referencegroup. If it is large (and that will be the case in most instances) that is of little use since the confidence bands will then be very small. (The program has an option whether or not to compute and draw confidence intervals.)

It should be noted that it is not sufficient to judge whether or not the confidence interval includes the central line in order to detect statistical significant differences at a 5% level. In that case it would be sufficient to have a profile per scale indicating: "p<.05" or "p>.05". By visual inspection of the scales one should take into account the magnitude of the difference between the referencegroup and the group of interest. A small difference (say 25% versus 26% on a given question) may be statistically significant because many employees answered the question, but may also be of no relevance. On the other hand with small groups the confidence intervals can be very wide and include the central reference-line. Although this means that the difference is not statistically significant it still can be of substantial value and large differences (15% or 20% or more) may exist. This can be an indication for further research.

The added value of a profile with confidence-bands is that it visualizes the means and confidence-intervals of a studygroup and of a referencegroup. This presentation forms the quantitative backbone of differences between studygroups and a referencegroup. Take for example the scores of a studygroup that are all far from the means of a referencegroup but the confidence-intervals include the centerline. If the studygroup consists of a company or a department and the referencegroup represents a group of employees from a large number of comparable companies, there exists strong evidence that additional research into work and working conditions of the studygroup-employees is desired.

Which intrinsic consequences are to be drawn from the differences that are found in an actual situation is an other question. Is a difference of 1% important enough to take further actions (specific further research, intervention) or are only differences of 20% or more of interest ? For making these decisions, it is usually necessary to have available all other information on health and work (e.g. specific workplace research, biometry and absence data).

9.3 COMPUTATION OF THE CONFIDENCE INTERVALS

Confidence intervals for the dichotomous items and the subscales that are constructed from them cannot be computed by simply using the usual on the normal distribution based symmetric method. The computation of the confidence intervals used by V-PROF is according to the method used by Fleiss (1981) for the computation of confidence intervals for proportions. The confidence intervals for the subscales, which in most cases are very skewed distributed, are computed in a similar way. The subscalescore is transformed into a fraction. For this fraction the confidence interval is computed and then this 'pseudo-confidence interval' is transformed back to the original scale. A statistical foundation of this intuitive method and an analytically correct method is given by Brand & Radder (1992).

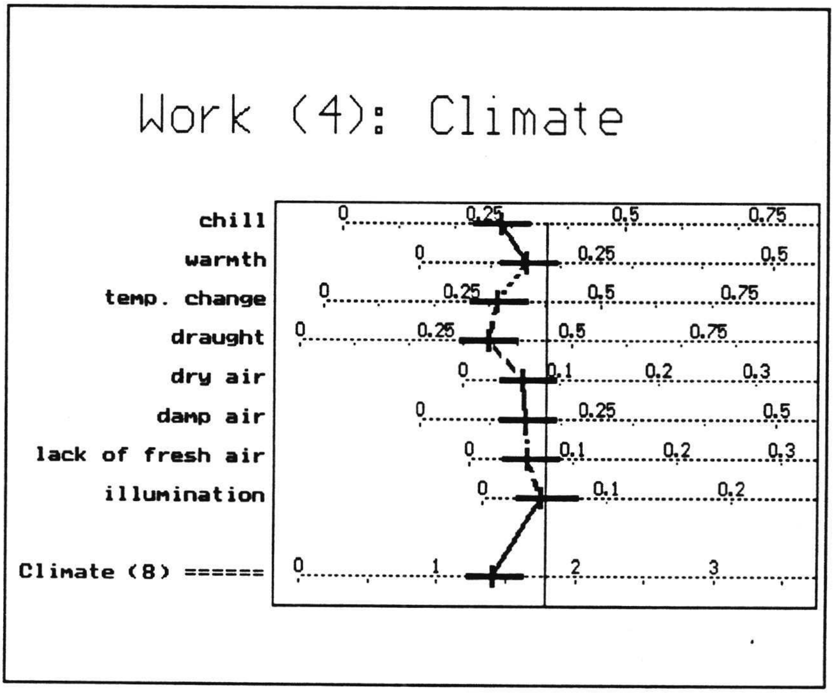
9.4 DIRECT AND INDIRECT STANDARDISATION

Usually the structure of the studygroups will not be exactly the same as the referencegroup. This may lead to interpretation problems when differences are found: is a high score in an examined group to be explained by possible differences in the distribution of a confounding variable (e.g. age) or by differences in worksituation. This problem can be solved by direct or indirect standardization. The method of GRP offers both methods. With direct standardization the distribution of the confounder variable in the referencegroup is used when computing the standardized values for the studygroup(s). Indirect standardization means that when computing centerline- and scalevalues for the referencegroup, the composition of the studygroup is taken into account. The advantage of direct standardization is that different studygroups can be compared to each other. With indirect standardization this is not the case. A disadvantage of standardization is a decrease in precision of the computed statistics when the studygroups are small. This is especially the case with direct standardization.

9.5 ANALYSIS OF SEPARATE ITEMS

It is possible in the program to generate first the overall mean of a subscale (see figure 9.1 where different subscales are put together). After that, one can examine the answers on the separate questions that make up the subscale. In this way one first starts with a quick, global impression of the results and then secondly study them more in depth. An example of a profile of a subscale with its separate items is given in figure 9.2.

Figure 9.2 Item-profile for company X (curved line) and a referencegroup with the 95% confidence interval for the studygroup



9.6 ADVANTAGES OF GRAPHICAL REFERENCE-PROFILES

- The program is able to present a bulk of information (a large number of questions and subscales) in a coherent way. All variables are shown, not only the ones "that matter". This diminishes the risk of falsely drawing conclusions from accidental differences (for instance based on a single item out of a long list of questions) or missing differences that are not expected (blind spots).
- The presentation of the confidence intervals together with the means gives at the same time the statistical significance of a difference, together with the magnitude (intrinsic relevance) of that difference. These together form the quantitative backbone for judging the relevance of the encountered differences.
- When a large, sufficiently diverse referencegroup is available, it is often possible to come to conclusions on the basis of small studygroups.
- There is flexibility in constructing the structure of the referencepopulation. It is simple to make several comparisons, for instance with a "representation of the dutch working population", and afterwards with "the same group in which differences in e.g. age-composition have been corrected for by standardization".
- The results are well organized. It is easy to illustrate many statistical principles by means of Graphical Reference-Profiles: direct- and indirect standardisation, confidence intervals, statistical testing, etcetera.

REFERENCES

BRAND, J & RADDER J.J. Betrouwbaarheidsintervallen voor somscores van dichotome items in V-Prof, Department of Statistics and Computer Science, Leiden, NIPG-TNO, 1992.

FLEISS, J.L. Statistical methods for rates and proportions. New York, Wiley, 1981.

PUTTEN, VAN D.J. De vragenlijst voor Periodiek Bedrijfsgeneeskundig Onderzoek. Leiden, NIPG-TNO, 1989.

DENTAL CLINICAL TRIALS AND MULTIVARIATE ANALYSIS

D.N. Geary*, E. Huntington, R.J. Gilbert**

* University of Oxford, Oxford, United Kingdom

** Unilever Research, Port Sunlight Laboratory, Merseyside, United Kingdom

Abstract

Some background on dental disease is presented. Four dental clinical trials conducted between 1974 and 1986 are considered. The trials compare toothpastes, with three or six pastes per trial. Between 1400 and 3000 children enter each trial. Complete data for a given child consist of five oral health variables measured on each of three or four occasions, over a period of three years. The five variables are plaque, DEFS (decayed surfaces), DEFT (decayed teeth), gingivitis and calculus.

The complete data are transformed towards multivariate normality. Multivariate analyses of covariance (MANCOVAs) show up significant effects.

The MANCOVA analyses are based on data from 72% of the children who entered the trials: only those who provided complete data measured by the same clinician each year. Missing data are examined for any lack of randomness which might invalidate conclusions from the MANCOVAs. It turns out that no serious imbalance occurs between pastes in any of the trials, with respect to missing values, though a significantly greater proportion of boys than girls have missing values in one of the trials.

If a trial could somehow be limited to involve only subjects who are most likely to show up differences between toothpastes, this could lead to some economy. This possibility, of economizing by involving only selected subjects in the trials, is investigated. In the given trials, pre-selection of only girls, or only children who brush their teeth more than twice per day on average, seems to discriminate almost as well as when no selection is imposed and data from all subjects are used.

Differences in paste effects with respect to decay might become statistically significant in a caries trial before the standard three years' duration. If so, a decision may be made

to stop the trial early, claiming a significant effect and saving time, effort, money and tooth decay.
Sequential testing for paste differences, based on (univariate) caries increments, is carried out retrospectively for each of the four trials. This testing takes account of the repeated measurements of caries from year to year. In two of the trials it may have been possible to draw 'safe' conclusions about paste differences in less than three years.

Keywords: *dental clinical trials, diagnostic error, multivariate analysis of covariance, sequential testing, transformations to multivariate normality.*

10.1 INTRODUCTION

Plague bacteria can be removed from teeth by brushing. If there are not removed, several ill-effects may ensue. Plaque attached to tooth enamel can ferment sugars to produce acids which demineralize the enamel, causing *decay*. Plaque accumulated at the margins of teeth and gums can generate toxic waste products leading to gingivitis, inflammation of the gums. And, especially near outlets of salivary ducts, plaque can acquire salivary salts and become a stone-lime deposit known as *calculus* (or *tartar*), not directly related to decay or gingivitis, but cosmetically undesirable. This paper examines plaque, decay, gingivitis and calculus data obtained from four dental clinical trials sponsored by Unilever Research between 1974 and 1986. Typically, attention in the analysis of such trials has focused on the decay variable. Examination of the multivariate response gives a clearer overall view of trial outcome. First, some background on dental disease is presented.

The Unilever data are described, and transformations of the data to multivariate normality are considered. Standard multivariate analyses of covariance address some interesting questions, and results from these are described. Other issues are discussed in less detail; missing values, economizing by excluding selected subjects from the trials, estimation of the effect of diagnostic error, and (univariate) sequential testing based on caries increments.

10.2 DENTAL DISEASE

Dental caries is tooth decay. Murray (1989) writes (p. 2) that "At present, in Britain, only one adult in every thousand is caries free, and 25% of adults aged 16 years and over have no natural teeth at all". However, according to Sir James Crichton Browne (1892). "It is impossible to believe that the British Empire would have become what it is to-day if, amongst those hardy Norsemen, who pushed up their keels on the shore at Ebbsfleet and entered upon the making of England, there had been only one sound set of teeth in every ten", much less in every thousand; and "no nation has ever climbed to pre-eminence on carious teeth, or can retain its pre-eminence when its teeth are no more".

It certainly seems that caries was not so prevalent in earlier centuries as in ours. Ruffer (1921) observes (p. 315) that "Dental caries occurred at all periods of Egyptian history", but qualifies this by quoting (p. 293) Elliot Smith and Wood Jones (1910); "Both in Nubi and Egypt the ordinary form of caries is exceedingly rare in predynastic and protodynastic people, and among the poorer classes it never became at all common until modern times... Dental caries became common as soon as people learned luxury".

The address by Crichton Browne (1892) to a Branch meeting of the British Dental Association, was remarkable for its 19th-century identification of fluorine as a possible remedy for caries. "In as far as our own country, at any rate, is concerned, this is essentially an age of white bread and fine flour, and its an age therefore in which we are no longer partaking to anything like the same amount that our ancestors did of the bran of husky parts of wheat, and so are deprived to a large degree of a chemical element which they received in abundance, namely fluorine.. I think it well worthy of consideration whether the re-introduction into our diet, and especially into the diet of child-bearing women and of children, of a supply of fluorine in some suitable natural form.. might not do something to fortify the teeth of the next generation." (See also Murray, p.115). It was not until 1945 that fluorine (in the form of sodium fluoride) was added to communal drinking water, at Grand Rapids in the USA. Seven-year results from this experiment, reported by Arnold et al (1953), presented clear evidence for the efficacy of fluoridation in suppressing caries. In the UK, fluoridation of water supplies was first carried out in 1955.

Trials of fluoride *toothpastes* were first reported in the mid-1940s. At first, fluoride in toothpaste seemed ineffective. However by the mid-1960s, certain combinations of fluoride compounds in toothpaste had been established as beneficial. Since then, many clinical trials of fluoride in toothpaste have been conducted (see eg, Jenkins, 1985); and since the late 1970s, trials involving placebo toothpastes have generally been avoided on ethical grounds.

According to an FDI/WHO report (1985), nine developed countries, including the UK, showed reduction of 30% to 50% in prevalence of dental caries in 5- and 12-year-old children of the decade (approximately) 1973-83. The report suggests the reductions are strongly associated with use of fluoride in water supplies and in toothpastes. Jenkins (1985) points out that in the UK only about 10% of the population receives fluoridated water, and he suggests that widespread use of fluoride toothpastes has been the main reason for recent reduction in UK caries.

Periodontal disease is disease of the tissues, including the gums, surrounding the teeth. Murray (1989) writes that "dental caries and periodontal disease are so widespread that virtually everybody in the world, certainly every adult, has either one or both of these conditions". For nearly 5000 years, calculus was thought to be the main cause of periodontal disease (Mandel and Gaffar, 1986); but since the 1960s, plaque has taken over as the *sine qua non* of both calculus and periodontal disease (cf Section 1). Gingivitis is inflammation of the gums. *Periodontitis* is the inflammation of structures supporting the gums, and is generally more serious than gingivitis. Rate of tooth-loss due to periodontal disease tends to increase with increasing age; and there is evidence to suggest that in developed countries over about 40 years of age lose more teeth because of periodontal disease than because of caries (cf Jenkins, 1989, pp. 333,334). By comparison with caries, relatively few studies involving large samples of people and standardized methodology have been conducted on periodontal disease (FDI/WHO, 1985).

Incidentally, Mattila et al (1989) identify an association between dental disease and acute myocardial infarction, evident from two case-control studies involving 81 and 121 subjects respectively. They point out that further prospective studies are necessary to confirm the association.

10.3 THE DATA

It has become common for caries clinical trials to be conducted over three years, with the co-operation of children typically aged 12 to 15 (cf FDI, 1982, p. 295); the relatively high caries activity observed in children between these ages helps to identify ant significant toothpaste effects. Four trials are considered here. They were conducted in Berkshire (BERKS), 1974-1977; Münster (MUNST), West Germany, 1976-1979; Somerset (SOMER), 1976-1979; and Lanarkshire (LANAR), 1983-1986. Summaries of caries-related findings in these trials are given by Murray and Shaw (1980), Buhe et al (1984), Andlaw et al (1983) and Stephen et al (1988), respectively. For each trial, two dental clinicians were appointed to share the work of making the relevant measurements, each accounting for about half the subjects involved. (Note that occasionally other clinicians made the measurements, in the absence of the two main trial-clinicians). Numbers of (roughly) 12-year-old children volunteering to participate in the trials are presented in Table 10.1. It was expected that some 20-25% of children would not provide complete data - families sometimes move away from the study area of children are absent on data of clinical examinations - and the trials were designed to take account of such data-losses.

Table 10.1 Numbers of subjects involved in the trials

	BERKS	MUNST	SOMER	LANAR	total
na	947	1280	1073	2111	5411
nb	325	273	317	890	1805
nc	159	9	96	4	268
total	1431	1562	1486	3005	7484

na subject with complete data, recorded by the same clinician (one of the two main trail-clinicians each year
nb subjects with incomplete data
nc subjects with complete data but differing clinicians

The trials were designed to compare toothpastes; the BERKS, MUNST and SOMER trials each involved three pastes, while LANAR involved six pastes. After initial examinations, subjects were stratified, according to different criteria in the different trials (such as sex, clinician and baseline caries). Pastes were then allocated using modified randomized permuted blocks within strata (children from the same family were assigned to the same paste, and in LANAR the trail design specified half as many subjects on two pastes as on the other four). Children used the same toothpaste for three years. Neither children nor clinicians knew who was using which paste (the trials were *double-blind)*. Free toothpaste was supplied to participants and their families during the trials.

Five variables were measured on each subject in each trial: plaque (P), DEFS (S), DEFT (T), gingivitis (G) and calculus (C). DEFS and DEFT are standard measures of tooth decay, denoting the numbers of decayed, missing or filled *tooth surfaces* and *teeth* respectively. In MUNST, the variables were measured at the outset, after 18 months and after three years; in the other three trials, the variables were measured at the outset and each year for three years. In addition, brushing frequency (B) data were recorded in SOMER at the outset, and in LANAR at years one, two and three.

Clinicians count either four or five surfaces on each tooth, so DEFS is at least as big as DEFT, and both are integers. Plaque was measured using the Greene and Vermillion (1964) index, on different numbers of tooth surfaces in the different trials, with each surface being given a plaque score of zero, one, two or three (three representing a high level of plaque). Thus in BERKS, for instance, a total P score for each subject was obtained by adding P measurements from two surfaces on each of 12 teeth. Gingivitis scores were obtained similarly, using the Löe and Silness (1963) index. In each trial, a search for calculus was made on each of six particularly calculus-prone teeth. Absence of calculus on all of these teeth was registered as zero; presence on any of the six teeth was registered as one in LANAR, one or two in SOMER (according to severity) and one, two, three or four in BERKS and MUNST. Brushing frequencies were recorded on scales between zero and nine; but for the purpose of analysis, were later transformed to binary scales. Ranges of variables in the distinct trials are given in Table 10.2.

Table 10.2 Ranges of variables

	BERKS	MUNST	SOMER	LANAR
P	0-53	0-65	0-16	0-17
S	0-68	0-79	0-66	0-79
T	0-22	0-25	0-27	0-25
G	0-35	0-48	0- 8	3-43
C	0- 4	0- 4	0- 2	0- 1
B	----	----	0- 4	0- 9

10.4 TRANSFORMATIONS "TO MULTIVARIATE NORMALITY"

The (P,S,T,G,C) data described in section 3 are certainly *not* multivariate - normally - distributed: univariate plots reveal skewness and discreteness incompatible with univariate normality. How then should the data be analysed ?

One answer, pursued in what follows, is to transform the data towards multivariate normality, and use standard normal-theory approaches to the analysis. An appeal to the robustness properties of normal-theory techniques can be made in respect of any residual non-normality in the transformed data. The disadvantage of this approach is that the

extent to which results are influenced by departures from multivariate normality is uncertain and unmeasurable, and may be large. (Attempts to gauge the effects of such departures can be made, in given circumstances, by simulating data). One important advantage is that the approach is well documented and computer programs for its implementation are readily available in several statistical packages. How then should the data be transformed ? They are extensive, with (in three of the trials) four measurements, one per year, of each variable, on each subject. It was decided to transform each variable by a separate transformation, aiming at five separate sets of multivariate normal data, in the hope that these might be jointly approximately multivariate normal. Increases or decreases over time in any one variable can then be easily interpreted. Thus, for instance, the plaque transformation would be based on P at years zero, one two and three.
The following transformation was considered:

$$\underline{x} \rightarrow \log(\underline{\mathbf{x}} + c\underline{\mathbf{1}}), \qquad 0.01 \leq c \leq 200.00;$$

How should a value of c be chosen to provide the best transformations ?
A given test of multivariate normality may be used to choose between values of c. Several such tests exist, with no one test being uniformly best (see eg, Small, 1985). The test due to Hawkins (1981) is considered here. Hawkins' (1981) test was developed in the context of discriminant analysis, as a test of multivariate normality *and* homoscedasticity; and has been used to advantage in this context by, for example, Mclachlan and Basford (1988). Each vector measurement has an associated scalar value *A*, which is distributed as Uniform (0,1), under the null hypothesis of multivariate normality and homoscedasticity. Asymptotically, the *A*s are independent, and the Anderson-Darling statistic is used to test their fit to Uniform (0,1). While Hawkins' test assumes data come from different groups, it affords a statistic W_T appropriate for testing multivariate normality when data are combined into one group.

Even when transformed, the data are not multivariate normal in any of the trials. A rough ordering of trials, in terms of having data with improving normality, is: LANAR, SOMOR, BERKS and MUNST.

10.5 MULTIVARIATE ANALYSES OF COVARIANCE

10.5.1 Pastes, Contrasts and Significance Results

Denote transformed plaque at year three by P3, and so on. Multivariate analyses of covariance were carried out, using statistical packages SAS and SPSS, with variates P3, S3, T3, G3 and C3 adjusted for covariates P0, S0, T0, G0 and C0. Consideration of data at intermediate years (one and two in BERKS, SOMER and LANAR, one-and-a-half in MUNST) is deferred to Section 5.6. Factors involved were PASTE (with three or six levels), SEX (1=girl, 2=boy) and CLINICIAN (two clinicians in each trial, with different clinicians in different trials). Pastes used in each of the trials are indicated in Table 10.3.

Table 10.3 Pastes in the trials

paste	BERKS	MUNST	SOMER	LANAR
1	placebo	placebo	TMP	1000 ppm F
2	1000 ppm F, *a*	1000 ppm F	1000 ppm F, *c*	1000 ppm F, plus ZCT
3	1000 PPM F, *b*	1500 ppm F	1000 ppm F, *d*	1500 ppm F
4				1500 ppm F, plus ZCT
5				2500 ppm F
6				2500 ppm F, plus ZCT

Abbreviations:
ppm - parts per million
F - fluoride
TMP - sodium trimetaphosphate
a and *c* - normal-abrasion toothpastes
b and *d* - low-abrasion toothpastes
ZCT - zinc citrate trihydrate

The ZCT component was included in some LANAR pastes as a possible inhibitor of plaque and calculus. Numbers of subjects receiving the various pastes in the different trials, excluding those with incomplete data, with changing clinicians and with outlying data, are presented in Table 10.4.

Table 10.4 Number of subjects on each paste

paste	BERKS	MUNST	SOMER	LANAR
1	302	422	345	425
2	331	435	354	397
3	307	416	367	424
4				421
5				207
6				209
total	940	1273	1066	2083

Roughly equal numbers of subjects were allocated to each of the three pastes in the BERKS, MUNST and SOMER trials. In LANAR, the trial was designed to have about half as many subjects on paste 5 and 6 as on pastes 1, 2, 3, of 4. (This was because it was anticipated that at the highest fluoride level, 2500 ppm, fewer subjects would be necessary to show a significant fluoride effect.) The overall PASTE effects for the MANCOVA computations were divided into orthogonal single-degree-of-freedom contrasts. Thus for BERKS, MUNST and SOMER, two PASTE contracts were defined:

	contrast	
paste	(1)	(2)
1	0	-2
2	-1	1
3	1	1

Contrast (1) represents the difference between pastes 3 and 2, and contrast (2), the difference between the average of 3 and 2 (both fluoride pastes) and paste 1 (placebo in two of the trials). Similarly, in LANAR, five PASTE contrasts were initially identified:

contrast

paste	(1)	(2)	(3)	(4)	(5)
1	-4	2	1	-4	2
2	-4	2	-1	4	-2
3	-1	-3	1	-1	-3
4	-1	-3	-1	1	-3
5	5	1	1	5	1
6	5	1	-1	-5	-1

Of particular interest in LANAR are contrasts (1) and (3); (1) representing a linear effect of fluoride level, and (3) representing the difference between non-ZCT and ZCT pastes. Contrasts (4) and (5) are interactions of (1) and (3), and (2) and (3), respectively. Contrast (2) is a comparison between a weighted average of 1000 and 2500 ppm F, and 1500 ppm F. At first, all factor main effects and interactions were included in the models. If highest-order interactions were non-significant at the 5% level they were eliminated from the given model, and the corresponding sums of squares and cross-products were absorbed into residuals. This process was repeated if appropriate, given the significance levels of Table 10.5.

Table 10.5 Significance levels from MANCOVA analyses

effect	BERKS	MUNST	SOMER	effect	LANAR
PASTE (1)	0.008	0.348	0.880	PASTE (1)	0.026
PASTE (2)	0.000	0.000	0.089	PASTE (2)	0.603
SEX	0.006	0.000	0.000	PASTE (3)	0.000
CLINICIAN	0.000	0.000	0.000	PASTE (4)	0.456
				PASTE (5)	0.422
				SEX	0.000
				CLINICIAN	0.000
				PASTE*SEX	0.717
				PASTE*CLINICIAN	0.220
				SEX*CLINICIAN	0.000

The main effects of SEX and clinician are highly significant in all trials. In BERKS and MUNST, no interactions appear significant. In SOMER, the PASTE*SEX*CLINICIAN and SEX*CLINICIAN interactions verge on significance, at the 5% level; while in LANAR, the SEX*CLINICIAN interaction is highly significant. The SOMER trial is the only one with no significant PASTE contrast. The nature of the significant effects is explored in Sections 5.3 and 5.4.

10.5.2 Correlations and regressions

Correlations between transformed variables at year three in the trials, adjusted for covariates at year zero, are presented in Table 10.6.

Table 10.6 Correlations between transformed variables

	P/S	P/T	P/G	P/C	S/T	S/G	S/C	T/G	T/C	G/C
BERKS	0.04	0.01	0.51	0.15	0.85	0.01	-0.05	0.02	-0.06	0.18
MUNST	0.08	0.08	0.48	0.04	0.87	0.10	-0.05	0.10	-0.03	0.13
SOMER	0.02	-0.00	0.46	0.06	0.90	0.05	-0.12	0.03	-0.12	0.18
LANAR	0.06	0.08	0.40	0.08	0.88	0.09	-0.05	0.11	-0.05	0.06

Correlations for the S/T decay variables are consistently high (around 0.88) as might be expected. Those for the P/G pair are also relatively high (around 0.46). Otherwise, correlations are less than 0.2 in absolute value, usually positive for P/S, P/T, P/C, S/G, T/G and G/C, and negative for S/C and T/C. Negative correlations between decay and calculus are generally accepted as existing: decay results from demineralization of enamel, while calculus arises from the "opposite" process, deposition of minerals of enamel. In MUNST and LANAR there are highly significant ($P < 0.005$) *negative* regressions of S3 and T3 on C0, further reflecting the negative relationship between decay and calculus. Other similarly significant, but *positive*, regressions were those of P3 on P0, S3 on S0, T3 on T0, G3 on G0 and T3 on T0 (in all trials); P3 on G0, T3 on S0 and G3 on P0 (each in three trials); P3 on S0 and S3 on T0 (each in two trials); and S3 on P0, S3 on G0 and T3 on P0 (each in one trial).

10.5.3 Canonical variate coefficients

Canonical variates comprise in a sense, the best linear combinations of variables for separating groups. They are obtained by maximizing the ratio of the between- to the within- groups sums of squares, with respect to the coefficients of the variables. To obtain the canonical variates referred to here, the original variates are first standardized to have unit *within*-group variance. This allows a direct comparison between coefficients, illustrated below, by ensuring that the relative sizes of the variable-coefficients reflect the relative abilities of the corresponding variates to distinguish *between* groups.
For example, the canonical variate corresponding to the significant PASTE (1) effect in BERKS (cf Table 10.5) is:

$$V = 1.05*P3 - 0.43*S3 + 0.07*T3 - 0.56*G3 - 0.38*C3$$

Recall that PASTE (1) represents the contrast between pastes 3 and 2. The variate V suggests that, in discriminating between pastes 3 and 2, the variable P3 is particularly important, because its associated coefficient in V is the largest. However proper interpretation of the canonical variate coefficients is not quite so straightforward; allowance must be made for correlations between the variables. Thus since plaque and gingivitis are positively correlated (with correlation 0.51, cf Table 10.6), they may tend to "cancel one another out" in the canonical variate if their coefficients are of opposite signs, which indeed they are, in V. Even allowing for this though, the relative sizes of the variable-coefficients in V suggest that plaque is the best discriminating variable for distinguishing pastes 3 and 2. Canonical variate coefficients were evaluated using SPSS (output from which refers to them as "standardized discriminant function coefficients"), for each of the significant effects of Table 10.5, and these are presented in Table 10.7. In each case, the coefficient for the variable informally judged to the best discriminator for the corresponding effect, is underlined; and, where there is potential ambiguity in judging which is best, coefficients tending to cancel one another out are printed in bold.

Table 10.7 Canonical variate coefficients for significant effects

effect	variable	BERKS	MUNST	SOMER	LANAR
PASTE (1)	P3	1.05			0.14
	S3	-0.43			-0.97
	T3	0.07			0.09
	G3	-0.56			-0.31
	C3	-0.38			-0.40
PASTE (2)	P3	-0.32	-0.12		
	S3	-0.58	-1.21		
	T3	-0.29	0.28		
	G3	-0.06	-0.05		
	C3	0.43	-0.03		
PASTE (3)	P3				-0.39
	S3				**-0.90**
	T3				**-0.96**
	G3				-0.15
	C3				-0.72
SEX	P3	-1.03	0.52	1.07	-0.41
	S3	-0.40	**0.36**	0.30	-0.48
	T3	0.36	**-0.61**	-0.43	0.43
	G3	0.15	0.62	-0.35	-0.74
	C3	-0.02	-0.18	-0.19	0.05
CLINICIAN	P3	-0.15	0.15	0.56	0.90
	S3	-0.70	0.19	0.36	**-1.07**
	T3	0.58	-0.44	-0.41	**0.95**
	G3	-0.06	0.60	0.30	-0.66
	C3	0.95	-0.82	-0.77	-0.07
SEX*CLINICIAN	P3				0.44
	S3				-0.36
	T3				-0.07
	G3				-0.96
	C3				0.08

As noted above, plaque is the best discriminating variable for the PASTE (1) effect in BERKS. For three of the other four significant PASTE effects, DEFS is the best discriminator; with two of these, the associated contrasts are between fluoride toothpaste and placebo, and in the third, between different levels of fluoride in toothpaste. The remaining significant PASTE effect is the ZCT effect in LANAR, and the best discriminator for this is calculus. Plaque and gingivitis are the best discriminating variables for two each of the significant SEX effects; and calculus and plaque discriminate best for three and one of the four significant CLINICIAN effects, respectively.

The directions of the significant effects are indicated by the estimated marginal means of Section 5.4.

10.5.4 Estimated marginal means

Searle et al. (1980) propose the name "population marginal mean" (PMM) as preferable to the term "least squares mean", used, for instance, in output of the SAS general-linear-models procedure GLIM. The PPM is a parametric function, with one PPM for each level of each factor in a linear model design. For a given factor-level, the PPM is independent of the number of measurements on this level over levels of other factors. These PMMs are estimated by best linear unbiased estimators. They can be adjusted for covariates in the usual way; the adjusted means being, informally, those which would have been observed had all covariates in the different groups been the same. SAS least squares means, or "estimated marginal means", of P3, S3, T3, G3 and C3, adjusted for P0, S0, T0, G0 and C0, are presented in Table 10.8. Columns in Table 10.8 corresponding the variables selected from Table 10.7 as the best discriminators for significant effects, are underlined.

Table 10.8 Estimated marginal means (approximate standard errors in brackets)

		P3	S3	T3	G3	C3
BERKS						
PASTE	1	3.68 (0.01)	3.08 (0.01)	2.73 (0.01)	16.39 (0.25)	0.66 (0.03)
	2	3.65 (0.01)	3.03 (0.01)	2.69 (0.01)	16.22 (0.24)	0.76 (0.03)
	3	3.68 (0.01)	3.00 (0.01)	2.67 (0.01)	16.09 (0.25)	0.71 (0.03)
SEX	1	3.65 (0.01)	3.03 (0.01)	2.70 (0.01)	16.00 (0.22)	0.70 (0.03)
	2	3.69 (0.01)	3.04 (0.01)	2.70 (0.01)	16.47 (0.19)	0.72 (0.02)
CLINICIAN	1	3.67 (0.01)	3.05 (0.01)	2.70 (0.01)	16.18 (0.23)	0.57 (0.03)
	2	3.67 (0.01)	3.02 (0.01)	2.69 (0.01)	16.29 (0.21)	0.85 (0.02)
MUNST						
PASTE	1	3.12 (0.01)	3.71 (0.01)	3.34 (0.01)	20.27 (0.31)	0.61 (0.02)
	2	3.09 (0.02)	3.64 (0.01)	3.31 (0.01)	19.79 (0.30)	0.61 (0.02)
	3	3.08 (0.02)	3.61 (0.01)	3.30 (0.01)	19.60 (0.31)	0.62 (0.02)
SEX	1	3.04 (0.01)	3.66 (0.01)	3.32 (0.00)	18.88 (0.25)	0.62 (0.02)
	2	3.16 (0.01)	3.65 (0.01)	3.31 (0.00)	20.90 (0.26)	0.61 (0.02)
CLINICIAN	1	3.07 (0.02)	3.66 (0.01)	3.32 (0.00)	19.11 (0.26)	0.69 (0.02)
	2	3.13 (0.01)	3.65 (0.01)	3.31 (0.00)	20.66 (0.25)	0.54 (0.02)
SOMER						
PASTE	1	2.46 (0.01)	2.68 (0.02)	2.47 (0.01)	2.05 (0.01)	-0.74 (0.04)
	2	2.44 (0.01)	2.62 (0.02)	2.42 (0.01)	2.05 (0.01)	-0.77 (0.04)
	3	2.45 (0.01)	2.64 (0.02)	2.44 (0.01)	2.05 (0.01)	-0.76 (0.04)
SEX	1	2.41 (0.01)	2.65 (0.01)	2.45 (0.01)	2.05 (0.01)	-0.73 (0.04)
	2	2.49 (0.01)	2.64 (0.01)	2.44 (0.01)	2.05 (0.01)	-0.78 (0.04)
CLINICIAN	1	2.49 (0.01)	2.66 (0.01)	2.45 (0.01)	2.07 (0.01)	-0.92 (0.04)
	2	2.41 (0.01)	2.64 (0.01)	2.44 (0.01)	2.03 (0.01)	-0.59 (0.04)
LANAR						
PASTE	1	2.83 (0.01)	3.17 (0.01)	2.88 (0.01)	24.76 (0.10)	0.22 (0.02)
	2	2.82 (0.01)	3.18 (0.01)	2.89 (0.01)	24.52 (0.10)	0.19 (0.02)
	3	2.83 (0.01)	3.18 (0.01)	2.88 (0.01)	24.71 (0.10)	0.25 (0.02)
	4	2.82 (0.01)	3.17 (0.01)	2.88 (0.01)	24.73 (0.10)	0.16 (0.02)
	5	2.84 (0.01)	3.14 (0.01)	2.86 (0.01)	24.93 (0.10)	0.22 (0.03)
	6	2.82 (0.01)	3.13 (0.01)	2.87 (0.01)	24.63 (0.10)	0.13 (0.03)
SEX	1	2.80 (0.00)	3.15 (0.01)	2.87 (0.00)	24.24 (0.07)	0.19 (0.01)
	2	2.85 (0.00)	3.17 (0.01)	2.88 (0.00)	25.19 (0.07)	0.20 (0.01)
CLINICIAN	1	2.85 (0.00)	3.15 (0.01)	2.88 (0.00)	24.56 (0.07)	0.19 (0.01)
	2	2.80 (0.00)	3.17 (0.01)	2.88 (0.01)	24.87 (0.08)	0.20 (0.01)
SEX*	1*1	2.82 (0.01)	3.13 (0.01)	2.86 (0.01)	23.85 (0.10)	0.19 (0.02)
CLINICIAN	1*2	2.78 (0.01)	3.17 (0.01)	2.88 (0.01)	24.62 (0.10)	0.19 (0.02)
"	2*1	2.87 (0.01)	3.17 (0.01)	2.89 (0.01)	25.27 (0.10)	0.19 (0.02)
"	2*2	2.82 (0.01)	3.17 (0.01)	2.87 (0.01)	25.11 (0.11)	0.20 (0.02)

Plaque is significantly greater on paste 3 (low-abrasion 1000 ppm F) than on paste 2 (normal abrasion 1000 ppm F) in the BERKS trial. The difference may be related to the different abrasion levels. In LANAR, the significant PASTE (1) effect arises mainly because there is less DEFS on pastes 5 and 6, with 2500 ppm F, than on pastes 1, 2, 3 and 4, with 1000 of 1500 ppm F. More fluoride, in the toothpastes tested, leads to less decay. Incidentally, for fear of negative consequences (especially in infants) of having *too much* fluoride in toothpaste, the current legal limit in over-the-counter UK toothpaste is 1500 ppm F.

In BERKS and MUNST, the PASTE (2) effects spring mainly from smaller DEFS measurements on the fluoride toothpastes than on placebos. The LANAR PASTE (3) effect represents a clear inhibition of calculus associated with addition of the ZCT ingredient. Greater plaque and gingivitis for boys (SEX 2) than for girls characterizes the SEX effect in all the trials, except that in SOMER, there appears to be little difference between sexes on the gingivitis variable. Girls seems generally to make more effort than boys in the realm of oral hygiene (cf Section 5.5). (The fact that such effort is not rewarded more substantially, by less caries for girls than for boys, may be related to earlier tooth-eruption (and therefore more teeth at risk of decay) for girls than for boys.

In three of the four trials, there are large differences between clinicians in respect of calculus measurements. Table 10.8 suggests that clinicians generally agreed well in their assessments of DEFS and DEFT.

The SEX*CLINICIAN interaction in LANAR is most clearly signalled by gingivitis; clinician 1 records less gingivitis for girls than clinician 2 and marginally more gingivitis for boys than clinician 2. A similar pattern is observed in the DEFT measurements.

10.5.5 Brushing frequencies

Tooth-brushing frequency, B was recorded for each subject in the SOMER and LANAR trials; in SOMER at year zero, and in LANAR, at years one, two and three. Different coding systems for B were used in the two trials, but there, these are reduced to a common binary system: 1 representing "once or less, per day, on average", and 2, "more than once per day on average".

The significance levels of main effects and interactions not involving B0, the brushing frequency at year zero, are not changed substantially by introduction of B0 (cf Table 10.5); nor are the canonical variate coefficients for the significant main effects of SEX and CLINICIAN (cf Table 10.7).

Consider now the B data from LANAR.

In LANAR, B was recorded at years one, two and three. At each year, girls brushed more than boys, on average. Chi-square statistics on one degree of freedom for differences between girls and boys with respect to B, based on dichotomizing B as in SOMER, were 119.2, 162.0 and 176.1 for years one, two and three respectively. At year three, 292 girls brushed once or less per day on average, and 794 brushed more than once per day. The corresponding numbers for boys were 553 and 444 respectively. In the absence of brushing frequency data for year zero, it seems best to employ B1 as a factor in MANCOVA analyses. Data B3 may be more influenced than B1 by the outcome of the trial and may represent a kind of response to the various paste used, inappropriate as criteria by which to classify subjects "prior to" the trial.

Effects of PASTE (1) (the linear Fluoride effect), PASTE (3) (the ZCT effect), SEX, CLINICIAN and SEX*CLINICIAN remain significant (cf Table 10.5) when B1 is included in the model. The only new significant effect involving B1 is the main effect of B1 itself. Examination of the canonical variate coefficients corresponding to significant effects reveals that best discriminators for PASTE (1), PASTE (3), SEX, CLINICIAN and SEX*CLINICIAN remain unchanged with the inclusion of B1. The main effect of B1 is highlighted by plaque measurements, but estimated marginal means of Table 10.9 shows that all variables have lower means associated with the higher level of brushing frequency.

Table 10.9 Estimated marginal means for B1 in LANAR (approximate standard errors in brackets)

B1	P3	S3	T3	G3	C3
1	2.846 (0.005)	3.184 (0.008)	2.891 (0.005)	24.874 (0.078)	0.209 (0.013)
2	2.812 (0.005)	3.139 (0.007)	2.862 (0.005)	24.595 (0.069)	0.183 (0.012)

10.5.6 Analyses incorporating data at intermediate years

Results described so far in Section 5 relate only to data measured at the beginning and end of each trial. Thus data at years one and two in BERKS, SOMER, and LANAR, and data at year one-and-a-half in MUNST, have so far been overlooked.

Although conclusions from a trial are usually drawn from results at the end of the trial, intermediate data may help answer some important questions. Would be apparent beneficial effect on decay on pastes two and three (as compared with one) in BERKS for instance, still be apparent if decay at years one and two were "added" to decay at year three to give a measure of "average decay" over the period? Or do pastes two and three "start weakly and finish strongly" ? Or again, do girls and boys accumulate tooth decay at the same rate over time for a given paste ?

Attempts to answer such questions may be made using multivariate repeated measures analysis of covariance, in which TIME is introduced as a factor *within* subjects.

In general, it is difficult for different clinicians to standardize measurements of calculus and gingivitis, by comparison with those of decay and plaque. In all four trials, the main SEX effect arises from plaque measurements - as noted in Section 5.4, girls have less plaque on average than boys. A highly significant TIME effect exists in each of the trials and for each trial, the main contribution to this is the linear increase in DEFS. The DEFT measurements likewise show an increasing trend in each trial. These results are as expected, since DEFS and DEFT are progressively increasing with time, whereas plaque, gingivitis and calculus formation may be reversed. Other trends in variables over time are less pronounced.

10.6 MISSING VALUES

Two general approaches for dealing with incomplete (multivariate) data are possible. One is to estimate the missing values (see eg, Beale and Little, 1975), and so "fill up" the data. However, such estimation can be computationally complicated and confidence in subsequent analyses of the overall data may be compromised by uncertainty about the imputed data. Another approach is to discard *all* data from subjects having *any* missing value and to analyse only the complete data. This is the approach used in Section 5. It is valid only if subjects having missing values occur randomly among the original subjects. About 72% of the children who start the four Unilever trials considered here finish the trials with complete data measured by the same clinician each year (cf Table 10.1). Analyses of Section 5 are based on data from these children. A further 24% of the children, for one reason or another, do not provide complete data; and almost 4% provide complete data measured by different clinicians on different occasions.

The questions are pertinent: do the 28% of subjects with data not used in the analyses of Section 5 represent random samples from all the trial-subjects; and if not, what, if any, biases arise in consequence, in the conclusions of Section 5 ? It is impossible to answer these with certainty; but answers to other incomplete-data questions shed some light on the randomness-and-bias questions, and are more accessible.

For example, are children on paste 1, say, more likely to miss an examination in the trial than children on paste 2 ? Or are boys more likely to miss examinations than girls ? Questions of this kind may be answered by analysis of contingency tables which cross-classify subjects according to completeness or otherwise of associated data and other factors such as paste or sex. These tables may best be analysed in the framework of generalised linear models, assuming Poisson error and log link function. Analyses were performed using GLIM. They were restricted to subjects who were examined at year zero by either clinician 1 or 2; numbers of subjects involved are given in Table 10.10.

Table 10.10 Numbers of subjects with complete versus incomplete data, excluding those not examined by clinician one or two at year zero (cf Table 10.1)

	BERKS	MUNST	SOMER	LANAR	total
complete	1064	1286	1146	2115	5611
incomplete	319	273	306	890	1788
total	1383	1559	1452	3005	7399

Having established answers to some questions on *numbers* of children with missing values, it is still worth asking: do children who miss examinations have more or less plaque, DEFS and so on, at the outset, than those who finish with complete data ? The answer to this may have implications for the design of such trials and may suggest possible directions of bias in the results of Section 5.

Multivariate analyses of variance were carried out for the four trials, with vector response P0, S0, T0, G0 and C0, transformed as for the complete data in Section 5 and with factors PA, SE, CL0 and INC. Numbers of subjects involved were as in Table 10.10.

In summary, imbalances occur in numbers of boys and girls with missing values in BERKS and LANAR; in numbers of subjects having missing values with clinicians one and two in LANAR; and in average baseline characteristics of complete-data versus incomplete-data subjects. However, no serious imbalance exists between pastes in any of the trials, with respect to missing values. It seems likely then that the main conclusions of Section 5 with regard to pastes are unaffected by ignoring data from subjects with missing values.

10.7 SELECTIVE INCLUSION OF STUDY SUBJECTS

If a trial could somehow be limited to involve only subjects who are most likely to show up differences between toothpastes, this could lead to some economy. But how many subjects who are most likely to show up differences be identified before a trial begins ? Downer et al (1975) discuss this question with particular reference to DEFS measurements. They examine data from three selected caries trials, observing that for these trials, girls consistently give higher mean DEFS differences between test and control pastes than boys. The suggestion is that caries trials would be more efficient if restricted to involve only girls. In addition, Downer et al (1975) suggest including only girls with initial DEFS greater than or equal to six (say), and with all six (say) "key surfaces at risk" (KSAR) initially present in the mouth or "unerupted". They select the KSAR by determining which tooth-surface measurements show the greatest average difference between treatment and control pastes, in their selected trials.
Significance levels for the four Unilever trials were recomputed using two restricted subgroups: first, girls only; and second, girls with SO $\geq$ 6. (KSAR measurements for the Unilever trials were not readily available). Results were obtained using the same transformations to normality for the subgroups as were obtained for the complete datasets (Near-)significance is most closely associated with decay measurements. It seems reasonable that differences between toothpastes with respect to caries become clearer with increased brushing frequencies. Pre-selection according to brushing frequency should be less potentially-controversial than simply choosing all girls.

10.8 DIAGNOSTIC ERROR

It is generally agreed that once a tooth has been correctly diagnosed as decayed, missing or filled, it cannot on a subsequent occasion be caries-free. Formation of caries is irreversible. In most caries trials however, a small proportion of teeth is diagnosed as being carious at one time and free from caries at a later time. This may happen because of error in diagnosis on the part of the clinician or error in recording on the part of the data-encoder. Elsewhere in this paper analyses are based on the data as recorded, without taking account of such diagnostic error. In this section, an attempt is made to gauge the extent of diagnostic error in the Unilever trials and to assess possible implications of this error for the results presented in other sections.

Reed and McHugh (1979) discuss estimation of caries incidence in the presence of diagnostic error. They recognise different ways in which diagnostic error may arise/ Denoting *true* caries states at two consecutive times as S_1 = CC (carious at both times) S_2 = NN (non-carious at both times) and S_3 = NC (non-carious at time one and carious at time two), and *diagnosed* states as S_1 = cc (carious at both times), S_2 = nn, S_3 = nc, S_4 = cn, they define probabilities $P_{\alpha\beta}$ of a tooth actually being in state S_α but being diagnosed in s_β, $\alpha = 1, 2, 3$, $\beta = 1, 2, 3, 4$. Then P_{11}, P_{22} and P_{33} correspond to correct diagnoses, while all other probabilities correspond to incorrect diagnoses. The *true* caries state S_α of a tooth is unknown and it is impossible to estimate $P_{\alpha\beta}$ sensibly without making some assumptions about the way in which diagnostic error arise. Reed an Mchugh (1979) consider a model, proposed by Carlos and Senning (CS, 1968). Of particular interest is the influence of incorrect diagnoses on the estimated caries incidence rates. The CS model assumptions are that: (a) the probability of misdiagnosis at either time is independent of the true state at that time; (b) the probability of misdiagnosis at time two is independent of the correctness or otherwise of diagnosis at time one; and (c) the probability of misdiagnosis at time one is equal to the probability of misdiagnosis at time two. The model has a set of associated parameters which may be estimated under the given assumptions. The parameter estimates are in turn used to estimate $\{P_{\alpha\beta}\}$. Reed and McHugh (1979) show how to construct point and interval estimates of the caries incidence rate under the CS model. They conclude (p. 478) by suggesting that "a large-scale empirical study, providing detailed information on the matrix of the $P_{\alpha\beta}$ ... would provide a direct basis for model selection". Matrices of $P_{\alpha\beta}$ are obtained from subjects with complete data recorded by the same clinician each year. Numbers of subjects involved are in Table 10.1, row *na*. Data were obtained for presence or absence of decay on individual teeth at each year. Define the estimated probability of correct diagnosis *(epcd)* as $\hat{P}_{11} + \hat{P}_{22} + \hat{P}_{33}$, where the ^'s denote estimates such as $P_{\alpha\beta}$. Then the CS *epcd's* for BERKS, MUNST, SOMER and LANAR are 0.974, 0.996, 0.999 and 0.007 respectively. Point estimates of caries incidence over three years for the four trials, based on the CS model, are 1.85, 3.89, 1.93 and 2.67 respectively. If diagnostic error is ignored, corresponding unadjusted point estimates are 1.80, 3.88, 1.93 and 2.66. (These are based on average numbers of teeth 22.29, 24.33, 22.22 and 24.14 in the four trials respectively). It is clear that for three of the four trials, the estimated extent of diagnostic error is very small, and adjusting for such error hardly changes estimates of caries incidence. In BERKS however, there is relatively more error. It is instructive to examine the effect of adjusting for this, via the CS model, on comparisons between pastes in the BERKS trial. Results are presented in Table 10.11.

Table 10.11 Estimates of caries incidence (with corresponding estimated 95% confidence intervals in brackets) unadjusted for diagnostic error, for the three pastes in BERKS

paste	unadjusted	adjusted using CS
1	2.12 (1.82, 2.42)	2.18 (1.88, 2.48)
2	1.72 (1.46, 1.99)	1.76 (1.50, 2.03)
3	1.58 (1.32, 1.84)	1.62 (1.36, 1.88)

Examination of Table 10.11 suggests that the pastes are separated slightly better with respect to caries incidence when the CS adjustment for diagnostic error is used. However the difference between adjusting and not adjusting is only slight. Analyses in this section suggest that diagnostic error (at least with respect to DEFT), in the four trials considered here, can safely be ignored.

10.9 POSTSCRIPT

"Horrible, no doubt, in its way is the dentist's chair, excruciating are the associations that cluster around it, but a timely resort to it robs it of its terrors, and converts it into a benefactor that lifts us from purgatorial pains into paradisiacal tranquillity and ease.. "Taken in its incipient stages caries is readily extirpated and stayed, and one of the most pleasing features in your profession at the present stay, if I may say so, is the anxiety you manifest to save teeth in which it has already made serious inroads, and to sacrifice as little as possible of natural dental structure. Marvellous is your skill, but you cannot yet supply artificial teeth with the genuine vital polish, and firm and nice in grip, like the natural organ planted in its living alveolus, and your efforts are therefore directed to preserving that natural organ, even when badly damaged, wherever it is practicable to do so. Extraction, which was at one time the opprobrium of your art, just as amputation was of surgery, is now comparatively rarely practised, except by old-fashioned or questionable practitioners.." (Sir James Crichton Browne, 1892, cf Section 2).

REFERENCES

ANDLAW, R.J. PALMER, J.D., KING, J. & S.B. KNEEBONE. Caries preventive effects of toothpastes containing monoflurophosphate and trimetaphosphate: a 3-year clinical trial. Community Dent. Oral Epidemiol. 11 (1983) 143-147.

ARNOLD, F.A. Jr., DEAN, H.T. & J.W. KNUTSON. Effect of fluoridated public water supplies on dental caries prevalence: results of the seventh year of study at Grand Rapids and Muskegon, Mich. Public Health Rep. 68 (1953) 141-148.

AZZALINI, A. Estimation and hypothesis testing for collections of autoregressive time series. Biometrika 71 (1984) 85-90.

BEALE, E.M.L. & R.J.A. LITTLE. Missing values in multivariate analysis. J.R. Statis. Soc. B 37 (1975) 129-145.

BOX, G.E.P. & D.R. COX. An analysis of transformations (with discussion). J.R. Statis. Soc. B 26 (1964) 211-252.

BUHE, H, BÜTTNER, W. & B. BARLAGE. Über einen dreijährigen klinischen Zahncremetest mit Zahnpasten unterschiedlicher Fluoridkonzentration: 0.8% and 1.2% Natriummonofluorphosphat. Die Quintessenz 35 (1984) 1-9.

CARLOS, J.P. & R.S. SENNING. Error and bias in dental clinical trials. J. Dental Res. 47 (1968) 142-148.

CRICHTON BROWNE, J. Address to the annual general meeting of the Eastern Countries Branch of the British Dental Association. J. Brit. Dental Ass. (1892) 404-416.

DOWNER, M.C., HOLLOWAY, P.J. & T.G.H. DAVIES. Efficiency of clinical trials: selective inclusion of study subjects. J. Dental Res. 56 (1975) C123-C129. (Special Issue C)

ELLIOT SMITH, G & F. WOOD JONES. Archaeological survey of Nubia, 1907-8, Volume II, "Report on the Human Remains". Cairo, 1910.

FDI. Principal requirements for controlled clinical trials of caries preventive agents and procedures. Technical Report no.1. 3rd ed. 1982.

FDI/WHO. Changing patterns of oral health and implications for oral health manpower: part I. Int. Dental J. 35 (1985) 235-251.

GEARY, D.N. Repeated tests on serially correlated data. Oxford, University of Oxford, 1986. D. Phil. Thesis.

GEARY, D.N. Sequential testing in clinical trials with repeated measurements. Biometrika 75 (1988) 311-318.

GREENE, J.C. & J.R. VERMILLION. The simplified oral hygiene index. J. Amer. Dental Ass. 68 (1964) 7-13.

HAND, D.J. & C.C. TAYLOR C.C. Multivariate analysis of variance and repeated measures: a practical approach for behaviourial scientists. London, Chapman and Hall, 1987.

HAWKINS, D.M. A new test for multivariate normality and homoscedasticity. Technometrics 23 (1981) 105-110.

HUNTINGTON, E. Increasing the sensitivity of caries clinical trials by applying logistic modelling to specific teeth. Community Dent. Oral Epidemiol. 13 (1985) 264-267.

JENKINS, G.N. Recent changes in dental caries. Brit. Med. J. 291 (1985) 1297-1298.

JENKINS, W.M.M. The prevention and control of chronic periodontal disease. In: J.J. Murray (ed.) The Prevention of Dental Disease. Oxford, Oxford University Press, 1989. pp. 327-372.

KRAZNOWSKI, W.J. Principles of multivariate analysis: a user's perspective. Oxford, Clarendon Press, 1988.

LÖE, H. & J. SILNESS. Periodontal disease in pregnancy. I - prevalence and severity. Acta Odontol. Scand. 21 (1963) 533-551.

MANDEL, I.D. & A. GAFFAR. Calculus revisited: a review. J. Clinical Periodontol. 13 (1986) 249-257.

MARDIA, K.V. Applications of some measures of multivariate skewness and kurtosis in testing multivariate normality and robustness studies. Sankhya 36B (1974) 115-128.

MARDIA, K.V. Mardia's test of multinormality. In: Encyclopedia of Statistical Sciences, 5 (1985) 1 pp. 217-221.

MARDIA, K.V. & P.J. ZEMROCH. Measures of multivariate skewness and kurtosis. Algorithm AS84. Applied Statistics 24 (1975) 262-265.

MATTILA, K.J., NIEMINEN, M.S., VALTONEN, V.V., RASI, V.P., KESÄNIEMI, Y.A., SYRJÄLÄ, S.L., JUNGELL, P.S., ISOLUOMA, M., HIETANIEMI, K., JOKINEN, M.J. & J.K. HUTTUNEN. Association between dental health and acute myocardial infarction. Brit. Med. J. 298 (1989) 779-781.

McLACHLAN, G.J. & K.E. BASFORD. Mixture models: inference and applications to Clustering. New York, Marcel Dekker, 1988.

MURRAY, J.J. The prevention of dental disease. 2nd ed. Oxford University Press, 1989.

MURRAY, J.J. & L. SHAW. A 3-year clinical trial into the effect of fluoride content and toothpaste abrasivity on the caries inhibitory properties of a dentifrice. Community Dent. Oral Epidemiol. 8 (1980) 46-51.

REED, G.F. & R.B. McHUGH. The estimation of dental cries incidence in the presence of diagnostic error. Biometrics 35 (1979) 473-478.

RUFFER, M.A. Studies in the palaeopathology of Egypt. Chicago, University of Chicago Press, 1921.

SEARLE, S.R., SPEED, F.M. & G.A. MILLIKEN. Population marginal means in the linear model: an alternative to least squares means. Am. Statisn. 34 (1980) 216-221.

SMALL, N.J.H. Testing for multivariate normality. In: Encyclopedia of Statistical Sciences, Vol. 6 (1985) pp. 95-100.

STEPHEN, K.W., CREANOR, S.L, RUSSEL J.L., BURCHELL, C.K., HUNTINGTON, E & C.F.A. DOWNIE. A 3-year oral health dose-response study of sodium monofluorophosphate dentifrices with and without zinc citrate: anti-caries results. Community Dent. Oral Epidemiol. 16 (1988) 321-325.

SPECTRAL ANALYSIS IN CLINICAL AND PUBLIC HEALTH RESEARCH

B. Goldfarb*

* Laboratoire de Biostatistique et d'Informatique Medicale, Hôpital Necker, Paris, France

Abstract

Spectral analysis is presented as a basic tool for research in public health and related fields. This is illustrated by an analysis of mortality rates in France and the disentangling of the circadian and a dialysis-induced rhythm in dialysed patients, using bio-incompatible material. Limits and difficulties in interpreting results are discussed.

Keywords: *spectral analysis, mortality rates, repeated measurements, clinical trials.*

11.1 INTRODUCTION

Widely used in physics or in economics, spectral analysis might be of a considerable interest for biostatisticians concerned by series of time-ordered measures arising frequently from epidemiologic research but sometimes also from clinical research. We will focus on the usefulness of this statistical method in health research (public health and clinical trials).

The evolution or progression of a phenomenon may be appreciated through modelling the time-ordered data, and it might be realized for a disease as well as for well-known health indices or rates (incidence, mortality rate, birth rate, etc...) By means of modelling, one can look for a secular trend, a seasonal variation, a perturbation or a break-point. The final purpose is to identify the constitutive factors on which one would like to act in order to improve the future. In the search of cycles, spectral analysis appears to be the most widely used method; it has been developed on the basis of Fourier analysis, formerly by A. Schuster in the XIXth century to look for hidden periodicities in a deterministic environment. Further developments, mainly due to M.S. Bartlett and R.A. Fisher, lead to its application in a random context.

First we present the main tools of spectral analysis in the search of periodicities. In the second part we focus on the limits and difficulties in interpreting the results. In the two next sections, we describe and discuss briefly two applications. We conclude with short comments on some other developments of spectral analysis.

11.2 STATISTICAL ASPECTS

11.2.1 Spectrum and periodogram

Repeated measurements along the time axis define a time-ordered series of highly correlated data. Such a series must be viewed as one realization of a random process denoted by $\{ X_t \}$. A cross section at a specified time gives a random variable X_t, and a random process is thus a time-ordered collection of random variables, while the observations made at different times on the same unit forms a time series. We will restrict to ourselves random processes verifying the weak stationarity:

$E(X_t) = 0$ and $E(X_t \, . \, X_{t+k}) = \gamma(k)$ only dependent on k, not on t

The underlying phenomenon is supposed to vary along time, and may include periodic components which express is reproducibility. Cross-sections are not suitable for analysing these periodicities since they may not use the information provided by the correlations between observations within the same series. Through these autocorrelations:

$$\gamma(k) = \text{cov}\,(X_t, X_{t+k}) = E(X_t \, . \, X_{t+k})$$

one can recognize the similarities between values lagged by any constant period. The spectral density function, or spectrum, sums up these autocorrelations:

$$f(\omega) = \frac{1}{\pi} \sum_{k=-\infty}^{k=+\infty} c(k).\, e^{-i\omega k}$$

Letting k=0, we see that the area under the curve of the spectrum equals the variance $\sigma_X^2 = \gamma(0)$ of the process (under the assumption of weak stationarity). Thus a peak in the spectrum shows an important contribution to the total variance by the corresponding frequencies, suggesting a cycle whose period (or length) would be equal to the reciprocal of the frequency.

Univariate spectral analysis is concerned by the estimation of the spectral density function, mainly to detect hidden periodicities, using the periodogram:

$$I(\omega_j) = \frac{1}{\pi} \sum_{k=-(N-1)}^{k=(N-1)} c(k).\, e^{-i\omega_j k}$$

where the c(k) is the estimate of the $\gamma(k)$ obtained with the observed time series. To assess the existence of a cycle, tests are built for a null hypothesis of no periodicity. This implies the reference of a process generated by mutually independent random variables, which are supposed to be all normally distributed, centred and with a common variance (weak stationarity). This is the so-called purely discrete random process or white noise. Its autocovariance function is given by:

$$\gamma(k) = 0 \text{ if } k \neq 0$$

$$\gamma(0) = \sigma^2$$

so that its spectrum is constant $f(\omega) = \frac{\sigma^2}{2\pi}$ and obviously has no peak.

The periodogram of such a process is chi-square distributed with two degrees of freedom. It is then easy to test whether the non-constancy of the periodogram of the time series is only due to random sampling or whether it can be attributed to a periodicity. A significance test of a peak in the periodogram has been proposed by R.A. Fisher using the ratio of the maximum ordinate to the sum of all ordinates. This test has strong statistical properties in the case of a unique periodicity. When a significant period has been identified, one would like to know if other cycles remain to be determined. We use the filter obtained by differentiation with order equal to the identified significant period $T_1 : Y_t = X_{t+T_1} - X_t$. The resulting series is a realization of a process presenting the same maxima in the spectrum (thus the same periodicities) except the one used in filtering and may be analyzed through its periodogram. This procedure may be repeated as often as the filtered process is found to be significantly different from a white noise. In univariate spectral analysis, this iterative procedure provides a simple and efficient statistical method to identify cycles. Most of these steps are implemented in some statistical packages. Graphical representations of the periodogram enhance the applications as useful analysis options.

11.2.2 Some limits and difficulties of application

There appear to be at least six points that may lead to misunderstanding and/or confusion when interpreting on the results obtained with univariate spectral analysis. Most of them are taken into account in the two examples presented below.

a) Spectral analysis is built for stationary processes; thus it is necessary to examine the initial series and to apply adapted transformations if hypotheses of weak stationarity are violated.

b) The periodogram is an asymptotically unbiased estimator for the spectral density function, but it is not consistent and it may present important fluctuations in the neighbourhood of a peak, due to the independence of its ordinates. In this context, the use of smoothing and of truncation and weighting has been proposed; one would have to choose between some classical filtering windows.

c) The size of the time series is a limiting factor to the analysis, especially in case of multiple periodicities. If we suppose a series of 36 points (three years of monthly observations) in which a significant trimester-period of length 3 is identified, the application of the differentiating filter leads to a series of only 30 points $N\text{-}2T_1$ in which it will be difficult to look for a half-a-year-period of length 6.

d) Another difficulty when looking for hidden cycles with spectral analysis is aliasing. This problem arises when the sampling rate of the phenomenon is too low, namely less than the double of the highest Fourier frequency. A simple case is given by the two series $x(t) = \cos(2\pi ft)$ and $\gamma(t) = \cos(4\pi ft)$ when the times of sampling are $t = N/f$ (where N is the sample size).

e) Fisher's significance test for a peak in the periodogram is uniformly the most powerful test in case of a unique periodicity. Unfortunately, because the denominator of the statistic being tested may be artificially enlarged by multiple peaks (significant or not), the power decreases with the number of periodicities.

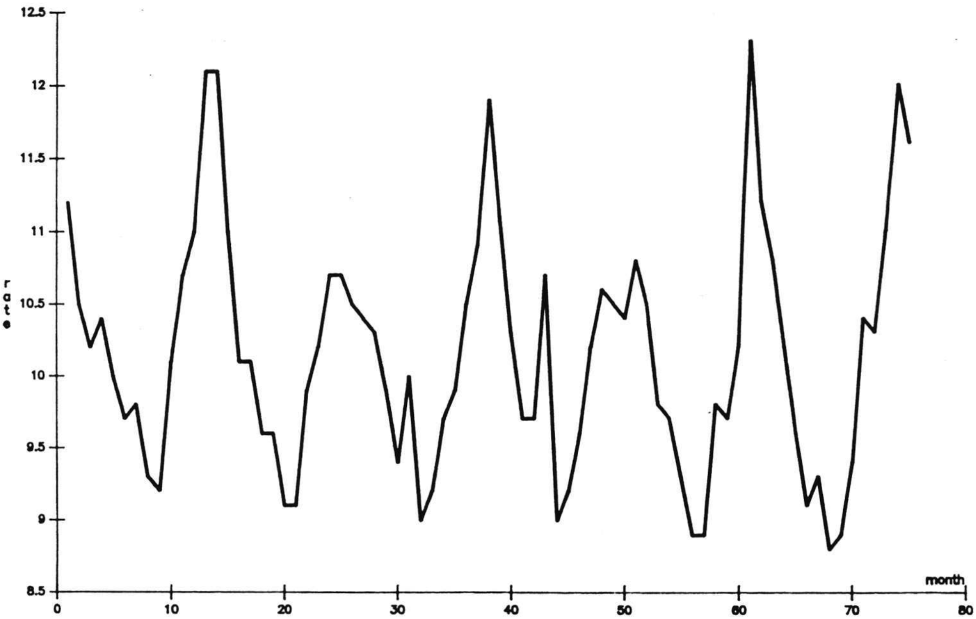
f) In this iterative procedure, only the first test (overall test for a white noise of the initial series) is unconditional. All other tests give a conditional result (given the results of the previous tests), and this must be taken into account in interpreting significancies.

11.3 AN APPLICATION TO THE EVOLUTION OF MORTALITY RATE

The first application of spectral analysis presented here is devoted to monthly mortality rates in France from January 1980 to March 1986, as calculated by the French National

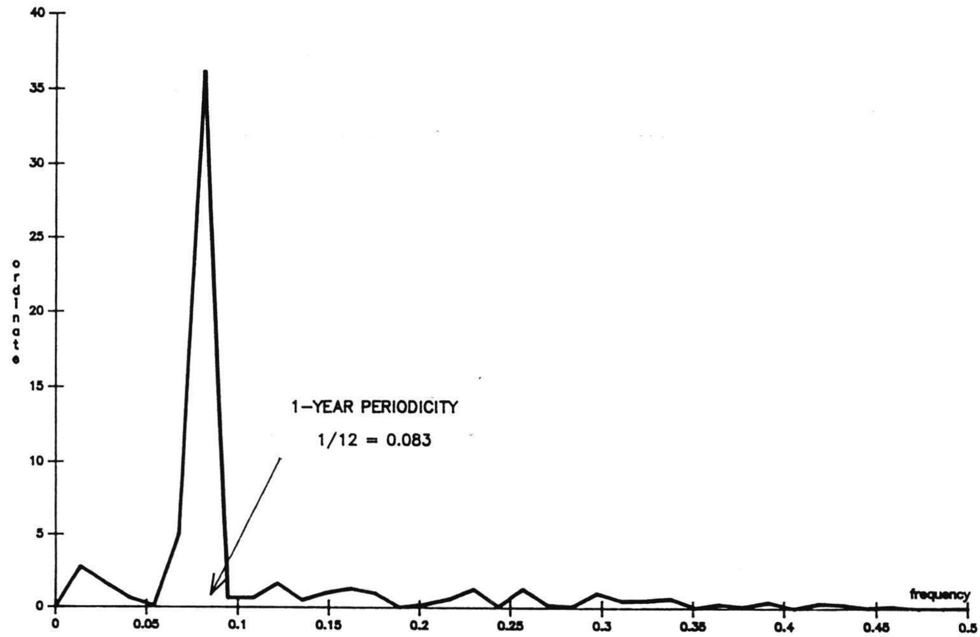
Institute of Statistics (INSEE). This series contains 75 values. When plotting data against time the graph (fig. 1) shows both aspects of stationarity and of periodicities.

Figure 11.1 Mortality rate in France between January 1980 and March 1986



The periodogram obtained on these raw data is plotted on figure 11.2.
It is obviously non-constant (thus confirming the existence of periodicities), and exhibits a strong peak corresponding to a frequency of 0.0833 which is exactly the inverse of 12. Thus the main period is well identified as to the expected one-year period. Fisher's test gives a highly significant result, which is not surprising due to the appearance of the periodogram. Although the periodogram seems to have only one real peak, we apply the differentiating filter with lag 12 to the initial data and compute the periodogram of the filtered series.
When dealing with such situations, we have to keep in mind that the initial series was formed by 75 points, and that only 51 values remained in the filtered series, losing twice the length of filtering.

Figure 11.2 Mortality rate in France, periodogram of raw data

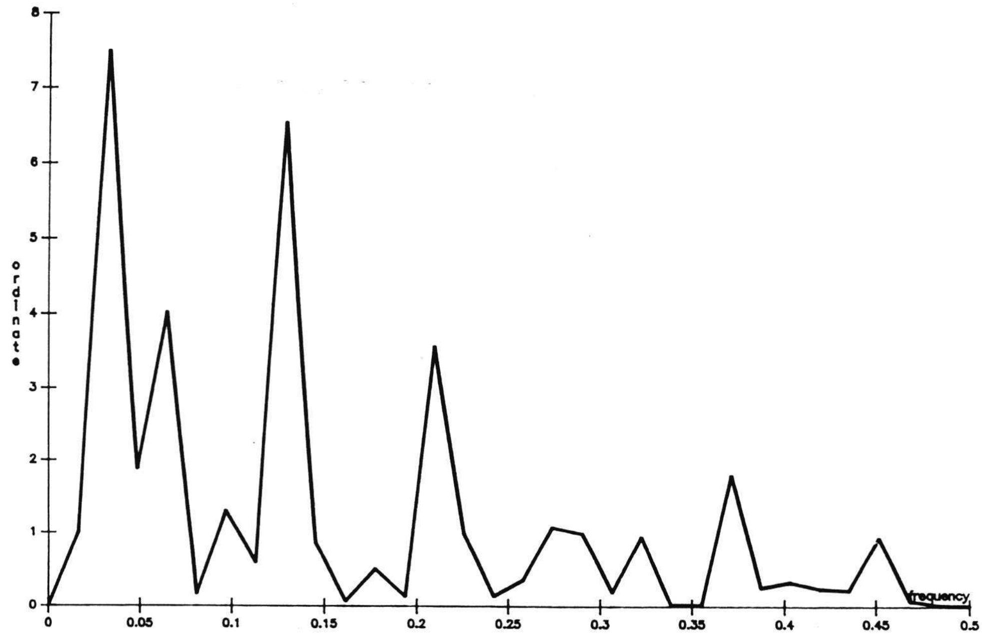


This is plotted against frequencies on figure 11.3, and shows a very irregular aspect with a succession of decreasing peaks.

Should Fisher's test conclude to non-significance for the main peak, this might be the result of the number and importance of other peaks which hardly raise the denominator of Fisher's statistic.

Unfortunately, that series of mortality rates does not allow us to postulate another conclusion than that of a unique periodicity: the periodogram obtained after filtering the 1-year periodicity is not significantly different from a chi square with 2 degrees of freedom. This result may be attributed to an artificial generation of fluctuations in the neighbourhood of a peak and a lack of consistency of the periodogram; there is not enough data (series length and/or sampling rate) to underpin a conclusion.

Figure 11.3 Mortality rate in France periodogram after filtering the 1-year period

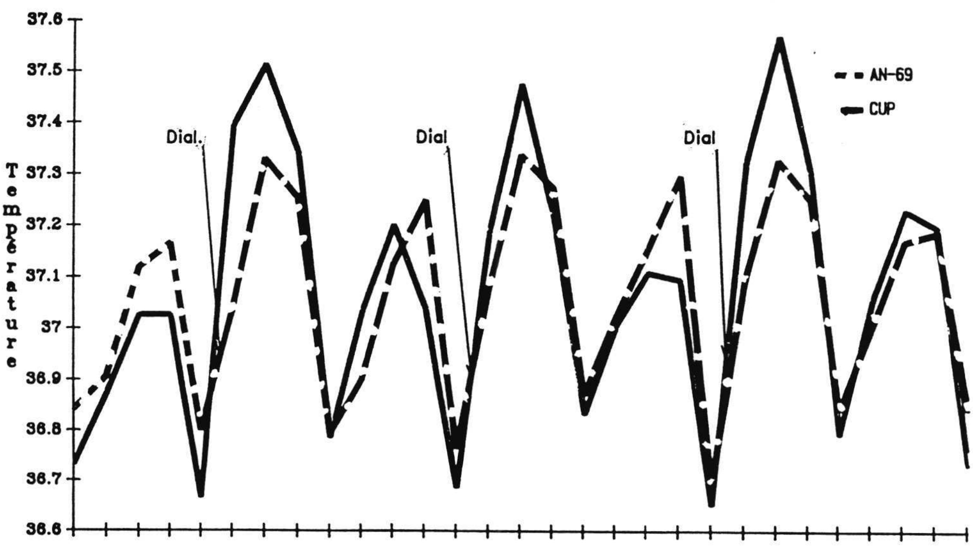


Such a series is not long enough to look for low frequencies (or high periods) and this may explain the discrepancy between the appearance of raw data and the absence of a significant cycle other than the annual rhythm. Moreover we must notice that raw data give an impression of a two-year cycle, which may lead to frequency confounding (or aliasing as defined above). Almost all limits of application of univariate spectral analysis are met in this example.

11.4 AN APPLICATION IN CLINICAL RESEARCH

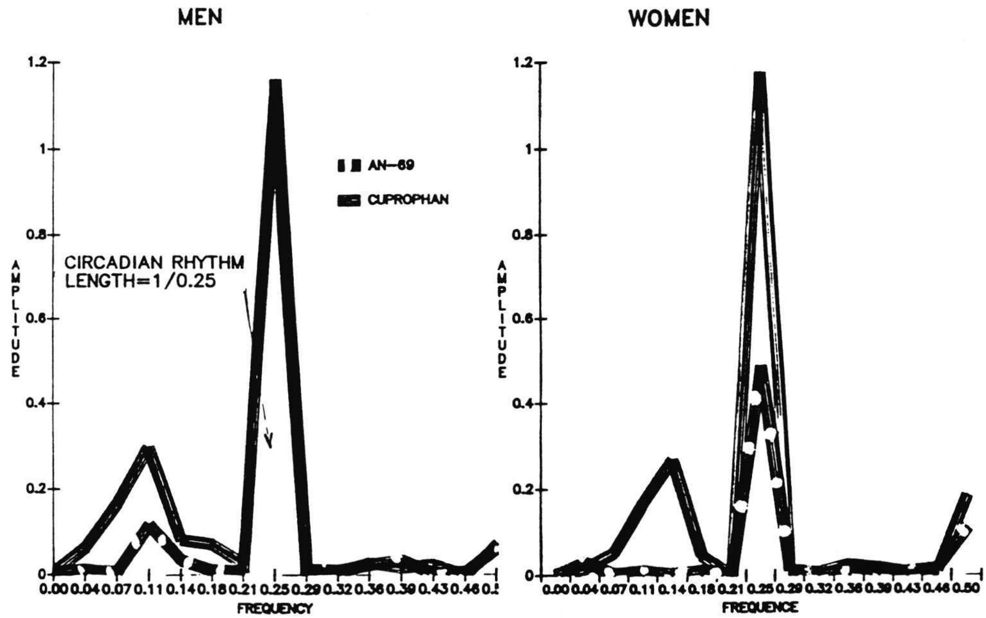
The most frequent use of biomaterials in therapeutics is dialysis whose generalisation to the elderly raises some questions of public health policy. The notion of bio-incompatibility of materials is based on observations within the dialysis session only (interdialytic phases are not taken in account mainly because there is no need of any hospitalisation for dialysed patients between two sessions) and only on laboratory findings on blood samples. We designed a trial to investigate the hypothesis of the existence of a dialysis-membrane induced criteria of thermic dysregulation. The clinical criterion of body temperature was based on the knowledge of biological perturbations allowing for recording of data between dialysis sessions. Two groups of respectively 15 and 17 patients - defined by their type of membrane, cellulosic (CUPROPHAN) or synthetic (AN69) - were matched on different background characteristics (sex, age, initial disease and duration of follow-up in dialysis maintenance). Their rectal temperature was recorded four times daily with a regular schedule, during one week, plus one additional measure, leading to a series of 29 observations.

Figure 11.4 Average rectal temperature on 29 points of time for two group of patients



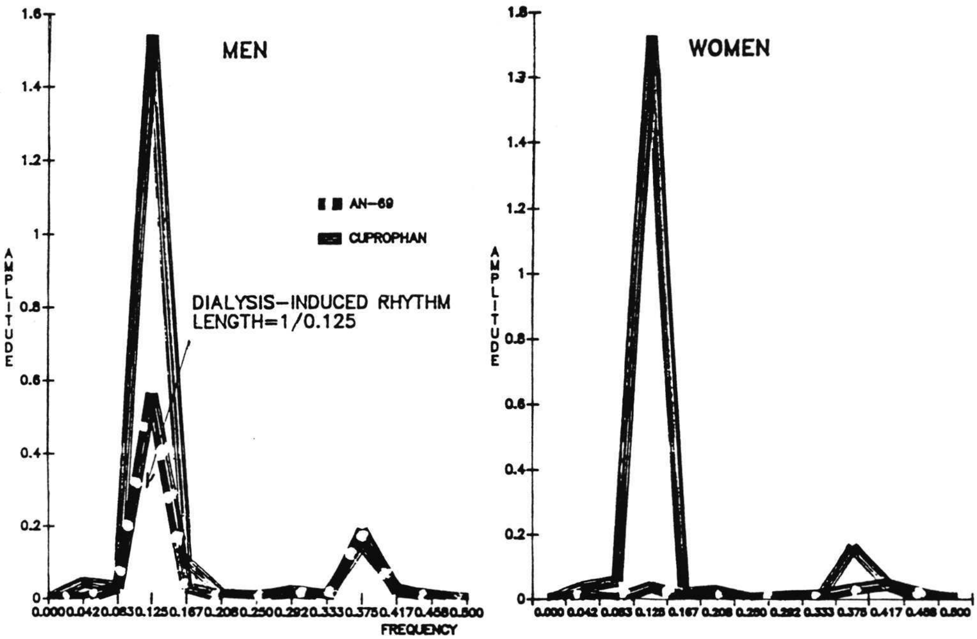
Group curves of temperature plotted in figure 11.4 appear to be sampled from stationary processes, and exhibit the well known circadian rhythm. They show also rises in temperature within the twelve hours following the end of the dialysis session, significantly higher in the CUPROPHAN group than in the AN69 group. The same significant difference is observed for daily thermic amplitudes. Because women undergoing chronic dialysis maintenance are in a hormonal situation identical to early menopause, we analyzed the data for men and women data separately.

Figure 11.5a and 11.5b Men and Women periodogram on raw data



The men's periodogram on raw data plotted in figure 11.5a shows the circadian rhythm (highly significant) with no difference due to the membrane, and a secondary peak to analyze. The women's periodogram (figure 11.5b) shows the circadian rhythm but with a difference explained by the presence of two patients in that subgroup who had pathological conditions (breast neoplasia, and neuroleptic treatment) not revealed during the trial and known to give central thermic disorder. A secondary peak is observed in the CUPROPHAN subgroup of women only.

Figure 11.6a and 11.6b Men and women periodogram after circadian rhythm filtering



After filtering with lag 4 for the daily rhythm, the men's periodogram (figure 11.6a) shows a highly significant peak in the CUPROPHAN group and a less marked and not significant peak in the AN69 group. The frequency equal to 0.125 indicates a period of 8 measures length, i.e. a two-days period corresponding to the rhythm of successive dialysis sessions. None of the other fluctuations is significant. In the women's periodogram after the same filtering (figure 11.6b), we see the same significant 2-days period in the CUPROPHAN subgroup, the series of the AN69 subgroup being not different from a white noise. Although not significant, the main fluctuation observed after filtering in the AN69 subgroup of men is due to three patients that were not always dialysed on the same material; they have been transferred from CUPROPHAN to AN69, but the effects of CUPROPHAN may have lingered for a while. Following a second differentiation with lag 8, none of the residual series obtained could be considered as significantly different from a white noise.

A dialysis-induced rhythm is thus added to the physiologic circadian cycle for patients dialysed on a bio-incompatible material, while on another type of material there is no extra rhythm added. Moreover, this thermic dysregulation appears to be more or less acquired with the time of bio-incompatible maintenance.

11.5 CONCLUSION

Spectral analysis must be viewed as a basic tool in health research and social sciences. Iterative applications of periodogram analysis after differentiation in most cases allow easy identification of the main cycles of a process although some difficulties of interpretation may occur. An other development of spectral analysis is a more sophisticated method allowing study of interrelationship (within a causality context). The univariate case may be considered for detecting outliers.

STATISTICAL ANALYSIS OF POINT OBSERVATIONS FOR DETECTING CLUSTERS OF DISEASE INCIDENCE

P.A. Burrough*

* Netherlands Expertise Center for Spatial Information Processing, Rijksuniversiteit Utrecht, NL

Abstract

Two methods to analyze spatial clustering of rare events are discussed: Schulman's more simple method of density equalized map projections -DEMP- and Openshaw's geographical analysis machine -GAM-. The latter method is promising but still rather experimental. DEMP has some difficulties in relating clusters of rare events to geographical locations and it's use is therefore more confirmatory than exploratory.

Keywords: *data aggregation, rare disease, spatial clustering, data base, GAM, DEMP.*

12.1 INTRODUCTION: The problem of detecting clusters

Rare diseases, such as certain cancers, can be caused by many factors. In older people who have worked under many different conditions, who have travelled and maybe, moved house several times, it is almost impossible to establish exactly any direct link between a geographical source of a potential cancer-inducing factor in the environment and incidence of disease. With rare child cancers, however, the possibility of establishing some kind of link is potentially feasible because children are more likely than adults to have spent most of their lives in one geographical area (a house, street, district).

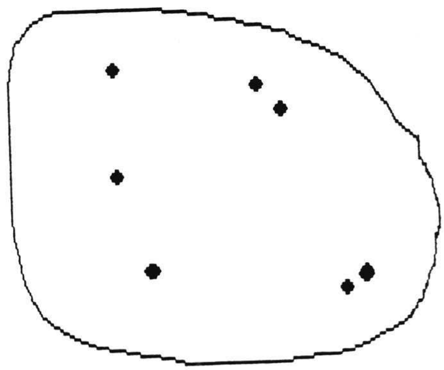
When apparent clusters of rare diseases occur, people are often quick to attribute them to nearby sources of perceived evil - the nuclear reprocessing plant, the garbage incinerator chimney, the dioxene polluted effluent from the garbage tip that drains into the local pond. Before such intuitive and often emotional associations can be accepted as even moderately likely, there are at least two important points to be considered. The first is: is it scientifically feasible that the perceived evil is capable of causing the observed disease ? This is a problem for the medical specialists and I do not propose to go further with it. The second is: is there an unusually high (possibly statistically significant) incidence of the disease at the location under consideration ? The problems of detecting and determining the significance of clusters of rare events is the subject of this paper.

12.2 POINT PATTERN ANALYSIS

When a rare child cancer is detected, the incidence can be recorded as an event occurring at a given location in geographical space (the child's home). In principle the data of all events for a given time period can be seen as a set of all points with coordinates X and Y, with an attribute A covering the geographic area under consideration. If A=0 then the disease did not occur; if A=1, then the disease has been detected. Mapping all points with each point coloured for A=0, and black for A=1 will give a point distribution map which can be analysed by eye (Figure 12.1).

Although such a map may apparently show clusters, we cannot say how significant or important they are unless we know something about the way the data have been collected and stored (in particular the spatial units of aggregation), and the probability of the disease occurring. We should be particularly careful to avoid deriving hypotheses from a data distribution such as Figure 12.1, and then testing them because of bias. It is much better to use an independent test of significance to test for the presence of clusters.

Figure 12.1 Examining random point data by eye only gives a first impression of clustering



12.2.1 Data aggregation

Before the days of electronic databanks it was not possible to record all data about a population with reference to a detailed geographical coordinate system. Therefore data about all aspects of the population were linked to local authority areas, census districts etc. This caused local variations in incidence to be smoothed away. If the number of incidences of a condition in these aggregated units was large, and they were distributed evenly over the unit then the units were a good way of reducing data complexity. Units such as local authority areas or census tracts areas are good ways of displaying the variations in density of numbers of jobless, per capita income, house prices, political views, etc. at a sub-national or regional scale. They are not suitable for displaying the variation of properties that may vary spatially within the aggregated units. For example, a map of Europe showing population densities on a country-by-country basis will not show the clustered population areas of S.E England, Paris, the western Netherlands and the Ruhr.

Clearly, if we wish to determine if there are spatial coincidences between particular kinds of disease and the environment, then it is desirable to map both the places where the disease has been observed and all possible environmental factors that may be associated with that disease. This is an impossible task and so it is sensible first to isolate clusters of disease, and then to search the environment surrounding those areas for possible causal factors. This at least restricts the search to a few areas.

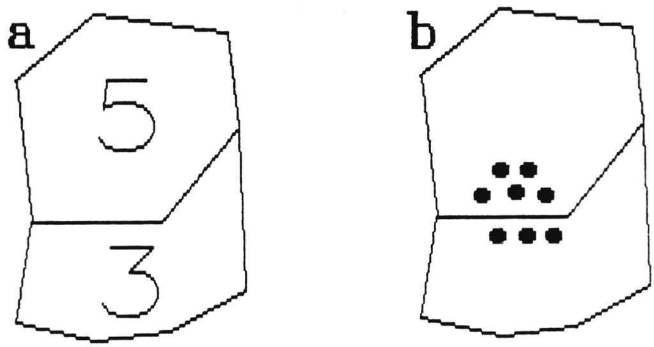
A major problem with this approach lies in the recognition and definition of clusters. When there are sufficient events and these are plotted on a map then clusters can be detected by eye. Many diseases that are suspected to be caused by environmental causes are quite rare, however, and it is by no means a simple matter to decide whether a given number of instances in a certain area constitutes a significant cluster or not. This is a stochastic problem involving the size of the target population, the area in which it is concentrated and the chances that any particular person has of contracting a given disease.

12.3 THE DISTRIBUTION OF THE TARGET POPULATION

Because people are not spread evenly over the earth's surface, determining the geographical location of the target population is far from straightforward. In principle one could map the location of every single person but not only is this very expensive to do, it is considered that such accuracy is offensive and so few data are readily available at this level. It is much more likely that general census information is available for given areas or tracts. Then the estimated population at any given location is a function of both the census data and the method used to map the data.

For example, if a disease is estimated to occur in one person in 10.000, then if two enumeration areas A and B have populations of 100.000 and 20.000 respectively, then there are likely to be 10 and 2 incidences, respectively. If these incidences are smeared out over the whole area as in Figure 12.2a, then no clusters can be detected. If, as in Figure 12.2b, the location of the disease incidence is more precisely known, then it is possible to ask whether or not the incidences are clustered.

Figure 12.2a Reporting rare diseases by administrative area may hide clusters
and
Figure 12.2b If the location is known then clusters can be more easily detected



As data recording and data storage techniques improve it is not longer necessary, nor desirable to aggregate social and epidemiological data before analysis. Many countries are building large, detailed data sets so the problem of spatial aggregation before analysis can be removed. But it is essential to consider spatial aggregation during analysis. A single incidence, tied to a single household, is not a cluster. Ten incidences spread over a city of 500.000 people may be too small to be thought of as a cluster. Clearly we should adopt a spatial unit that is independent of the bureaucratic organization. Langford et al (1990) have demonstrated the effects of different mapping techniques on the resulting estimates of population density as a continuous function rather than as the discretized, lumped model commonly used in coropleth (dasymetric) maps.

When target populations are distributed unevenly over space there are at least two ways in which clusters can be sought for:

1. Transform the geographical space so that the population space is uniform, and then

test for clusters in the population space. Appropriate tests could be the average, minimum or maximum distances to a potential source of influence such as a nuclear reactor as compared to a null hypothesis of no clusters.

2. For areas of a given size in real space find the population and then test if the number of occurrences found exceeds the number that would be expected by chance alone for a given level of significance. Because population distribution is uneven, the analysis is repeated for areas covering a range of sizes in order to investigate how the stability of clusters varies with spatial resolution.

Note that in both cases the following assumptions must be made:

a) The geographical locations of the occurrences of disease are recorded as accurately as possible (post codes, census tract or preferably address) in order to minimize the diluting and distorting effects of lumped areal classification.

b) The population sampled for disease is the population that was exposed to the external cause. If during the period between exposure (e.g. a radiation leak) and appearance of symptoms the population partially migrates, either because of people moving away from the area, people moving into the area, or general interchange, then it will be much more difficult to establish unequivocal results because the "clusters" may diffuse with time

Note that with approach (1) one begins with a suspected source, though this is not strictly necessary. Approach (2) is neutral with respect to sources because it is only looking for clusters that occur with a small statistical probability. Once clusters have been found their location can be compared with that of suspected causal factors in order to explore any possible linkages between the environment and the incidence of the disease.

Approach 1 - the anamorphic map method

This approach recognizes that population density is a continuous function. The method of density equalized map projections - DEMP - (Schulman et al, (1988) eliminates the distortion produced by the geographical distribution of the target population by transforming the geographical area in such a way that it is proportional to population density. The technique has been previously used to analyse voting patterns across large countries. When population is recorded by administrative unit then the areal magnification factor of each unit is given by:

$$M = \frac{pop_A \cdot \sqrt{pop_{total}}}{area_A \cdot \sqrt{area_{total}}}$$

where pop_A, $\sqrt{pop_{total}}$, $area_A$, $\sqrt{area_{total}}$ are the populations and areas of a given polygon (administrative unit) A within a larger total region. This factor ensures that the total area of the map is unchanged by the transformation and that "distance" on the DEMP is proportional to the square root of numbers of persons, not kilometres or miles. Therefore any distance statistics used to detect clusters are working in "people space" rather than in geographical space.

Schulman et al. test for significance of clustering using a distance statistic based on the differences between the locations in the transformed space of the centroids of the whole target population and the cases. The experimentally observed distance is compared with the probability density functions of minimum and maximum distances to a fixed point, thereby allowing a measure of statistical significance to be assigned. In this way hypotheses about the clustering of disease in the population can be tested. For example, suppose that a given target population lives within 30 km of a nuclear plant. If the centroid of the incidence of a specific leukaemia occurs at a distance from the total population centroid that has only a small chance of occurring and the direction of that difference points to the plant, then we may conclude that there may be some influence that needs further investigation.

The advantage of Schulman et al's method is that it is fairly easy to implement on even personal computers and they claim it to be useful even with very sparse data.

Figure 12.3 Some results of anamorphic map analysis for clusters

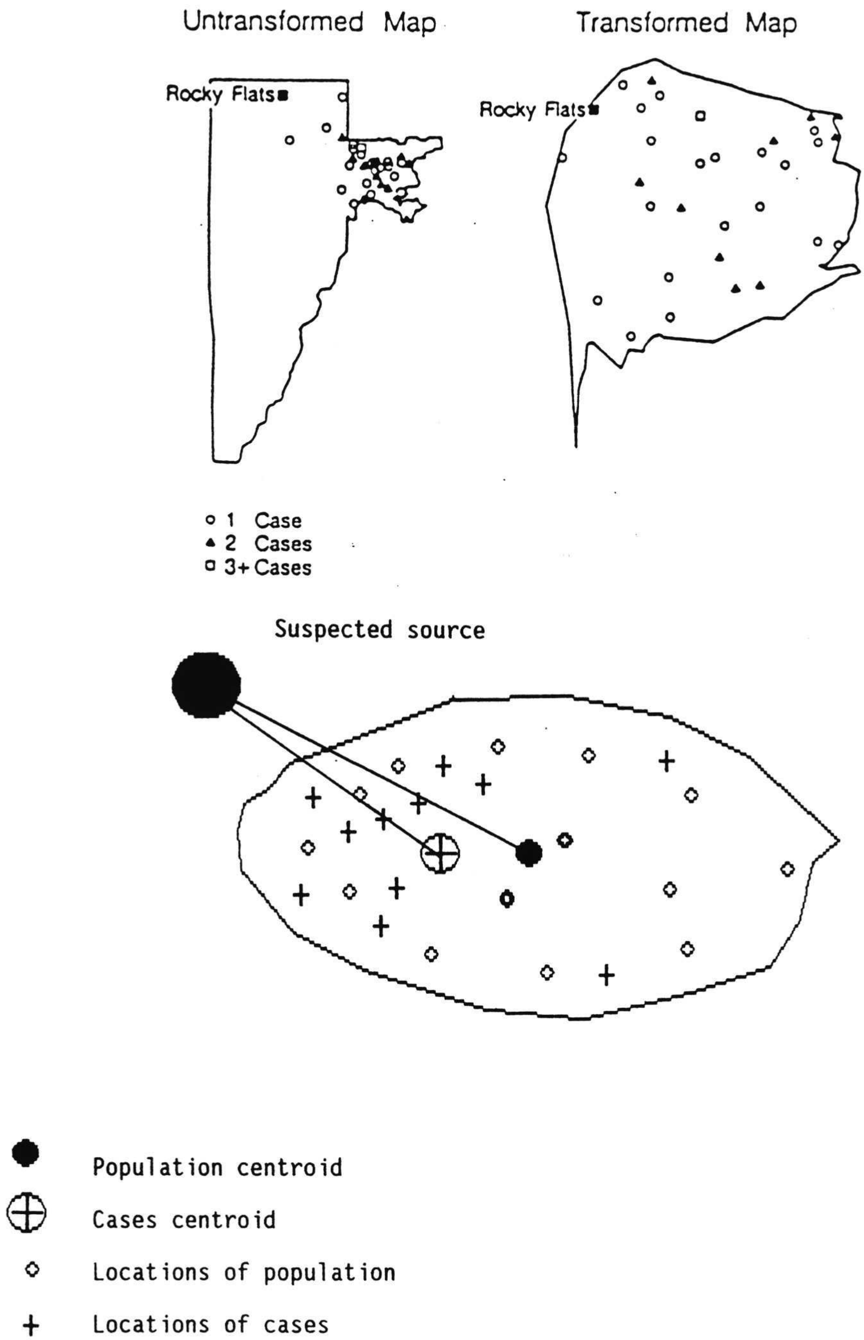


Figure 12.3 gives some results taken from the paper by Schulman et al (1988).

Approach 2 - the Geographical Analysis Machine

Openshaw (1987) has adopted a diametrically opposite approach to the detection of disease clusters. His method requires the geographical location of disease incidence and population to be known and to be held in a database. The method then searches for "clusters" in geographical space, where a cluster is defined as a number of occurrences that are significantly different from the null hypothesis when corrected both for scale and size of local population. The method involves database searches, scale-corrected density adjustments and significance testing, and it makes heavy demands on computer time. The incidence of disease in a population is assessed for circles of different radius. A grid of points is set up over the area in question and a circle of radius r is centred at the first point. The number of incidences falling within the circle and tied to the point. The procedure is repeated for each point on the grid. In order to examine the stability of the "clusters" with respect to geographical area and population, the analysis is repeated for a series of circles of increasing radius. Each circle of larger radius results in a different set of counts. This is computationally tedious, but quite possible on modern computers.

The statistical significance of each count at each grid point for each circle radius can be obtained by comparison with a statistical model expressing the probabilities that the disease would occur purely at random. Although more statistical work needs to be done here. Openshaw (1988) suggests that a Poisson distribution is suitable as the basis for the test statistic. Each count at each grid point is then compared with the test statistic in order to evaluate the likelihood that such a count value occurs by chance alone (Miller & Kahn 1962, p. 380). Values that have a low probability of occurring (say 2 in 1000 or p = 0.002) can be highlighted and displayed as a map.

Figure 12.4 Results of Openshaw's method applied to acute lymphoblastic leukaemia - each circle represents a significance of p= 0.002

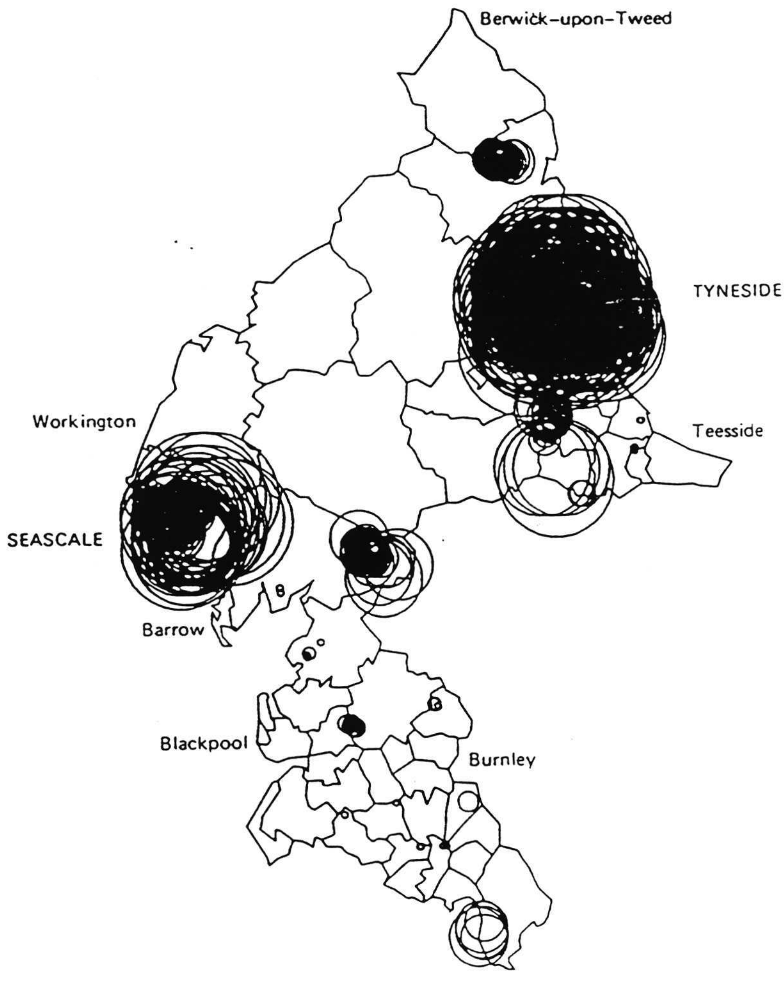


Figure 12.4 shows some results taken from Openshaw (1987). When probability levels do not exceed 0.002, then no circle is drawn. Regions of densely overlapping circles, on the other hand, indicate places where clusters not only occur, but where the clusters are relatively insensitive to spatial aggregation as given by circle diameters. Openshaw considers that these "hot spots" are the real clusters that have been determined independently without reference to external hypotheses or local perceived evils. In his

examples he not only detected the much disputed Sellafield "hotspot" for lymphoblastic leukaemia, but also found a very strong suggestion of a previously unnoticed cluster in the area of north east England known as Tyneside. The presence of such stable clusters allows hypotheses to be set up about the relationship between the cancer incidence and environmental factors which can be examined further.

12.4 CRITICISM OF GAM AND FURTHER DEVELOPMENTS

No method of analysis can be better than the data used, and it is likely that all large data sets may contain the mis-typed data or poorly-referenced site information. In many cases the data used for analysis are only a sample. Unless point counts are related to geographical coordinates there will always be an extra source of error, if for example, locations are recorded by postal codes or other non-exact means. So far, very little work seems to have been done on comparing data collected at different points in time, or of exploring the consequences of using data in which disease incidence changes through time.

The method of significance testing is an area for statistical research. Although the Poisson distribution seems to be a sensible replacement for the Monte Carlo simulation used originally, there is as yet, no unambiguous means of setting up the test statistic. This is clearly an area of further research.

Openshaw (1988) has suggested various developments of GAM, including a super-computer version to speed analysis, and a system that would work automatically, albeit much slower, on an 80383-based personal computer. He has also suggested modifying the spatial search to include certain target populations around each point in order to overcome rural-urban differences. Another alternative is to adjust circle diameters so that they always contain a fixed number of incidents, and then to assess significance. In both cases, simulation is used to handle the multiple significance testing problem.

12.5 CONCLUSION

Two contrasting methods of searching for clusters have been examined. The method of density equalized map projections - DEMP - (Schulman et al, 1988) is simple in concept. Although it seems likely that the method can detect clusters readily, the anamorphic map transformation makes it difficult to relate these clusters to the geographical locations of suspected sources. Consequently, the method may be most useful for testing a priori hypotheses about whether a given suspected site has a significant link with the locations of the incidence of the disease.

Openshaw et al (1987) have demonstrated the construction and use of a methodology for analysing spatial point patterns and determining the significance of spatial clusters. The method is independent of spatial aggregation and is free of intuitively developed hypotheses. The method is still experimental but is worthy of further study because, if proved, it could be of immense value in detecting the spatial occurrence of rare, but potentially avoidable diseases. It would be sensible if funds were made available to conduct research into the feasibility of the methodology in the Netherlands.

REFERENCES

LANGFORD, M., UNWIN, D.J. & D.J. MAGUIRE. Mapping the density of population: continuous surface representations as an alternative to coroplethic and dasymetric maps. Leicester (UK), Midlands Regional Research Laboratory, University of Leicester, 1990. (Research Report Number 8)

MILLER, R.L. & J.S. KAHN. Statistical analysis in the geological sciences. Wiley, New York, 1962.

OPENSHAW, S., CHARLTON, M., WYMER, C., & A. CRAFT. A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets. Int. J. Geographical Information Systems 1 (1987) 335-358.

OPENSHAW, S. Building more refined geographical analysis machines for the automated analysis of cancer data: a review of progress, problems and opportunities. Proceedings of a Congress on Quantitative Applications in Medical Geography, University of Lancaster, September 1988.

SCHULMAN, J. SELVIN, S. & D.W. MERRILL. Density equalized map projections: a method for analysing clustering around a fixed point. Stat. Med. 7 (1988) 491-505.

STATISTICAL METHODS FOR MONITORING RARE HEALTH EVENTS

G. Gallus*

* Instituto di Biometria E Statistica Medica, Facoltà di Medicina e Chirurgia, Università degli Studi, Milan, Italy

Abstract

Health surveillance as a whole represents a fundamental prerequisite not only to be able to detect new environmental risk factors, but also to evaluate the efficacy of health interventions. It will become a topic of growing concern in developed countries.
Given the great variety of health conditions and their mutual relationships, the problem in itself appears to be rather complex. So far surveillance of health events has been considered for very specific problems, but in the future it is expected to become a highly structured system.
After the thalidomide epidemic, congenital malformations represent the first field in which health surveillance procedures have been set up, and in several countries they have been operating for many years.
The main emphasis of this intervention will be on the discussion of the problem from the statistical viewpoint.
The reference procedures remain those proposed for quality control in industrial environments, such as for example the CUSUM scheme, but some procedures, like the SETS method, have been specifically proposed for health applications. The most relevant methods will be briefly presented and properly compared. Their suitability to fit into the health requirements will be finally discussed.

Keywords: *rare health events, CUSUM, SETS, Health surveillance.*

The specific topic of this intervention is part of an important branch of public health. Public health surveillance, which, surprisingly, has received so far a rather limited attention. A brief introduction to the general problem will be given first.

13.1 PUBLIC HEALTH SURVEILLANCE

There isn't yet an agreed definition of the term "health surveillance", but it does imply collection of ongoing routine data to estimate the extent of disease, to follow routine data to estimate the extent of disease, to follow trends and to detect changes in disease occurrence. Important aspects related to surveillance in public health are also the problem of setting up objectives and, particularly, of evaluating intervention programs.
Public health surveillance includes four specific components (Mason, 1989). The first is ongoing, systematic collection of data. Analysis, interpretation and dissemination of the acquired data is the second step, even if often the process stops after collection and analysis. The third component is action based upon a plan to prevent or control a specific disease or adverse condition. Finally there is evaluation of assessment of the process. In spite of the recognition that "surveillance is the cornerstone of public health" (Berkelman, 1989) we have to agree that "the science of surveillance has only recently gone beyond its infancy" (Berkelman, 1989).
One of the most developed fields of application of health surveillance has been so far that of congenital malformations and indeed, after the thalidomide epidemics, population programs have been made operational since the sixties. This is the field to which this intervention will be addressed, and the interest will be mainly focused to the statistical aspects of the problem.

13.2 SURVEILLANCE OF CONGENITAL MALFORMATIONS

As previously mentioned, several countries have set up surveillance programs concerning congenital malformations: at present they are operating at national level in England, in Denmark and in Norway. In 1974 a cooperation international program was activated, the

Clearinghouse for birth defects Monitoring System: it coordinates now 20 centers, corresponding to 3 million births/year (3..). Since 1979 a program is operating in Europe, EUROCAT (De Wals, 1988). This network coordinates European registries, monitoring a total of about 300.000 births per year. Congenital malformation are characterized for being rare diseases and most of what will be discussed will be then relevant to any other infrequent condition besides malformations. Conceptually the problem is not different from that of quality control in industrial environments, since the aim is also there to signal as soon as possible a significant increase in the frequency of unwanted events, i.e. to detect an out of control situation. There are, however relevant differences. Firstly in the industrial case the unit to which the monitoring is to be applied is generally well defined, being normally a machine or a production line. This is not always true in the health application, as the ideal unit of observation should be restricted to the segment of population exposed to a given risk and susceptible to it, like babies of women assuming a specific drug or doing a given job. As it is generally difficult to define a prior which subsets to consider, the surveillance will normally regard the totality of events of a given area, while the unknown risk factors will in general act on only a small fraction of the population. This implies observing effects much more diluted with respect to their real consistency, leading to the need to consider monitoring system particularly efficient in detecting small increases with respect to baseline.

A further information may concern limitations on the available data. Typically they may consist of the number of events occurred in a given area during a given time interval rather than being related to a fixed number of new born babies. Besides the diagnosis of a given malformation may possibly represent a more complex problem than defining a defective item, for its dependency on the diagnostic criteria, on the care put in the visit etc. These added variability sources may make the most usual event distribution assumptions, like the Poisson distribution, inadequate. As a matter of fact in some practical instances the negative binomial distribution has shown to provide a more appropriate event distribution than the Poisson distribution.

The adaption in the medical field of techniques developed for industrial quality control without properly considering the peculiar characteristics of this new field has sometimes resulted in methodological inadequacies (Gallus, 1988). The presentation of the methods will be preceded by a problem formalization, with definition of the optimality criteria.

13.3 STATISTICAL ASPECTS

13.3.1 The formalization of the problem

A crucial point of any monitoring procedure is defining when the system is out of control.

Given X_1, X_2,... observations having a distribution F_0 under H_0, in the reference condition and a distribution F_1 under H_1, in the out of control situation, we check the out of control situation in terms of a stopping rule, which is considered each time an observation occurs. As soon as the stopping rule satisfies a given condition an alarm is given and the system is considered out of control. We define a monitoring system as "optimal" when it minimizes the expected number of observations for the first alarm when the process is out of control, conditioned to a given rate of false alarms. The expected number of observations before an alarm is signalled is called Average Run Length (ARL). The optimality criteria require than a minimum of the ARL under H_0, given a predefined value of ARL under H_1. Three different procedures which can be used for surveillance applications in the health field will be now briefly described. The performances of these techniques will then be properly compared.

13.3.2 The CUSUM scheme

Proposed by Page in 1954, the reference method compares the performance of a surveillance procedure, particularly when small increases from the baseline value are of interest. Let us consider the sequence of successive observation X_i of given size, where each observation provides the number of defective items (or, say malformed babies) in a sample. The system cumulates the differences between successive observations X_i and a value k. When the cumulative sum becomes negative, it is made equal to zero. When it equals or gets larger than a value h, an ALARM is signalled.

$$S_i = \max \{0; S_{i-1} + (X_i - k)\} \qquad i = 1, 2, \ldots$$

N = first integer $i \geq 1$ such as that $S_i \geq h$

ARL = E(N)

The ARL for the CUSUM scheme is given by the solution of an integral equation (Page, 1954) which in general is obtained by resorting to numerical methods. For the setting of the CUSUM procedure in health applications the poisson approximation of the event distribution is normally adopted. In this case an exact solution of the Page equation has been given (Vandeman et al, 1985), but in practice it is convenient to refer to tables, like those produced by Ewan and Kemp (Ewan et al, 1960) or by Lucas (Lucas, 1985).

13.3.3 The "Sets" method

This is an approach (Chen, 1978) which was specifically proposed for health applications, on the ground of being particularly simple to apply even at a local level, without requiring skilled personnel. An "optimal" procedure of parameter determinations was proposed by Gallus et al (1986). Let us define for each malformed birth i one observation X_i given by the distance of the event from the previous one. The distance can be considered in terms of time, or, more properly, of number of new born babies (SET). Each observation is compared with a threshold $T > 0$. An alarm is given when the last n observations are all less than T.
Formally we have:

N = first integer $i \geq n$ such that

$i - 1 - n \leq j \leq i$ for $\max X_i < T$

It can be shown (Gallus et al, 1986) that

$$E(N) = \frac{1 - p^n}{p \ (1-p)^n}$$

where $p = p_1 (x \leq T) = F(T)$

F (x) being the distribution function of the observations. Let us call E_0 and E_1 the ARL under the null and alternative hypothesis respectively.

13.3.4 "Optimal" determination of T and n

Given E_0, the ARL under H_0, which defines the rate of false alarms, and a distribution function F (t) under the alternative hypothesis, the optimal parameter definition can be obtained with the following iterative procedure (Gallus et al, 1986).

1. start with n=2
2. calculate a value of p consistent with the required E_0 (this defines a threshold value T)
3. calculate E_1
4. increase n by 1 and start again from step 2.

Choose the parameters corresponding to the minimum E_1 value. Note that under the Poisson approximation X_i is exponentially distributed, so that
$E_0 (t) = 1 - \exp (- \lambda_0 t)$
where λ_0 measures the malformation rate per time unit under the baseline conditions, and the alternative hypothesis E_1 (t) is specified by the value of $\lambda_1 = \gamma\lambda_0$

13.3.5 The CUSETS method

This is a new procedure, which is being proposed by our group. Considering again the sequence of events (malformed babies) occurring in a given area, we define an observation for each new malformed case given by the sum of the last n consecutive intervals. An alarm is signalled as soon as an observation is below a threshold T. As before, the threshold will be expressed either in terms of time or, more frequently, of the number of normal newborn babies. Note that differently from the previous procedures, consecutive observations in this case are not independent. In order to use the procedure previously reported in 3.4 to "optimally" determine the system, it is required to estimate E(N), the expected delay for the first alarm. Let us assume that the dynamics of the events are well described by a Poisson process. Under this hypothesis, the density probability function of the waiting time X_n for n events has a gamma distribution. It follows that the random variable $Y_n = 2\lambda X_n$ has a Chi-square distribution with 2n degrees of freedom, i.e.

$$P_u = P\ (Y_n \leq x) = P\ (\chi^2_{2n} \leq 2\lambda x).$$

The unconditioned alarm probability P_u, i.e. the probability that, given n events, the waiting time is less than a threshold $2\lambda x$ can than be obtained through the chi-square tables with 2n degrees of freedom. The expected delay is simply given by the reciprocal of the alarm probability. Unfortunately no analytical expression of alarm probability is however available under the monitoring situation, i.e. when the sequence of overlapping observations of n intervals each is considered. The expected delay under these conditions has been obtained through Monte Carlo simulation. The numerical results suggest that, for a given P_u the relationship between the expected delay and n is very well represented by a linear function (minimum $r > .9998$). These numerical results allow to "optimally" determine the value of n and of the threshold T, by the same procedure shown in 3.4, once the alternative hypothesis ($\lambda_1 = \gamma\lambda_0$) and the expected delay under null hypothesis E_0 have been specified.

13.3.6 Comparison of the three methods

The comparison (Table 13.1) is done on the bases of examples taken from tables presented by Ewan and Kemp (1960) which show a number of cusum schemes for a poisson variate.

Table 13.1 Comparison of the three methods in eight different examples

Example number	Example specifications		Results			% variation CUSETS from CUSUM
	γ	E_0	CUSUM E_1	SETS E_1	CUSETS E_1	
1	5.11	95	6.80	5.33	5.16	- 24.0
2	4.11	140	8.06	7.24	6.30	- 21.8
3	2.90	310	12.59	13.78	11.28	- 10.4
4	2.29	590	18.92	24.69	17.68	- 6.5
5	2.00	2620	31.44	52.55	32.87	+ 4.5
6	1.97	975	26.89	40.16	27.93	+ 4.1
7	1.77	4295	45.61	87.63	47.20	+ 3.0
8	1.42	4425	87.97	221.18	103.5	+ 16.6

The first two columns identify each example in terms of the value, which specifies the rate of increase under the alternative hypothesis ($\lambda_1 = \gamma\lambda_0$) and the E_0 value, the expected delay under H_0, which is imposed to be the same for all the three considered techniques. The next three columns report the expected delays for the first alarm for the three methods under comparison and the last indicates the per cent variation of the CUSETS result with respect to CUSUM.

The first observation is that, as expected, the SETS method is always less efficient than CUSETS. Considering then the comparison between these last two methods, the CUSETS approach appears to be more efficient than CUSUM to detect important variations with respect to the baseline (for $\gamma > 2$). For increases of γ below 2, however the CUSUM scheme proves to be more efficient.

13.4 CONCLUSION

The examples above considered seem to lead consistently to the conclusion that when the increase of interest is in the range of $\gamma = 2$ (a 100 per cent increase) the CUSETS method and the CUSUM scheme are basically equivalent. For higher increases the CUSETS appears to be superior, while the CUSUM is preferable when smaller changes are of main concern. The discussion in the literature of the relative merits of the different proposed methods has been so far rather unsatisfactory (Gallus, 1988) and apparently in the existing programs no particular attention has been given to the problem of selecting the most suitable statistical approach. Looking at the results obtained by the working systems in the last years (De Wals, 1986), we have to register a relatively large number of alarms, which have brought to identify therapies, infections, chemical products etc. as factors associated with increased risk of malformations. However none of these "true" alarms emerged as a consequence of a statistical surveillance program. They originated instead from evaluations made informally by carefully observing the collected data. This surprising result may be due in some cases to inefficiency in the procedure of statistical analysis; but in most instances the inefficiency may be perhaps at the design level of the systems themselves. These at the moment work in a rather independent way, without foreseeing any integration of evidence neither in terms of space or level of information, nor of related diseases. To conclude let us stress a first necessary step ahead: the need to a proper and comprehensive evaluation of the existing systems.

REFERENCES

MASON J.O. Opening remarks Symposium on statistics in surveillance. Stat. Med. 8 (1989) 253-254.

BERKELMAN R.L. Summary remarks Symposium on statistics in surveillance. Stat. Med. 8 (1989) 393-395.

International Clearinghouse for Birth Defects Monitoring Systems: 1984 Annual Report. Wellington, New Zealand, V.R. Ward, Government Printer, 1986.

DE WALS P., et al. La surveillance épidemiologigue des anomalies congénitales par le registre EUROCAT, Rev. Epidemiol. Santé Publique 36 (1988) 273-282.

PAGE E.S. Continuous inspection schemes. Biometrika 41 (1954) 100-115.

VANDEMAN S. & D. RAY. Average run lengths for CUSUM schemes when observation are exponentially distributed. Technometrics 27 (1985) 145-150.

EWAN W.D. & K.W. KEMP. Sampling inspection of continuous processes with no autocorrelation between successive results. Biometrika 47 (1960) 363-380.

LUCAS J.M. Counted data CUSUM'S. Technometrics 27 (1985) 129-144.

CHEN R. A surveillance system for congenital malformations. JASA 73 (1978) 323-327.

GALLUS G., MANDELLI C., MARCHI M. & G. RADAELLI. On surveillance methods for congenital malformations. Stat. Med. 5 (1986) 565-571.

GALLUS G., The relative efficiency of the SETS and the CUSUM techniques in monitoring the occurrence of a rare event. (Letter) Stat. Med. 7 (1988) 997-999.

DE WALS P. Investigations of clusters and epidemics of congenital anomalies: A review of published studies. Proc. EUROCAT-ICBDMS Workshop, Florence, December 4-5, 1986.

COMMON STATISTICAL PROBLEMS IN AIDS RESEARCH AND WHY A CO-ORDINATED EFFORT IS THEREFORE REQUIRED

D.A. Lievesley*

* International Statistical Institute, Voorburg, The Netherlands
(Presently at ESRC, DATA ARCHIVE, University of Essex, Wivenhoe Park, Colchester, Essex CO 4 3SQ, England)

Abstract

Many of the difficult questions about AIDS are essentially statistical. The purpose of statistical research on AIDS is to investigate what factors are important, what changes are needed in behaviour and attitudes, and what policies should be adopted, in order to reduce the spread of the infection. Statistical research on AIDS will thus fall into a number of different categories.

Keywords: *AIDS, HIV.*

14.1 INTRODUCTION

There is a great scope for statisticians to make a major contribution to our understanding of AIDS and of the effectiveness of ways to combat it. Many of the important and difficult questions about AIDS are essentially statistical.

The purpose of much statistical research on AIDS is to investigate what factors are important in determining the spread of the disease. This is vital in order to clarify what changes are needed in behaviour and attitudes and what policies should be adopted in order to reduce the spread of the infection.

By way of background information let us examine the size of the problem.

Aids is a nationally, and through the WHO, internationally notifiable disease. At the end of July 1989 172000 cases of AIDS had been reported world-wide from over 170 countries including

99000 from the USA

24000 from Europe

(it is however essential to treat such figures with some caution since the system of reporting and the diligence with which it is conducted differs from country to country despite WHO's admirable efforts to develop standardised procedures).

In Europe the highest number of cases per 100.000 population are in France, Switzerland, Spain, Denmark, Italy, Netherlands, FDR, Great Britain, Austria and Belgium. It is notable that 70% of all cases up to July 89 had been registered in North and South America. The other cases of AIDS were distributed over 41 countries in 2 continents. 47 African countries account for 16% of the world total. To data only a little over 1% of the reported cases were from Asia, Eastern Europe, Near East and North Africa.

In recent years the number of reported cases of AIDS has doubled almost every 8-10 months. In some parts of Central Africa, 15% of the population are already infected with HIV.

Having looked at the regional distribution, we should now examine the heterogeneity within a geographic area. If the AIDS cases are examined according to their 'risk' category, we find that about 60% of cases occur in homosexual or bisexual people, compared to 20% who are intravenous users of drugs and 8% who have contacted AIDS from blood transfusions.

Due to inadequate information about the base population it is difficult to translate these estimates into risks. However it seems that the risks of contracting AIDS amongst the heterosexual "non-risk" group is increasing. It is currently estimated at about 0.02% - though this estimate conceals a large amount of variability. For example in Italy it is estimated at 1.4%. Even within countries there can be considerable heterogeneity. For example in inner London the estimate of risk for the "non-risk" groups is 0.4%, but it is slightly lower (0.3%) for outer London and considerably lower (0.04%) in the other Thames Regions.

14.2 ESTIMATION OF CURRENT INCIDENCE

A major role exists for statisticians in making and improving the estimates of the current incidence of AIDS cases and of the number of people infected but not yet diseased (HIV seropositives), and in determining their distribution across space and subgroups.

In order to do this it is important to establish reliable monitoring systems collecting comparable data on a uniform basis. It is essential to have estimates of the time delay in diagnosis as well as the time lag in registration since in some situations both of these delays may be quite substantial.
Estimates are required of the extent of under-reporting of AIDS cases. Some sources indicate it could be as much as 50% overall (i.e. that there may have been more than 350.000 cases rather than by 172.000 cases by mid 1989). However it is not uniform: There is no doubt that the under-reporting depends upon the system established for reporting, the prominence given to AIDS, the social environment etc. It has been estimated to be less than 20% in the USA for example. The level of under-reporting within a country may not remain constant and thus elaborate longitudinal exercises are required in order to obtain good estimates which are up to date.

Statisticians also have a role to play in research to establish better estimates of AIDS deaths since often AIDS is only a contributory factor and it may not be mentioned on the medical records. Statisticians can also assist in research on the accuracy of tests used to diagnose AIDS cases and to identify those with HIV infection.

The problem of obtaining reliable estimates of the numbers of people with HIV infection remains extremely difficult. It is vital to share information amongst researchers and to pool the skills of different disciplines in order to tackle this problem. Statisticians can assist in setting up sampling schemes, in analysing the results from such samples and in estimating the effects of biases. They can play an advocacy role in order to argue for the resources and support to be devoted to the essential monitoring exercises. The Royal Statistical Society of Great Britain did just this by producing a statement which argued for random testing of blood. At first the statement received a rather mixed reception - many of the seemingly intractable difficulties being essentially ethical rather than statistical - but after debate its recommendations are to be implemented.

The statistical work on AIDS which has had the highest profile is the prediction of the future spread of the disease (Perhaps statisticians would be wise to heed Foedler's rule that forecasting is dangerous especially if it is about the future !). The research predictions has become increasingly complex and sophisticated over time. Initially it began with simple homogeneous mixing models involving just five parameters in modelling the sexual transmission of HIV infection and AIDS. The five parameters were: probability of transmission of infection from infected person to susceptible per partnership, probability that an infected person will develop AIDS, the rate of partner change, the incubation period for those with AIDS and the infectious period for non-AIDS seropositives.

As a greater understanding developed of the disease, the model for predicting AIDS numbers was applied to different subgroups (e.g. to heterosexual separately from homosexuals) and it resulted in a large increase in the numbers of parameters to take account of partnerships of different kinds, assymetries between the sexes, the non-sexual transmission of disease, migration between subgroups etc. etc. More sources of variation between individuals have been taken into account in an attempt to mimic actual behaviour. This increased realism in the models is reflected in an increased complexity.

14.3 IDENTIFYING MODELPARAMETERS

Dietz, for example, has 29 variables in his model, the equations for which involve 42 parameters. Unfortunately the difficulty of identifying and including all the relevant parameters and the marked shortage of information to provide reasonable estimates of level of these parameters, has led to major inaccuracies in the predictions. Compare for example three predictions made for numbers of AIDS patients in the U.K.:

Table 14.1 example for three predictions for numbers of AIDS patients in the U.K.

	Year of prediction		
Prediction for	Late	Mid	Early
Year	1987	1988	1990
1988	1540	1560	(990)
1990	4600	2940	1300
1992	13000	4500	2000

Inaccuracies of this level not surprisingly lead to scepticism about the value of forecasting. It is essential that the accuracy of forecasts should be increased and this depends crucially on obtaining better data as input to the models. I shall return to this later.

Greater understanding can be gained from examining, through simulation, which parameters are a source of sensitivity in the models and which are more robust. Dietz finds, for example that whether or not the epidemic is self sustaining depends critically on the value of the probability of transmission for a single contact between heterosexuals. If it is 0.02 for example the epidemic is self sustaining whereas for a value of 0.01 is not.

This sensitivity analysis can be rewarded since it can indicate where to concentrate efforts on parameter estimation. It could even, depending upon the variable, indicate which parameters any campaign to change behaviour/reduce the chance of infection should focus upon.

14.4 BEHAVIOURIAL ASPECTS

We mentioned the importance of improving the estimates of the parameters. In particular reliable information is required on sexual attitudes, behaviour and knowledge amongst different subgroups (by age, sex, sexual proclivity, etc). Quantitative data on patterns of sexual activity are difficult to acquire and AIDS has highlighted our comparative ignorance in this important area of our lives.

Surveys are needed to collect quantifiable information on the number of new sexual partners of each sex per person, the type and frequency of sexual activity, the duration of sexual partnerships etc. Mean levels of activity are not sufficient and may be misleading. They lead to inappropriate models, poor forecasts and badly focused campaigns. It is important therefore to identify the variability in activity in the population.

The detailed information is not only important to feed into the models for predictive purposes, but coupled with information on attitudes and knowledge it is vital to monitor the impact and effectiveness of public awareness campaigns.

Some evaluation of campaigns has been carried out. For example it has been found that, although the mass media communication programmes can disseminate information very effectively, extensive exposure to the correct information. (For example, over half of the adults in the US surveys in September 1988 thought it was very likely or likely to be transmitted by kissing).

On particularly dangerous misconception is that AIDS is curable: 2/3 of those studied in a survey in Kenya, 1/3 in Zaire and 1/3 in Uganda thought it was curable <u>now</u>. Over 1/3 in Zaire thought there is a vaccine. Few people interviewed knew it could be passed on to a foetus.

This research is ideal for cross country analysis since very different styles of campaigns have been operated. Large, unfocussed campaigns have been used in some countries whereas in others targeted campaigns have been used. In some countries the media campaigns are continuous ('tap dripping') whilst in others one off mass methods have been used. Rather abstract messages have sometimes been employed whereas other campaigns have been direct and explicit. The difficulty in examining the effect of these campaigns lies in the absence of good baseline data collected prior to campaigns.

14.5 IMPACT ON THE SOCIETY

As I have indicated above, statisticians are already making an input to a number of different aspects of the AIDS problem. There is however one area in which they ought to be making a contribution but which has been badly neglected: that is assessing the impact of the AIDS epidemic on society. It is essential to assess the likely burden it will put upon the medical, educational and social services of communities as well as the impact on national economies - particularly due to the loss of people of productive age. There is of course also the impact AIDS will have on society's structure and operation through what has become known as the third epidemic of fear.

The expenditure arising from AIDS will have both direct and indirect costs to the national economy as well as intangible consequences. The direct costs can be divided into personal and non-personal costs. The consumption of medical supplies due to AIDS comprises more than nursing care in hospital or treatment outside of hospital. It also includes the total private and public expenditure on research, money spent on health education, programmes on counselling, support services and so on.

Indirect costs are based on the consideration that persons with AIDS were productively and usefully occupied before falling ill or dying.

In contrast to the economic costs many of the consequences of AIDS are more intangible and cannot be measured by money units. These consequences consist of physical and mental pain, the anxiety pain and distress caused to relatives and friends, a sense of hopelessness and resignation in society, and the impact on the creativity and vigour of the population.

14.6 THE POTENTIAL ROLE OF THE ISI

We are all aware of the challenge of AIDS and what is needed. We have to consider what the ISI, as a worldwide community of statisticians and those interested in statistics, can do. The WHO has already taken a lead in co-ordinating a global programme on AIDS. The ISI should not - and has no wish to - duplicate their role.

We can play a major part by providing the opportunity for statisticians in different areas of the discipline and in different areas of the world to share their methodologies and their results. It is essential that unintended duplication should be avoided: each researcher should build on what has been done elsewhere and should not need to reinvent the wheel. This ensures that important developments can be rapidly absorbed and the alignment of methodologies permits comparability. Deliberate replication is also possible within a collaborative environment.

How can the ISI promote collaboration ? Through our regular conference programme, our journals and newsletters and by perhaps by establishing networks of researchers and holding special meetings.

The ISI and Section membership includes mathematical modellers, statistical computing experts, survey statisticians, official statisticians, epidemiologists, demographers and so on, who all have a role to play in researching the AIDS problem.

The ISI can promote collaborative efforts by enabling data to be shared. We already have extensive experience of data sharing through the ISI Dynamic Data Base which was set up in order to maintain, document, and disseminate data from the World Fertility Survey. The benefits of placing data on such an archive are manifold: researchers can have the benefit of larger samples; they can conduct comparative analysis; the data can be used for training purposes; researchers in relatively isolated positions (e.g. in the developing world) can be supported. It is particularly important to have access to these data from different countries given that the country boundaries are artificial (the spread of infection does not stop at the border) and given that secondary analysts may have a quite different perspective to the original researchers.

ISI's independence and non-political status together with its established reputation for protecting the confidentiality of data whilst maximising its use, make it an ideal base for an archive on AIDS data.
Institutionally, the ISI has considerable experience in carrying out cross country research on difficult sensitive topics, built up through the World Fertility Survey. The ethos of the ISI is to develop the capabilities within countries at the same time as collecting the data, so that the statistical system within countries is strengthened.

The World Fertility Survey, in addition to collecting invaluable data, also produced many guidelines and manuals as well as setting standards for the collection of these data. This model would seem to be an appropriate one for surveys on sexual behaviour and attitudes.

The ISI is also in an excellent position to act as an agency for statistical consultancy since good links have been established with both donor agencies and national statistical offices throughout the world.

The ISI can provide a forum for the discussion of ethical issues which arise in AIDS research due to the sensitivity of information and the importance to maintain confidentiality whilst ensuring that informed decisions are possible. The ISI has produced a Declaration on Statistical Ethics which should prove valuable in such debates and which could be reviewed in the light of this work.

Finally the ISI, like the Royal Statistical Society, can play an advocacy role by acting as an independent non-political pressure group to ensure that resources are devoted to statistical research and relevant data are collected and made available.1

REFERENCES

ISI Declaration of Professional Ethics, International Statistical Institute, Proceedings of the 45 Session 1985, Volume 5, 1985. pp. 319-345.

Royal Statistical Society Meeting on AIDS. J R Stat. Soc. (A) 151 (1988) Part 1

Statistical requirements of the AIDS epidemic statement by the Royal Statistical Society J R Stat. Soc. (A) 151 (1988) Part 1.

ANDERSON R.M et al. It is possible to predict the minimum size of the AIDS epidemic in the UK ? Lancet (i), 1987.

DIETZ, K. Epidemiological Models for sexually transmitted infections. Proceedings 1st World Congress Bernoulli Society, 1987.

MOLLISON, D. Spread of disease through heterogeneous populations. Bulletin of the ISI, Book 1, Paris 47th Session, 1989.

SHASTRI, G.N. AIDS knowledge and attitude in Botswana. Bulletin of the ISI, Book 1, Paris 47th Session, 1989.

INDEX OF KEYWORDS

LIST OF REFERENCES (alphabetical)

AARTS, L.J.M & P.R. DE JONG. Economic aspects of disability behavior. Rotterdam, Erasmus University, 1991. (Dissertation)

ALTMAN, D.G. Statistics in medical journals: developments in the 1980s. Stat. Med. 10 (1991) 1897-1913.

ANDERSON R.M et al. It is possible to predict the minimum size of the AIDS epidemic in the UK ? Lancet (i), 1987.

ANDLAW, R.J. PALMER, J.D., KING, J. & S.B. KNEEBONE. Caries preventive effects of toothpastes containing monoflurophosphate and trimetaphosphate: a 3-year clinical trial. Community Dent. Oral Epidemiol. 11 (1983) 143-147.

ANDRIES, F. Automatiseren is mensenwerk [Automation is the work of humans]. Leiden, NIPG-TNO, 1990.

ARNOLD, F.A. Jr., DEAN, H.T. & J.W. KNUTSON. Effect of fluoridated public water supplies on dental caries prevalence: results of the seventh year of study at Grand Rapids and Muskegon, Mich. Public Health Rep. 68 (1953) 141-148.

AZZALINI, A. Estimation and hypothesis testing for collections of autoregressive time series. Biometrika 71 (1984) 85-90.

ARMSTRONG, D. An outline of sociology as applied to medicine. Wright, Bristol, 1980.

BEALE, E.M.L. & R.J.A. LITTLE. Missing values in multivariate analysis. J.R. Statis. Soc. B 37 (1975) 129-145.

BEBBINGTON, A.C. The expectation of life without disability in England and Wales. Soc. Sci. Med. 27 (1988) 321-326.

BERKELMAN R.L. Summary remarks Symposium on statistics in surveillance. Stat. Med. 8 (1989) 393-395.

BLOEMHOFF, A. & C.R. DE WINTER. De invloed van sociaal economische status op arbeidsongeschiktheid: een longitudinale analyse. Leiden, NIPG-TNO, 1990.

BOX, G.E.P. & D.R. COX. An analysis of transformations (with discussion). J.R. Statis. Soc. B 26 (1964) 211-252.

BRAND, J & RADDER J.J. Betrouwbaarheidsintervallen voor somscores van dichotome items in V-Prof, Department of Statistics and Computer Science, Leiden, NIPG-TNO, 1992.

BUCK, S.F. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. J. R. Statis. Soc. B22 (1960) 302-306.

BUHE, H, BÜTTNER, W. & B. BARLAGE. Über einen dreijährigen klinischen Zahncremetest mit Zahnpasten unterschiedlicher Fluoridkonzentration: 0.8% and 1.2% Natriummonofluorphosphat. Die Quintessenz 35 (1984) 1-9.

BURG, E. VAN DER. CANALS. Leiden, University of Leiden, Department of Data Theory, 1985.

BUUREN, VAN, S. & W.J. HEISER. Clustering *n* objects into *k* groups under optimal scaling of variables. Psychometrika 54 (1989) 699-706.

BUUREN, VAN, S. & J.L.A. VAN RIJCKEVORSEL. Fast linear least squares imputation of missing data. RR-01-91, Leiden, University of Leiden, Dept. of Psychology, 1991.

BUUREN, VAN, S. & J.L.A. VAN RIJCKEVORSEL. Imputation of missing categorical data by maximizing internal consistency. Psychometrika 1992A. (to appear)

BUUREN, VAN, S. & J.L.A. VAN RIJCKEVORSEL. Data augmentation and optimal scaling. To appear in: R. Steyer & K.F. Wender (eds) Proceedings of the 7th European Meeting of the Psychometric Society, July 29-31, Trier. Heidelberg, Springer-Verlag, 1992b.

CARLOS, J.P. & R.S. SENNING. Error and bias in dental clinical trials. J. Dental Res. 47 (1968) 142-148.

CENTRAAL BUREAU VOOR DE STATISTIEK (CBS). Automatiseringsstatistieken. Particuliere sector [Automation statistics. Private sector]. 1989. 's-Gravenhage, CBS, 1989.

CHEN R. A surveillance system for congenital malformations. JASA 73 (1978) 323-327.

CLARK, D.W. A Vocabulary for preventive and community medicine. In: Clark, D.W. & B. MacMahon, Preventive and Community Medicine. Boston, Little, Brown and Co., 1981. pp. 3-15.

CLOGG, C.C. & S.R. ELIASON. Some common problems in log-linear analysis. In: J. Scott Long (ed) Common problems/proper solutions: avoiding error in quantitative research. Newbury Park etc., Sage, 1988, pp. 226-257.

COLVEZ, A. & M. BIANCHET. Potential gains in life expectancy free of disability: a tool for health planning. Intern. J. Epidemiol. 12 (1983) 224-229.

COUGER, J.D. & R.A. ZAWACKI, Motivating and managing computerpersonnel. Chichester, Wiley, 1980.

COX, T. The nature and measurement of stress. Ergonomics 28 (1985) 1155-1163.

CRICHTON BROWNE, J. Address to the annual general meeting of the Eastern Countries Branch of the British Dental Association. J. Brit. Dental Ass. (1892) 404-416.

CRONBACH, L.J. Coefficient alpha and the internal structure of tests. Psychometrika 16 (1951) 297-334.

DEAR, R.E. A principal component missing data method for multiple regression models, SP-86. Santa Monica Cal., System Development Corporation, 1959.

DEMPSTER, A.P. & D.B. RUBIN. Overview. In: W.G. Madow, I. Olkin & D.B. Rubin (eds) Incomplete data in sample surveys, vol. 2, Theory and annotated bibliography. New York,
Academic Press, 1983. pp. 3-10.

DE WALS P., et al. La surveillance épidemiologigue des anomalies congénitales par le registre EUROCAT, Rev. Epidemiol. Santé Publique 36 (1988) 273-282.

DE WALS P. Investigations of clusters and epidemics of congenital anomalies: A review of published studies. Proc. EUROCAT-ICBDMS Workshop, Florence, December 4-5, 1986.

DIETZ, K. Epidemiological Models for sexually transmitted infections. Proceedings 1st World Congress Bernoulli Society, 1987.

DIJKSTRA, A., GRINTEN, VAN DER, M.P., SCHLATMANN, M.J.Th. & C.R. DE WINTER. Functioneren in de arbeidssituatie; uitgangspunten, ontwerp en handleiding voor onderzoek onder werknemers naar gezondheid, werk en werkomstandigheden. [Functioning at the workplace; starting points, design and guidelines for research amongst employees into health, work and working conditions]. Leiden, NIPG-TNO, 1986.

DODGE, Y. Analysis of experiments with missing data. New York, Wiley, 1985.

DOWNER, M.C., HOLLOWAY, P.J. & T.G.H. DAVIES. Efficiency of clinical trials: selective inclusion of study subjects. J. Dental Res. 56 (1975) C123-C129. (Special Issue C)

ELLIOT SMITH, G & F. WOOD JONES. Archaeological survey of Nubia, 1907-8, Volume II, "Report on the Human Remains". Cairo, 1910.

EVANS, J. Women, men, VDU-work and health: a questionnaire survey of British VDU-operators. Work and Stress 1 (1987) 271-283.

EWAN W.D. & K.W. KEMP. Sampling inspection of continuous processes with no autocorrelation between successive results. Biometrika 47 (1960) 363-380.

FDI. Principal requirements for controlled clinical trials of caries preventive agents and procedures. Technical Report no.1. 3rd ed. 1982.

FDI/WHO. Changing patterns of oral health and implications for oral health manpower: part I. Int. Dental J. 35 (1985) 235-251.

FLEISS, J.L. Statistical methods for rates and proportions. New York, Wiley, 1981.

FORD, B.L. An overview of hot deck procedures. In: W.G. Madow, I. Olkin & D.B. Rubin (eds), Incomplete data in sample surveys, vol. 2, Theory and annotated bibliography. New York, Academic Press,
1983.

FRENCKEN, J.E., VAN 'T HOF, M.A., TRUIN, G.J., LEMBARITI, B.S. & K.G. KÖNIG. Cohort effects in the prevalence of caries in child populations in Tanzania. J. Dental Res. 68 (1989) 1777-1780.

GALLUS G., MANDELLI C., MARCHI M. & G. RADAELLI. On surveillance methods for congenital malformations. Stat. Med. 5 (1986) 565-571.

GALLUS G., The relative efficiency of the SETS and the CUSUM techniques in monitoring the occurrence of a rare event. (Letter) Stat. Med. 7 (1988) 997-999.

GEARY, D.N. Repeated tests on serially correlated data. Oxford, University of Oxford, 1986. D. Phil. Thesis.

GEARY, D.N. Sequential testing in clinical trials with repeated measurements. Biometrika 75 (1988) 311-318.

GEER, J.P. VAN DE. Analyse van kategorische gegevens. Deventer, Van Loghum Slaterus, 1988.

GIFI, A. Nonlinear multivariate analysis. Chichester, Wiley, 1990.

GINNEKEN, J. VAN, BANNENBERG, A. & A. DISSEVELT. Loss of Health as Result of a Number of Important Groups of Diseases in The Netherlands in 1981-1985: Methodological Aspects and Results (in Dutch). Internal report. Leiden and Voorburg, Netherlands Institute of Preventive Health Care and Netherlands Central Bureau of Statistics, 1989.

GLEASON, T.C. & R. STAELIN. A proposal for handling missing data. Psychometrika 40 (1975) 229-252

GREENE, J.C. & J.R. VERMILLION. The simplified oral hygiene index. J. Amer. Dental Ass. 68 (1964) 7-13.

GREENACRE, M.J. Theory and applications of correspondence analysis. New York, Academic Press, 1984.

HAND, D.J. & C.C. TAYLOR C.C. Multivariate analysis of variance and repeated measures: a practical approach for behaviourial scientists. London, Chapman and Hall, 1987.

HARTLEY, H.O. & R.R. HOCKING. The analysis of incomplete data. Biometrics 27 (1971) 783-808.

HAWKINS, D.M. A new test for multivariate normality and homoscedasticity. Technometrics 23 (1981) 105-110.

HAWKINS, A. Healthy Parastatisticians ? News & Notes, The Royal Statistical Society 18 (1992) 1-2.

HEDGES, B. & I. OLKIN. Selected annotated bibliography. In: W.G. Madow, I. Olkin & D.B. Rubin (eds) Incomplete data in sample surveys, Vol. 2, Theory and annotated bibliography. New York, Academic Press, 1983. pp. 3-10.

HOF, M.A. VAN 'T, PRAHL-ANDERSEN, B. & C.J. KOWALSKI. A model for the study of developmental processes in dental research. J. Dental Res. 55 (1976) 359-366.

HUNTINGTON, E. Increasing the sensitivity of caries clinical trials by applying logistic modelling to specific teeth. Community Dent. Oral Epidemiol. 13 (1985) 264-267.

International Clearinghouse for Birth Defects Monitoring Systems: 1984 Annual Report. Wellington, New Zealand, V.R. Ward, Government Printer, 1986.

ISI Declaration of Professional Ethics, International Statistical Institute, Proceedings of the 45 Session 1985, Volume 5, 1985. pp. 319-345.

JENKINS, G.N. Recent changes in dental caries. Brit. Med. J. 291 (1985) 1297-1298.

JENKINS, W.M.M. The prevention and control of chronic periodontal disease. In: J.J. Murray (ed.) The Prevention of Dental Disease. Oxford, Oxford University Press, 1989. pp. 327-372.

KALSBEEK, H. & G.H.W. VERRIPS. Dental caries prevalence and the use of fluorides in different European countries. J. Dental Res. 69 (1990) Special issue 728-732.

KALTON, G. & D. KASPRZYK. Imputing for missing survey responses. Proceedings of the Section of Survey Research Methods, 1982, American Statistical Association, 1982. pp. 22-33.

KLEINBAUM, D.G., KUPPER, L.L. & H. MORGENSTEIN. Epidemiologic research, principles and quantitative methods. New York, Van Nostrand Reinhold, 1982.Netherlands Central Bureau of Statistics. Health Interview Survey 1981-1985. The Hague, Staatsuitgeverij, 1988.

KRAZNOWSKI, W.J. Principles of multivariate analysis: a user's perspective. Oxford, Clarendon Press, 1988.

LANGFORD, M., UNWIN, D.J. & D.J. MAGUIRE. Mapping the density of population: continuous surface representations as an alternative to coroplethic and dasymetric maps. Leicester (UK), Midlands Regional Research Laboratory, University of Leicester, 1990. (Research Report Number 8)

LILIENFELD, A.M. & D.E. LILIENFELD. Foundations of epidemiology. New York, Oxford University Press, 1980.

LITTLE, R.J.A. & D.B. RUBIN. Statistical analysis with missing data. New York, Wiley, 1987.

LITTLE, R.J.A. & D.B. RUBIN. The analysis of social science data with missing values. In: J. Fox & T. Scott Long (eds) Modern methods of data analysis. London, Sage, 1990. pp. 374-409.

LÖE, H. & J. SILNESS. Periodontal disease in pregnancy. I - prevalence and severity. Acta Odontol. Scand. 21 (1963) 533-551.

LORD, F.M. Some relations between Guttman's principal components of scale analysis and other psychometric theory. Psychometrika 23 (1958) 291-296.

LUCAS J.M. Counted data CUSUM'S. Technometrics 27 (1985) 129-144.

MADOW W.G., OLKIN I. & D.B. RUBIN (eds). Incomplete data in sample surveys, Vol. 1, 2 and 3. New York, Academic Press, 1983.

MANDEL, I.D. & A. GAFFAR. Calculus revisited: a review. J. Clinical Periodontol. 13 (1986) 249-257.

MARDIA, K.V. Applications of some measures of multivariate skewness and kurtosis in testing multivariate normality and robustness studies. Sankhya 36B (1974) 115-128.

MARDIA, K.V. Mardia's test of multinormality. In: Encyclopedia of Statistical Sciences, 5 (1985) 1 pp. 217-221.

MARDIA, K.V. & P.J. ZEMROCH. Measures of multivariate skewness and kurtosis. Algorithm AS84. Applied Statistics 24 (1975) 262-265.

MASON J.O. Opening remarks Symposium on statistics in surveillance. Stat. Med. 8 (1989) 253-254.

MATTILA, K.J., NIEMINEN, M.S., VALTONEN, V.V., RASI, V.P., KESÄNIEMI, Y.A., SYRJÄLÄ, S.L., JUNGELL, P.S., ISOLUOMA, M., HIETANIEMI, K., JOKINEN, M.J. & J.K. HUTTUNEN. Association between dental health and acute myocardial infarction. Brit. Med. J. 298 (1989) 779-781.

McLACHLAN, G.J. & K.E. BASFORD. Mixture models: inference and applications to Clustering. New York, Marcel Dekker, 1988.

MEULMAN, J. Homogeneity analysis of incomplete data. Leiden, DSWO Press, 1982.

MILLER, R.L. & J.S. KAHN. Statistical analysis in the geological sciences. Wiley, New York, 1962.

MINTZBERG, H. The nature of managerial work. New York, Harper & Row, 1973.

MOLLISON, D. Spread of disease through heterogeneous populations. Bulletin of the ISI, Book 1, Paris 47th Session, 1989.

MURRAY, J.J. The prevention of dental disease. 2nd ed. Oxford University Press, 1989.

MURRAY, J.J. & L. SHAW. A 3-year clinical trial into the effect of fluoride content and toothpaste abrasivity on the caries inhibitory properties of a dentifrice. Community Dent. Oral Epidemiol. 8 (1980) 46-51.

NISHISATO, S. Analysis of categorical data: Dual scaling and its applications. Toronto, University of Toronto Press, 1980.

OBORNE, D.J. Computers at work. Chichester, Wiley, 1985.

OPENSHAW, S., CHARLTON, M., WYMER, C., & A. CRAFT. A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets. Int. J. Geographical Information Systems 1 (1987) 335-358.

OPENSHAW, S. Building more refined geographical analysis machines for the automated analysis of cancer data: a review of progress, problems and opportunities. Proceedings of a Congress on Quantitative Applications in Medical Geography, University of Lancaster, September 1988.

PAGE E.S. Continuous inspection schemes. Biometrika 41 (1954) 100-115.

POT, F.D., PADMOS, P. & A.A.F. BROUWERS. Achter de schermen [Behind the screens].
's-Gravenhage, Ministerie van Sociale Zaken en Werkgelegenheid, 1986.

PUTTEN, VAN D.J. De vragenlijst voor Periodiek Bedrijfsgeneeskundig Onderzoek. Leiden, NIPG-TNO, 1989.

REED, G.F. & R.B. McHUGH. The estimation of dental cries incidence in the presence of diagnostic error. Biometrics 35 (1979) 473-478.

REGTERING, H. & B. RIESEWIJK. ICON onderzoek Nederland [ICON research in the Netherlands] Nijmegen, International Computer Occupations Network, 1987.

RIESEWIJK, B. & J. WARMERDAM. Het slagen en falen van automatiseringsprojecten [The success and failure of automation projects]. Nijmegen, Katholieke Universiteit, Instituut voor Toegepaste Sociologie, 1988.

ROBINE, J.M., BROUARD, N. & A. COLVEZ. Les indicateurs d'espérance de vie sans incapacité (EVSI). Des indicateurs globaux de l'état de santé des populations. Rev. Epidemiol. Santé Publique 35 (1987) 206-224.

ROTHMAN, K.J. Modern epidemiology. Boston/Toronto, Little, Brown & Co., 1986.

ROSEN, G. Public Health. International encyclopedia of the social sciences. New York, London, Vol 13. (1972) pp. 164-170.

Royal Statistical Society Meeting on AIDS. J R Stat. Soc. (A) 151 (1988) Part 1.

RUBIN, D.B. Multiple imputation for nonresponse in surveys. New York, Wiley, 1987.

RUBIN, D.B. EM and beyond. Psychometrika 56 (1991) 241-254.

RUBIN, D.B. & N. SCHENKER. Multiple imputation in health-care databases: An overview and some applications. Stat. Med. 10 (1991) 585-589.

RUFFER, M.A. Studies in the palaeopathology of Egypt. Chicago, University of Chicago Press, 1921.

SAS Technical Report P-179. Cary, NC, SAS Institute Inc.

SCHNELL, R. Missing Data Probleme in der empirischen Sozialforschung [Missing data problems in the social sciences]. Inaugural dissertation, Bochum, Ruhr Universität, 1986.

SCHULMAN, J. SELVIN, S. & D.W. MERRILL. Density equalized map projections: a method for analysing clustering around a fixed point. Stat. Med. 7 (1988) 491-505.

SEARLE, S.R., SPEED, F.M. & G.A. MILLIKEN. Population marginal means in the linear model: an alternative to least squares means. Am. Statisn. 34 (1980) 216-221.

SHASTRI, G.N. AIDS knowledge and attitude in Botswana. Bulletin of the ISI, Book 1, Paris 47th Session, 1989.

SONSBEEK, J.L.A. Methodological aspects and content of the OECD-long-term indicator (in Dutch). Maandbericht Gezondheidsstatistiek 6 (1988) 4-17.

SMALL, N.J.H. Testing for multivariate normality. In: Encyclopedia of Statistical Sciences, Vol. 6 (1985) pp. 95-100.

SPRUIT, I.P. Social factors and disease in epidemiology and medical sociology: some foundations of measurements. In: Tengvald K. (ed.). Problem of Applied Epidemiology: an inner and outer perspective. F.R.N., Stockholm, in press.

Statistical requirements of the AIDS epidemic statement by the Royal Statistical Society J R Stat. Soc. (A) 151 (1988) Part 1.

STEPHEN, K.W., CREANOR, S.L, RUSSEL J.L., BURCHELL, C.K., HUNTINGTON, E & C.F.A. DOWNIE. A 3-year oral health dose-response study of sodium monofluorophosphate dentifrices with and without zinc citrate: anti-caries results. Community Dent. Oral Epidemiol. 16 (1988) 321-325.

SULLIVAN, D.F. Disability components for an index of health. Washington, U.S. Government Printing Office, 1971a. (PHS Publication No. 1000 Series 2, No. 42)

SULLIVAN, D.F. A single index of mortality and morbidity. HSMH Health Reports 86 (1971b) 347-354.

VANDEMAN S. & D. RAY. Average run lengths for CUSUM schemes when observation are exponentially distributed. Technometrics 27 (1985) 145-150.

VOGELS, T. & DANZ, M. Gezondheid, gedrag en relaties: Draaiboek. [Health, behaviour and relations: Scenario.]. Leiden, NIPG-TNO, 1989.

VOGELS, T., KLAAUW VAN DER, M.M., & M. VAN LAARHOVEN. Werken met APRI: handleiding en procedures. [Using APRI: manual and procedures.] Leiden, NIPG-TNO, 1989. (internal report)

VOGELS, T., VLIET VAN DER, R.W.F., DANZ, M.J. & M. HOPMAN-ROCK. Verslag vooronderzoek Gezondheid, Gedrag en Relaties: een onderzoek van de Peilstations Jeugdgezondheidszorg [Report preliminary research Health, Behaviour and Relations: a study by the Sentinel Stations for Youth Health Care]. Leiden, NIPG-TNO/NISSO, 1990.

WERKGROEP Sociaal Beleid van de stuurgroep actieplan Computer Service Industrie. Sociaal beleid computer service industrie [Committee on Social Policy of the steering committee Computer Service Industry. Social policy computer service industry].'s-Hertogenbosch, NEHEM, 1988.

WILKENS, R. & O.B. ADAMS. Health expectancy in Canada, late 1970s: demographic, regional and social dimensions. Am. J. Public Health 73 (1983) 1073-1080.

WINSLOW, C.E.A. The untilled fields of public health. Science 51 (1920) 23.

WORTEL, E., W.T.M. OOIJENDIJK & I. STOMPEDISSEL Preventie van privé-ongevallen bij 0-4 jarigen [Prevention of preschool children's accidents at home]. Leiden, NIPG-TNO, 1988.

Reprografie NIPG-TNO
Projectnummer 5038