

BERTIN, LEXIS AND THE GRAPHICAL REPRESENTATION OF EVENT HISTORIES

By Brian FRANCIS and John PRITCHARD

Centre for Applied Statistics, Fylde College, Lancaster University,
Lancaster, LA1 4YF, U.K.

e-mail: B.Francis@Lancaster.ac.uk

Abstract

After first examining the issues in visualising individual event histories, a description of a new method for viewing such data is given. The technique displays an event history as a pencil-like multi-faceted object in three-dimensional space, with changes of state being represented by changes in colour, texture or height along the length of the pencil. Viewing a single history allows a detailed examination of an individual's life events, and allows the relationship between changes of state in many variables to be examined. The extension of these ideas to viewing populations of event histories is then described. The Lexis diagram provides a suitable paradigm and is used to position the pencils in space. The resulting graphical representation is closely related to the ideas of Bertin, but with some important differences.

Introduction

Imagine a criminal career of an offender, collected since the individual reached the age of criminal responsibility. The career will consist of a sequence of offences at various dates throughout the offenders life, together with information about the sentence passed for each offence, time served. Other information may also be collected relating to the personal life of the offender. For example, information on his marital history (single, cohabiting, married, divorced etc.), family history (number of children in the household) and work history (unemployment or employment, salary) might also be collected over time.

This is an event history, as we are primarily interested in events (criminal convictions) and states (such as unemployment), the associations between them and changes over time. Although most examples arise naturally in the social sciences, there are many examples from medicine (examination of medical histories of a group of patients, looking at drug and surgical treatment, recurrence of disease etc.), from management (time-management studies) and psychology (observations of babies following stimulus for following five minutes).

Over the last ten years, techniques for the statistical analysis of event history data have been extended

dramatically. In the example above, models could be developed for the probability of reoffending or reconviction as a function of the age of the offender, current family and work history, and previous criminal history. This in general will lead to survival model methodology, and if multiple reoffending within an individual is taken into account, to survival frailty models where there is an additional unknown factor - the frailty - measuring the propensity of the individual to reoffend.

Although statistical analysis of event histories has recently had a lot of research interest, there has been little activity in the development of graphical methods to visualise an event history dataset before analysis. This is no doubt partly because of the complexity of such datasets - it is easy to be overwhelmed with the number of variables and different dimensions of a typical study. This work intends to rectify this imbalance by exploring the potential of modern interactive scientific visualisation systems for the initial examination of complex event history data.

Pencil representation of an event history

In developing a graphical representation of event history data, we have been guided by two principles. Firstly, the graphic should represent the full complexity of the data if this is what is required by the analyst - in Bertin's words «A graphic should not only show the leaves; it should show the branches as well as the entire tree» (Bertin, 1983, preface). In this context, this means representing the complexity of an individual event history, as well as representing a sample or collection of event histories. Our aim in producing a visual image of event history data is not to present summary information which will allow the viewer to interpret the graph in the way that the author wishes, but to allow the viewer to examine the data, looking for strange individuals or outliers, common relationships between measured variables, and patterns of groups of individuals who seem to behave in a particular way. Above all, we followed Tufte's invocation to «tell the truth about the data» (Tufte, 1961)

Our second principle was to allow the user full control over the nature of the graphic. The user, and not the developer makes decisions on the number of variables to examine, the method of alignment and the retinal variables used.

We begin by considering the types of variable which might exist in an event history dataset.

Firstly, there are variables as proxies for time, such as age of the individual, and calendar date. Secondly, there will be usually a large number of time-varying variables, representing changes of state or value over time. These may be continuous (such as earnings/week), ordinal (such as educational level reached) or categorical (such as type of crime committed).

Thirdly, there are time-constant variables which will not change over time - such as gender or ethnic group of the individual. Lastly, there will be pure events, either internal to the history of the individual (such as gaining a driving licence) or common to all individuals in the study (such as a change in government).

How do we best represent an individual history? Using Bertin's terminology, one way is to use a rectilinear construction (in either linear or circular form) for each time-dependent variable. A common time axis is defined representing the cumulative amount of time which the individual has been studied. For each variable, each state is represented by a different colour, density or pattern. Each part of the rectilinear construction is thus proportional to the amount of time the individual spends in that state before moving to a different state. With many time-dependent variables, we can produce a planar multi-faceted image, consisting of multiple rectilinear components side by side, or by multiple concentric rings. The latter idea has already been explored (Barry, Walby and Francis, 1989) where this representation was called a tulip diagram. However, such a diagram places undue visual emphasis on the variables contributing to the outermost rings, and this is not a desirable characteristic. We prefer here to consider multiple rectilinear components side by side, but to move into three dimensions, wrapping these multiple components into a solid object. This produces a pencil-shaped object, with each face of the pencil representing a different variable, and the length of the pencil representing the length of the event history. Changes in each of the face variables can be represented in many ways - in Bertin's terminology, these are visual variables and can be represented by the retinal factors of size, value, texture, colour, orientation or shape. Categorical variables can be represented by a set of colours, textures or patterns, with each categorical value represented by a different colour, texture or pattern. It is possible to represent continuous variables such as income as changes in height along a face of the pencil - however in practice we have found that the resulting graphic is difficult to interpret.

As an example, Figure 1 shows a perspective view of a typical pencil representing an employment event history for a married couple taken from a retrospective survey of three hundred couples living in Kirkcaldy, Scotland in 1985 (For further details of the dataset see Francis and Fuller, 1996). The event history for the couple begins at the date of marriage at the left hand side of the diagram, and continues until the survey date.

Figure 1 about here

The three faces, taken in order, and proceeding clockwise from the top of the pencil, are female employment status, male employment status, and the age of the youngest child in the household. Employment status for both husband and wife is coded 0 (dark blue) for not working and 1 (mid-blue)

for working. The age of the youngest child in the household is coded 3 (green) for no children, 4 (yellow) for a child aged under 1, 5 (red) from age 1 to under 5, 6 (magenta) from age 5 to under 11 and 7 (light green) from age 11 to under 16. In this history, we can see that the husband never worked throughout the survey period. However, the wife worked up until the birth of her first child, then stopped work until her youngest child was aged 10, when she returned to work, becoming unemployed for a while in the 1981 recession.

Other analysts would want to examine other variables over time, such as migration, housing tenure, educational level of the husband and wife and so on. The faces can be reassigned to new variables or extra faces can be added to the pencil display.

Comparing pencils - the Lexis diagram

Most studies consist of event history information from more than one individual, and we thus need methods to allow us to display more than one individual in the same graphic.

The simplest method of comparing pencils is to rank them by case order, and display them side by side. This is similar in concept to a set of pencils in a pencil box, but uses no further information in the dataset. A straightforward extension to this idea is to align the pencils according to age or according to calendar year; this then allows comparisons to be made more easily between pencils.

However, it is easily seen that these two simple displays are straightforward applications of the Lexis diagram (Lexis, 1875), the modern form of which is described in Pressat (1961).

The Lexis diagram is used extensively in demography and in survival analysis and has many useful statistical properties described in detail by Keiding (1990). A typical application would look at the survival experience of a group of patients entering a clinical trial. The x-axis represents the calendar date, and the y-axis represents the time spent in the study. Each patient is therefore represented by a solid line sloping at 45 degrees; the line slopes as both y and x increase with time. The line is anchored on the x-axis according to the date that the patient entered the trial. Figure 2 contains a typical Lexis diagram, showing eight individuals, with varying survival times. One of the individuals enters the study at time T days, and stays in the study until time T+A days - the time spent in the study (the y-value) is thus A days.

Figure 2 about here

The Lexis diagram can be modified in various ways. Firstly, information on events such as death can be added by placing an appropriate symbol at the end of the line. Secondly, changes of state can be introduced by using different colours to represent each state. Finally, the definition of the x-axis can be modified. If we change the x-axis to represent 'date of entry to the study' rather than 'calendar date', then the solid lines will then be vertical and will no longer slope at 45 degrees - as the patient proceeds in the study, y increases but x now stays constant. The x-axis can also represent other temporal variables (such as age of entry into the study) or any other variable - typically these will be marks of temporal variables

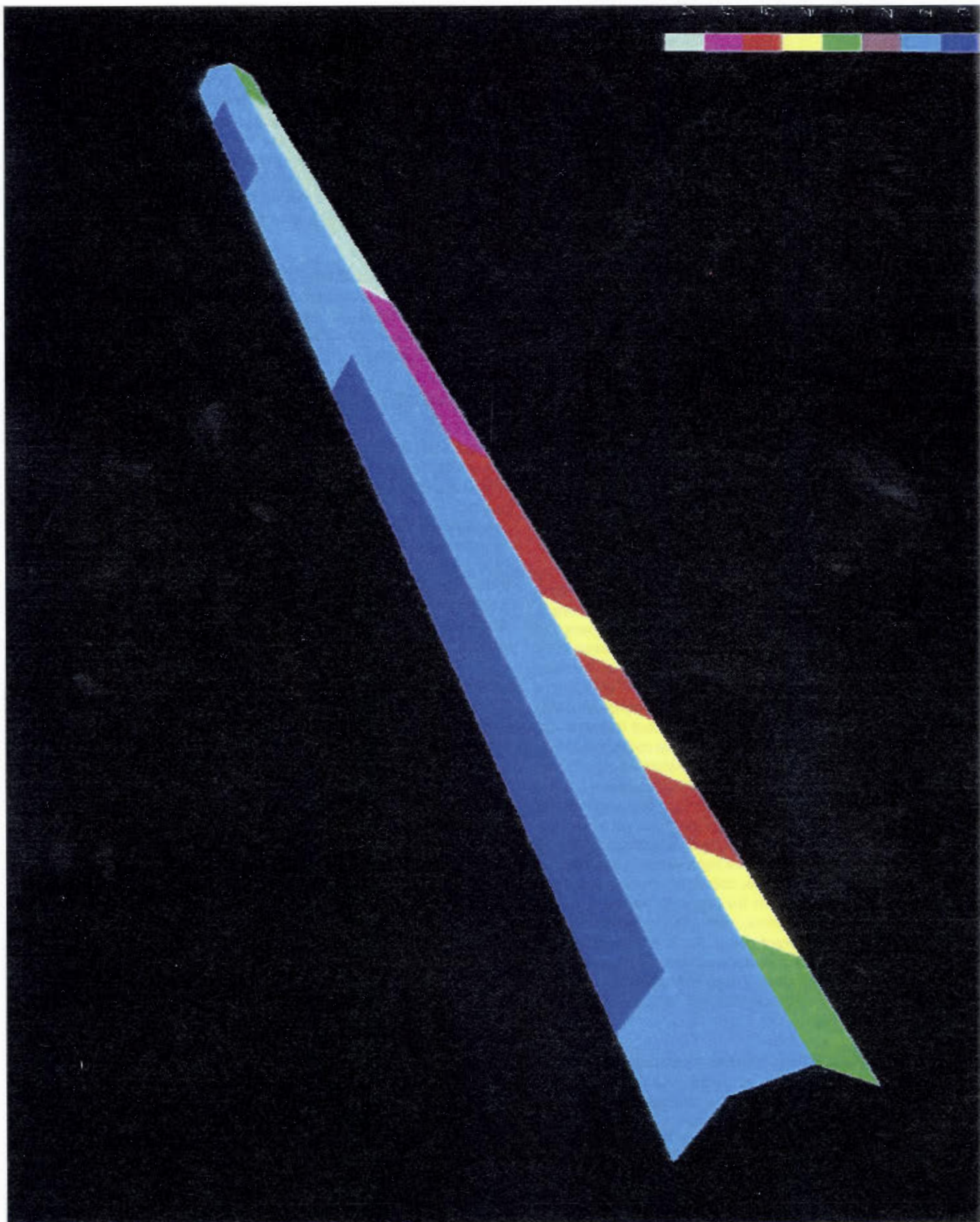


Figure 1 :

A perspective view of a pencil representation of the life history of a married couple. Time runs from left to right, starting at date of marriage and finishing at the survey date. Each face of the pencil represents a different variable. The top face represents the employment history of the wife, the middle face that of the husband, and the bottom face the age of the youngest child in the household. Explanation of the colours can be found in the text.

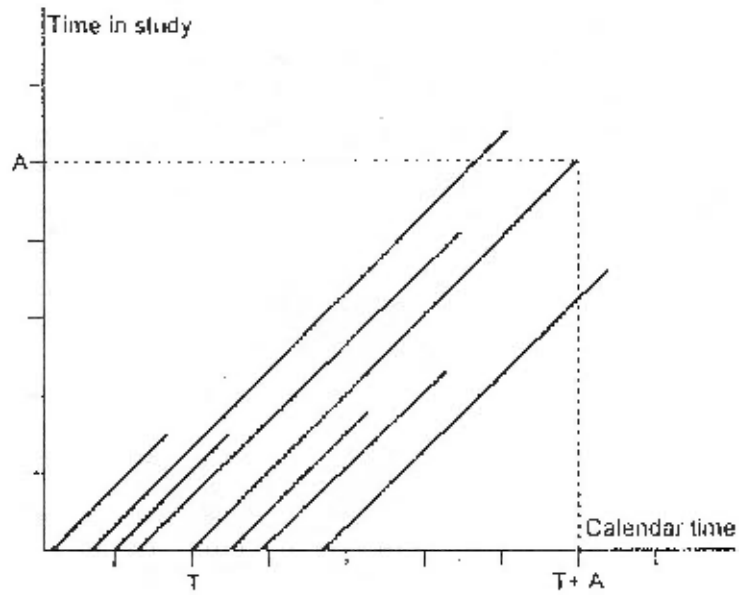


Figure 2:

A typical Lexis diagram. Each individual is represented by a 45 degree line.

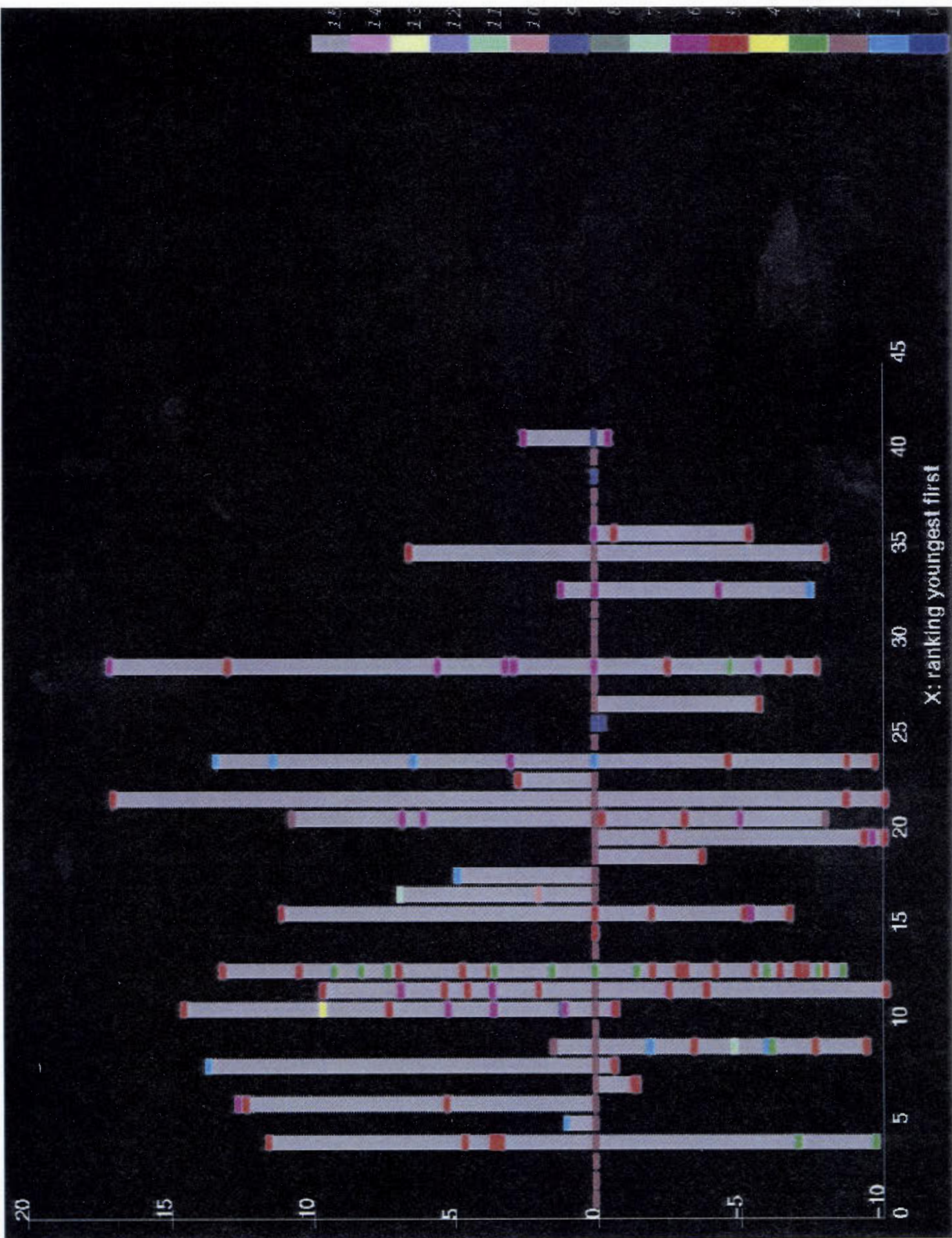


Figure 3:

A 2-D Lexis pencil diagram showing the criminal histories of 42 bigamy offenders who were convicted in England and Wales in 1973. The offenders are ranked with the youngest on the left and the oldest on the right. The representation of the criminal career starts with the first conviction and ends with the latest conviction in the study period. Each criminal conviction is represented by a bar of colour, with different colours representing the principal offence at that conviction.

such as 'rank order of age of entry into the study' - but might simply be a case ordering. Of course, if rank orderings are used to define the x-axis, the special statistical properties of the Lexis diagram are lost.

We can see that the two simple ideas for comparing event history pencils discussed above are therefore simply special forms of the Lexis diagram, with differing definitions of the x and y-axes, and with the lines replaced by pencils. In the first, the x-axis is the 'case order' or *index* of the individual, and the y-axis is 'time since start event'. The second graphic redefines the y-axis to be 'age' or 'calendar time', again keeping the x-axis as the index of the individual.

Returning to the original concept of a Lexis diagram, and replacing the Lexis lines by pencils we can see that a 'Lexis pencil' graphic would use a temporal variable such as 'age' or 'calendar date' along the x-axis, and use 'time in study' or 'calendar time' on the y-axis. We define this to be a two dimensional (2-D) Lexis pencil display- the dimensions refer to the number of axes.

As a first example, we consider the population of 7,442 sexual offenders in England and Wales in 1973, and examine their criminal history (which was obtained from the UK Home Office Offenders Index) over a thirty-two year period from 1963 until 1994. Choosing a subset of those who were convicted of a bigamy offence in 1973 gives us 42 individuals - 39 men and 3 women, with ages ranging from 20 years to 53 years at the time of the 1973 conviction. We examine this dataset using a 2-D Lexis pencil display. The x-axis is defined to be the rank order of the age of the individual at the 1973 conviction, and the y-axis is defined to be the time since the 1973 conviction. We display a single pencil face which represents the principal offence at conviction. Whenever an individual is convicted, a band of colour represents the type of conviction, and the remaining time the pencil face is white. The criminal histories are displayed from their first conviction to their last conviction within the 32 year period. Code 1 represents violence offences and is assigned the colour medium-blue, code 2 sexual offences (brown), code 3 burglary (mid-green), code 4 robbery (yellow), code 5 theft (red) and code 6 fraud and deception (magenta).

Figure 3 about here

Figure 3 shows the resulting display. At $y = 0$, all individuals have an offence displayed - this is their target bigamy conviction. Mostly the principal offence displayed is a sexual offence (code 2) although for ten cases other concurrent convictions, mainly of violence (code 1) and fraud (code 6) were judged more serious than the target bigamy conviction. What stands out from this display is that there are very few cases with other sexual convictions - only 2 individuals were convicted of another sexual offence (and in neither case was this another bigamy offence). For 16 individuals, the 1973 bigamy conviction was their only conviction. However, surprisingly, of the 26 individuals with other criminal convictions, 11 of these had principal convictions for fraud and forgery (code 6) - 27% of the whole sample. Finally, there seems to be little effect of age, with no obvious change in offence specialisation when tracking from the left of the figure to the right (that is from the youngest to the oldest bigamy offender). The initial analysis of this data raises questions as to whether bigamy should be considered as a sexual offence (as the

UK Home Office currently classify it) or whether it is better classified as a fraud offence (Soothill *et al*, 1997)

The 2-D Lexis pencil displays work well when the number of individuals is small, but with larger numbers of individuals, overlap of the pencils can easily occur. How can the display be improved?

Extending the Lexis display into three dimensions

First, we note that the 2-D Lexis pencil display is rather a strange concept, with 3-D pencil objects being plotted in a 2-D co-ordinate system. An obvious extension is therefore to use a 3-D co-ordinate system to position the pencils in space, with the y-axis defined as before and representing time spent in the study, but with a base plane defined by the x and z-axes rather than a base x-axis. We define this to be a 3-D Lexis pencil display. This extension into three dimensions corresponds to the characteristics of many datasets. Often event history studies have more than one way of representing time- typical variables will be the age of the individual and calendar date. If these are defined to be the x-axis and z-axis respectively, then the pencils will be anchored on the base plane according to age at first event and date of first event, and the pencils will slope at 45 degrees to both the x and z-axes. If the x-axis is defined to be 'age at first event' and the z-axis to be 'date of first event' then x and z are constant over time, and the pencils will instead be vertical. An example of a 3-D Lexis pencil display can be found in Francis and Fuller (1996).

If there is no obvious second temporal dimension, then any other continuous variable can be used to define the z-axis and to position the pencils. One special useful case is using a temporal variable such as age for the x-axis and the rank order of age for the z-axis. This will neatly space out the pencils along a curve which represents the cumulative distribution function of age when viewed perpendicular to the x-z base plane.

Interactivity and Visualisation

The 3-D Lexis pencil display is a complex diagram, and a single view of the 3-D world will not be sufficient to explore a dataset. We developed these ideas using AVS (Advanced Visual Systems, 1992), a scientific visualisation system which provides powerful user interactivity for exploring a three-dimensional world, such as zooming, panning and fly-through, as well as the programming environment to construct the Lexis pencils. The user has control over the assignment of variables to axes, the number of faces on each pencil and the variables assigned to them, the colour mapping of codes to colours, the width of the pencil face and the angle subtended by adjacent faces. Users can also rotate and spin the pencils around their own axes while keeping the viewpoint fixed, and can display identification information such as case number for any pencil by clicking on it.

Another useful facility is case selection. Subsets of cases may be defined either by selecting cases of the values of certain variables (which may or may not be displayed as part of the graph) or by specifying a set of case numbers.

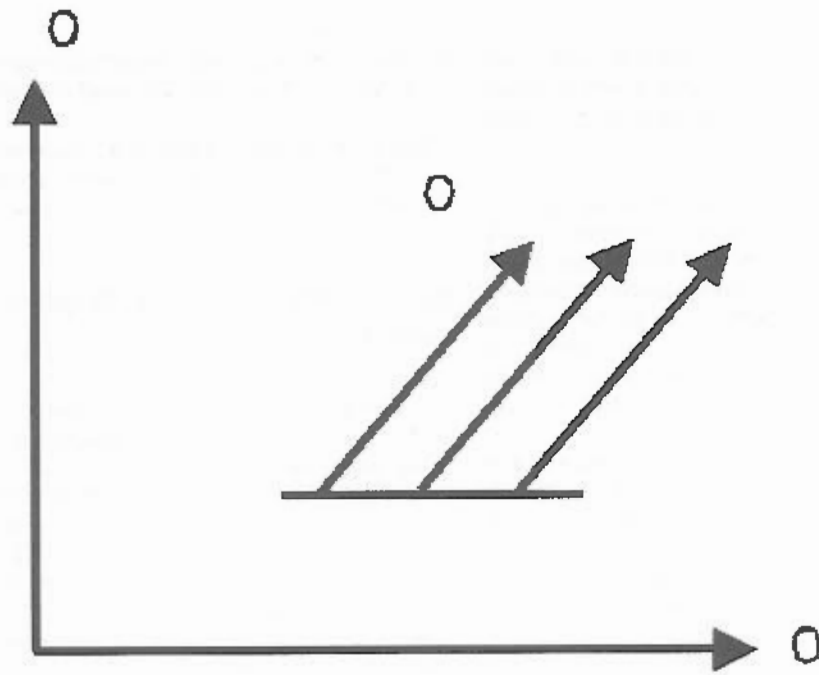


Figure 4:
Bertin's schema for the 2-D Lexis pencil display.

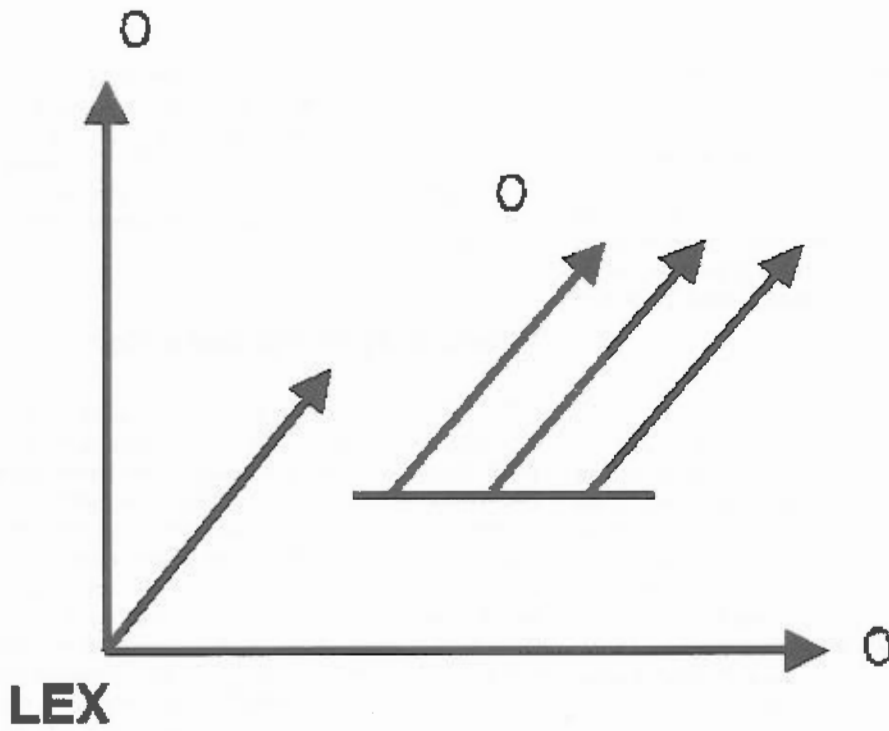


Figure 5:
The proposed schema for the 3-D Lexis pencil display.

This allows users to compare, for example, males and females in separate displays without the necessity of increasing pencil complexity by adding an additional pencil face.

Finally, we have considered carefully the problem of dissemination. The software allows postscript bitmaps of any view of the Lexis pencil world to be produced, and these can be published. However, it is often important to allow other users the ability to view the dataset themselves, and to identify their own special features of the data. We therefore provide the facility to write VRML (Virtual Reality Markup Language) files which define the 3-D Lexis pencil display. VRML viewers can now be purchased cheaply or downloaded free of charge on the World Wide Web to run on any platform. Other users can easily load a VRML Lexis pencil file into their VRML viewer and fly through the 3-D Lexis pencil world on their own desktop computer.

Bertin and the Lexis pencil display

We now return to Bertin (1983), which is essential reading for those interested in the grammar and philosophy of graphics. How does the concept of the Lexis pencil display relate to Bertin's semiology? There are many concepts of the Lexis pencil display which translate directly into his graphic theory. Bertin separates variables into two types - the two planar dimensions, which are denoted by x and y , and the visual variables, denoted by z . Each visual variable has an implantation (that is, it can be represented by a point, line or area) and values of the visual variable are represented by changes in retinal factors such as size, texture or colour. The 2-D Lexis pencil display uses temporal variables for the two planar dimension, and represents each additional variable as a visual variable, using area to represent a period of constant state, and colour or texture to represent the value of that state. The major difference is the use of a three dimensional solid object which is displayed in a two-dimensional axis system. Thus, the 2-D Lexis pencil display can be denoted by the schema given in Figure 4.

Figure 4 about here

Much of the time, the number of visual variables greater than one, which means that the 2-D Lexis pencil diagram is usually an inventory drawing, «displaying comprehensive information in a single figuration» (Bertin, p.172). In fact, the use of temporal variables as the planar dimensions suggests that the 2-D Lexis pencil display is close to an «inventory map», with time replacing the spatial dimensions of x and y .

Another important concept in Bertin's work was the ability to permute the rows and columns of the xyz matrix. This is allowed to a limited extent in the Lexis pencil plot, when

the x dimension is non-temporal. We have seen an criminological example in Figure 2 where the x dimension was defined to be the rank ordering of age. This variable could be easily replaced by the rank ordering of any other characteristic of the data, for example, the length of time between the first criminal conviction and the last criminal conviction. Those with a single conviction would then appear to the left, those with a long criminal history would appear to the right. Other permutations can easily be defined and used. However, in the Lexis pencil diagram, we do restrict the y dimension to represent the passage of time and do not therefore allow y to be permuted.

We now proceed to examine the 3-D Lexis pencil display. We have three 'planar dimensions', which should now properly be called 'spatial dimensions'. If these are now denoted by x , y and z , then we can define w to be the set of visual variables, giving us four components ($wxyz$) rather than Bertin's three (xyz). The 3-D Lexis pencil diagram can then be represented by the schema given in Figure 5:

Figure 5 about here

We have replaced the term GEO in Bertin's schema with the term LEX, to indicate that the resultant display is a temporal Lexis plot and not a spatial map.

In the Preface to the English Edition, Bertin (1983) makes the point that 'ten years of evolution' (since the book was written in 1973) 'has brought about an entirely different perspective'. Nearly fifteen years later, we can see that the perspective has again changed, with graphics developing into three dimensional visualisation, and dynamic graphics existing alongside static planar pictures. Communication of graphs need no longer be by means of the printed page, but can be through three dimensional worlds by using VRML technology. However, Bertin's grammar is robust, and we have shown that the semiology originally presented in 1965 and revised in 1973 (Bertin, 1973) can be developed further with some simple additions to deal both with the development of dynamic three-dimensional graphics and the complexities of event history data.

Acknowledgments

The authors are grateful for the support of the UK Economic and Social Research Council (ESRC), who partially funded this research under the Analysis of Large and Complex Datasets initiative (grant number H 519255029). The 1973 sample of bigamists used as an example comes from another ESRC research project *Criminal Careers and Sex Offending* (grant number R00023 6540), and we are grateful to Professor Keith Soothill for helpful discussions relating to this dataset.

Références

Advanced Visual Systems (1992) *AVS Users Guide*, AVS Inc., Waltham, Mass.

Barry, J. T., Walby, S. and Francis, B. (1990) Graphical exploration of work history data. *Quad. Statist. Mat. Appl. Sci. Econ. Soc.*, 12, 65-74

Bertin, J. (1973) *Semiologie Graphique*.

Bertin, J. (1983) *The Semiology of Graphics*. Translated by W. J. Berg. University of Wisconsin Press: Wisconsin.

Francis, B and Fuller, M (1996) Visualisation of event histories. *Journal of the Royal Statistical Society Series A*. 159, 2, 301-8

Keiding, N. (1990) Statistical inference in the Lexis diagram. *Phil. Trans. R. Soc. Lond. A*, 332, 487-509

Lexis, W. (1875) *Einleitung in der Theorie der Bevölkerungsstatistik*. Trübner: Strassburg

Soothill, K., Sanderson, B., Peelo, M. and Ackerley, E. (1997) Bigamy - Likelihood of reconviction and the place of bigamy in the pantheon of crime. *Centre for Applied Statistics working paper*, Lancaster University.

Pressat, R. (1961) *L'Analyse Démographique*. Presses Universitaires de France: Paris

Tufte, E. R. (1961) *The Visual Display of Quantitative Information*. Graphics Press, Cheshire.