# *The Historiographer in Computerland:*
## *A review Article*

by D.I. Pool *

I

Progress for some may seem a retrograde step to others, yet there must be some changes which are seen by most people as elements of progress. The case to be presented in this review is of the latter type: the majority of observers would accept the value of increasing the historiographer's data base, even if by methods which are new to the field in form if not function. A major qualification might be that the acquisition of this data base often involves the use of computers and other electronic equipment. A few practitioners might take a neo-Luddite stance and see the computer as inherently evil, yet even a moment's reflection will lead to the conclusion that this apparently new methodology is in reality central to the best traditions of historiography. Even with computer technology the practitioner must employ her/his own powers of criticism, they must carry out a rigorous content-analysis of documentary sources, and they must utilize the simplest possible method of data "reduction" and analysis. For analysis one must, when necessary, appeal to the available accumulated knowledge and methodologies in adjacent fields. Such then are features of the developments which are the subject matter of this brief commentary.

Specifically, I am concerned with a rapidly expanding field of "social" history: my attention will be focused of the reconstitution of related historical documents by individual record-linkage, a process most rapidly and efficiently realised by computer. To date this has primarily been at the micro-level—family reconstitution—but future developments need not exclude macro level studies, nor adaptation to a range of topics. Nor should one exclude other appropriate methodologies and substantive issues "economic" and cultural as well as social.

Before discussing in further detail some of the methods and research foci I must make one other general comment. These developments do imply acceptance of two positions neither of which, it would seem to me, should be an anathema to any historian. Firstly, most of the methods may lead to a revaluation of history, not so much in terms of grand themes, but in terms of the vantage point of the observer and the observed, as well as the universality of the records from which observations may be drawn. The concern will not be for example, with the relative significance of economic determinism as against epic figures and events in an explanation of aspects of sixteenth century English history, but rather what was

* Department of Sociology and Anthropology, Carleton University.

the behaviour of Mr. Sixteenth Century Everyman. As Wrigley has recently written:

> The 20th Century is sometimes called the century of the common man. Many will doubt the truth of this aphorism. But it may be that it will prove to be the century in which the history of the common man, not seen through the eyes of his betters but drawn out from the workaday records of the past, becomes a major part of historical work. And in this regard nominal record linkage, for all its technical trappings, is a means of discovering things about the lines of ordinary men which would otherwise remain obscure.[1]

Secondly, there is the marriage (perhaps one should say remarriage) of history to the allied social science disciplines of economics, sociology, anthropology and demography. Demography and economics have always paid attention to secular trends, so that this union is natural and of mutual benefit. Many of the methods come across to history from demography, yet demography will also benefit from the contract in that the viewpoint of the historian will aid in the analysis of problems of critical applied significance for demographers. By contrast, recently, sociology has tended to favour synchronic analysis to the extent of being almost ahistorical. This has sometimes been to the detriment of the subject, so that sociology stands to gain considerably from the union. Of course the analytical tools and even some theoretical perspectives of sociology will be of value to history.

This marriage (or remarriage) need not imply the divorce of history's present spouse, the humanities. Instead, as is evident from the range of backgrounds of contributors to Wrigley's volume[2]—history, philosophy, demography, genetics—polygamy is the only feasible conjugal state. Certainly, one would add the humanities disciplines associated with the analysis of literature to the list.

## II

Most of the methods to be described here imply the use of quantitative data, or data on attributes and characteristics handled in a quantitative manner. Moreover, they will be at a structural (the actual behaviour as measured by variables simply defined and delimited) rather that at a socio-psychological level, so may lack the "explanatory" power gained from a study of attitudes, norms and values—they will describe what a norm of "usual" behaviour, rather that what was a norm or "desired" behaviour. A study carried out today by sample survey can at the one

[1] E.A. WRIGLEY, "Introduction," in WRIGLEY, *Identifying People in the Past* (London: Edward Arnold, 1973), pp. 1-2. It is based on a conference on Nominal Record Linkage in History, Held at the Institute for Advanced in Princeton in May 1970. The essays discuss, in general terms or by reference to case-studies, many of the critical technical and even philosophical problems of record-linkage. This is essentially a "how to do it" book drawing on the experiences of various projects, whereas other studies give substantive results. Yet it is a readable and clear exposition which will make their techniques relatively comprehensible to the general reader.

[2] *Ibid.*

time attempt to collect data on both structural and socio-psychological variables relating to the same individuals. Obviously, this is not generally possible with historical reconstruction,[3] so that explanation must rest, as it does in economics and often in demography, at a structural level. By contrast the biographer or the general historian interpreting normal documentary evidence may easily introduce qualitative data and even information which suggest the influence of certain norms (desired), values and attitudes.

This situation need not deeply concern the social historian for three reasons. Firstly, the first order explanatory value of structural variables (e.g. place of residence, religion, occupation, age etc.) has been adequately demonstrated in economics, sociology and demography. Secondly, from this level of "explanation" one can then develop hypotheses which can be tested by incorporating into the analysis other perhaps qualitative data; for the historian these would include literary sources subjected to careful content analysis.[4] This may mean the integration of levels of data and modes of analysis which vary considerably, and obviously this poses problems for the scholar. Finally, it must be recognised that norms (desired behaviour) do not equal norms (actual behaviour). Indeed, one of the central problems of sociology is the relative lack of fit in general between attitudinal and structural level behavioural variables (e.g. between, for example, the aspirations of school-leavers and their subsequent performances which may be largely governed by structural factors such as income). Again, in anthropology a major problem is whether the key informant is reporting their ideals or the actual behaviour. But this latter example is close to the very problem posed to historical reconstruction —the reason for the detailed reconstitution is in fact to test whether the "key informant", usually the literate chronicler from upper middle or upper class, was correct in her/his assessment of events occurring to many of her/his contemporaries.

The various methods by which one attempts to solve this problem fall into two levels of aggregation: micro, where the unit of observation is one person, family, household, or even work unit, but including the sum of these observations; and macro, where the unit of observation is a society, a community, a parish, a county, an industry etc., and where the methodology is termed aggregative analysis.[5] Although one can, with prudence intermix these two levels of aggregation it does pose problems for the analyst. Very often we tend to construct theories and to formulate hypotheses at one level, but are forced, for practical reasons, to test at the other. The failure to recognize the discrepancy can cripple a research project. This may not be as true for other areas of historical research; in this sense they are closer to social reality. Thus, the psychiatric state

[3] An exception might be a random sample of court cases as recorded in newspapers or elsewhere.

[4] Ian WATT, "The New Women: Samuel Richardson's *Pamela*," in Rose L. COSER, *The Family Its Structure and Function* (N.Y.: St. Martin's, 1964); pp. 267-289.

[5] Glossary in WRIGLEY, *op. cit.*

of George III (micro level), may have affected decisions he took on behalf of the state (macro). A document written by one man (Durham) related to collectivities (Upper and Lower Canada), yet the historian can move back and forth between these with relative ease. By contrast, one cannot analyse the macro level relationship between two macro level variables, say the average family size in relation to the average size of holding for each Ontario county and for the Province in the year 1871, and come out with a conclusion that any particular farm family was likely to have geared their family size to their perceived well-being as measured by the size of their holding. Indeed, we may not even be certain that the relationship existed at all at the micro level. This may seem very much a truism yet a failure to recognise the problems of level of aggregation has been a persistent theme in the social sciences. In sociology, this has been termed the "ecological fallacy," but could relate equally well to time series data.

## III

A micro-level method receiving widespread attention in history and demography is the linkage of records relating to individuals. Generally, these have been derived from parish registers. Once account has been taken of migration, parish registers form a relatively closed system so that one can follow an individual from their baptism, to their marriage, to their offsprings' baptisms, to their burial. The procedure could be applied to other discrete data sources, as well as by linking data from various sources. For example, could one link census returns to parish registers? Again, geneological data, or data on any "closed" systems or specific institutions (e.g. the priesthood, company records, records on a trade union or professional organization) are susceptible to similar methods of analysis. The interesting thing is that the notion of record—linkage, although not new, has been picked up simultaneously and applied at an accelerating pace both for historical reconstruction and for the solution of contemporary problems of social data analysis, for vital statistics, criminal statistics, etc. There has been a relative pooling of experiences gained at both levels and thus feedback about problems and solutions. Indeed, the whole field has been revolutionised by two technical breakthroughs: (1) the use of phonetic equivalents to form links, which maximises the efficiency of the linking process and minimises risk of "clerical"—type errors, including variations in surname spellings;[6] (2) the use of computers with enormous storage capacity and also speed. But one must not overemphasize the role of the computer; it is merely a tool, and is only as valuable as the programme which governs its operations.

It would seem to me that record-linkage, once proven as a technique, might be adapted both to micro-data, in many fields of history: economic, social and even political. I do not wish to exaggerate this point, because the data must meet certain criteria. For example, is the system closed?

[6] Ian WINCHESTER, "On referring to ordinary historical persons," in WRIGLEY, *op. cit.*, pp. 17-40.

Do the means of identification change radically? Are the data amenable in any intelligent fashion to some form of quantitative analysis?

## IV

Obviously, record-linkage is merely one method of value to the historian. Any historical data source—documentary or not[7]—has its attendant methodology. But I would like to concentrate here on social history. At the micro level the range of alternative data sources and methods are very wide, and are being employed, for example in studies on the family.[8] If data sources are susceptible to content analysis, numerical analysis, record-linkage, inspired speculation, or whatever, singly or in combination an analyst will be found.

Equally well at the macro level sources and methods are legion, while analysis at this level has the advantage that it can sometimes indirectly "link" or correlate data from very diverse sources, including enumeration, vital registration, cartographic, general descriptive, agricultural and other economic and physical geographic. We could consider such studies as falling roughly into two categories:

1) *Univariate:* The analysis of aggregate data relating to one pattern of behaviour and only one unit of observation (e.g. one county). However, this can easily be broadened into numerous units, while it can be extended with facility into multivariate analysis.
2) *Multivariate analysis:* The analysis of diverse aggregate data for different patterns of behaviour, and often for numerous units of observation.

The first case is typified by demographic data drawn from the census and vital registration. A census is simply a single passage observation giving data on characteristics as they were at that moment. On its own it leads to synchronic analysis, while history is, by definition, diachronic. This is where vital registration comes in, (especially parish registration, the most common basis for record-linkage studies) for it provides the data on ongoing events. However, record-linkage links individuals whereas the normal vital analysis uses as a numerator the sum of the reported vital events (births, deaths, deaths by cause, etc.) related to a denominator usually drawn from the census and estimated for intercensal years. Although the analysis may be restricted to data relating to only one vital event a critical problem arises in reconciling the numerator (the events) and denominator (supposedly the population exposed to risk of those events). It is often the case that data will be available on the events but no suitable information on exposure to risk. A colleague and I ran into a particularly

[7] For a comprehensive description see T.H. HOLLINGSWORTH, *Historical Demography*, in the series: *The Sources of History, Studies in the Uses of Historical Evidence*. (London: Hodder & Stoughton, 1969).

[8] See for example, special issue *Journal Marriage and the Family*, Vol. 35, (April 1973); or Michael GORDON, *The American Family in Social-Historical Perspective* (N.Y.: St. Martin's, 1973).

trying example of this in a cross-comparative study of tuberculosis. Analysis of cohort[9] and period rates had shown that mortality from this cause was much lower in New Zeland (and each other Australian colony except Queensland) in the late nineteenth century than in England-Wales or Massachussetts. The problem was, why? Some data on deaths by cause, by birthplace, by duration of residence in New Zeland had the potential of helping answer this question; our hopes where dashed because we did not have the corresponding denominator, nor could we estimate one.[10]

Estimation often permits analyses to be carried out on macro-data in a way which would not be permissible for micro-data through record linkage. Estimation can run from adjustment to one or more elements (for under — or incorrect — enumeration or under —, mis-registration) to estimates of the vital rate. The latter methodology, now a normal part of the demographic armoury, requires the availability at least of a census giving broad age-groups and some notion of whether the growth rates and structures by age have changed radically or not.[11] They are of restricted value for populations exposed or "open" to migration. Even for "closed" populations the investigator must have some indication, that age-structure has not altered too drastically for a prolonged period (say 40-50 years prior to the date for which the estimate is to be made) from other causes such as high age-specific mortality from warfare or epidemics. In this instance a systematic content analysis of non-quantitative data sources is essential.

As an example, I wanted to estimate however crudely, the vital rates for the New Zealand Maori population during the nineteenth century and explain intercensal growth rates. In addition to the purely demographic techniques, support for the vital rates and for an understanding of growth rates came from a detailed analysis of standard historical sources: accounts and journals of missionaries, travellers, officials; British parliamentary papers and colonial bluebooks, and the reports of district commissioners and others in appendices to the journal of the House of Representatives. The clustering both in time and geographically of reports on some epidemic permitted me to get a qualitative notion of epidemic mortality. An intriguing fact was that these often paralleled epidemics in neighbouring regions of the south-west Pacific, and even pandemics (as for influenza about 1851, in the 1890s and in 1918).[12] In doing this analysis one had to go further afield even to biographical data on various rap-

[9] Record-linkage can very easily take the form of cohort analysis (eg. following the Baptisms of 1690, to the marriages of the same baptism-cohort in 1706, 1707...). However, cohort analysis normally uses macro data age-specific rates: eg. the rate for 0-4 year olds in 1880-84 refers to the same birth-cohort as the rate for 15-19 year olds in 1895-99.

[10] D.I. POOL and K.C. CHAN, "Differential Declines in Tuberculosis Mortality," *Proc. International Population Conference, London, 1969*, Vol. II (Liège: Int. Union. Sci. Study of Pop. 1971): 1006-1012.

[11] That is, whether the population is stable or at least quasi-stable. See UNITED NATIONS, *Manual IV: Methods of Estimating Basic Demographic Measures from Incomplete Data*, (N.Y.: U.N. 1967).

[12] D.I. POOL, "Estimates of N.Z. Maori Vital Rates from the mid-19th Century to W.W.I.," *Population Studies*, Vol. XXVII (March. 1973): 117-125.

porteurs, in order to evaluate their likely knowledge, reliability etc. In sum, mechanisation merely assists the historiographic artisan; their own hand tools and skills can never become obsolete.

A second case of quantitative analysis takes different data from diverse sources and correlates these by use of regression analysis or other multi-variate techniques. Here the variables are assumed to be independent, the linkage merely being joint occurence in some unit of observation, either an administrative district or a calendar period. From this joint occurence we assume, test for, and measure association, although we can never be certain that there was an association. Nor can we assume that the association at the macro level implies association at the micro, and anyway behaviour at the macro level generally will not be able to be defined, delimited and measured in the same way at the micro level, and vice versa. For example, to belong to a social class is a characteristic of individual or family delimited by subjective as well as objective criteria. For an area one could talk only of social grade, and restrict the analysis to readily measurable social structural variables, such as the percentage of the population with a given income, occupation or education.

This method has two very distinct advantages. Firstly, the possibility of indirectly "linking" data from diverse sources makes it particularly powerful as a means of interpretation and explanation. Secondly, the units of observation become increased by comparison with other types of macro analysis. Nevertheless, the advantage this bestows can be reduced if the researcher is forced to use less satisfactory or even "surrogate" indices. For example, for Ontario and Québec, Henripin calculated and estimated age-specific birth rates for 1851, and extended this to Nova Scotia in 1871. But for a county in one of these provinces in 1871 it might be difficult to get registration data.[13] The obvious ploy would be to employ the child-woman ratio (say, children 0-4 or 0-9 years of age, to women at reproductive ages). But this index does not measure fertility alone for it is contaminated by differential quality of enumeration, differential migration (e.g. a heavy "passive" migration of young children accompanying parents, an effect which could be most severe for a small unit subject to family migration over a very short period), and infant and early childhood survivorship. Once again let it be said that quantification can never replace judgement based on detailed knowledge.

## V

The methods outlined above are presently being applied in Canadian historical research. Thus, as Wrigley reports, record-linkage and family reconstitution "studies are either published or in train in France (where they may be numbered by the score), Belgium, *Canada*, England, Esto-

[13] Jacques HENRIPIN, *Trends and Factors of Fertility in Canada*, Appendix E. (Ottawa: Statistics Canada, 1972).

nia, Germany, Hungary, Italy, Japan, the Philippines, Poland, Scandinavia, South America and the United States."[14]

In Anglophone Canada one of the more well documented studies is the Hamilton project.[15] But Québec is perhaps one of the most logical sites for this genre of historical studies. This results from the fact that French historical demographers (Louis Henry and Michel Fleury, particularly Henry who trained many Canadians, "first fully worked out" the logic of this technique, while there is the availability of good record-keeping by parishes from the 17th century on what was often virtually a closed system. The classical work was, of course, carried out by Henripin,[16] but a systematic study is currently underway, directed by Hubert Charbonneau and Jacques Légaré which

> a pour but de reconstituer la population canadienne-française avant 1850, à partir des registres d'état civil et des recensements. Pour fins d'études démographiques, on veut établir, à l'aide des ordinateurs, un registre de population fait de dossiers individuels. Ce registre comprendra la liste des événements démographiques auxquels un individu a participé, soit comme sujet d'acte soit comme témoin, ainsi que les caractéristiques de cet individu, fournies par les sources exploitées.
>
> Un projet aussi ambitieux à propos de populations anciennes n'ayant été réalisé nulle part, il faut constamment innover. Ainsi, un important obstacle a dû être surmonté, que consistait à transformer les ordinateurs en généalogistes. Un code phonétique des noms de famille a été mis au point pour vaincre les difficultés dues aux variations orthographiques. De nombreux programmes d'ordinateur désignés sous le nom d'« Hochelaga», ont été créés afin d'automatiser les opérations de reconstitution des familles.
>
> Depuis le microfilmage des sources manuscrites jusqu'à la production de tableaux statistiques, toute une série d'opérations précèdent l'analyse démographique. Celle-ci portera d'abord sur le XVII$^{e}$ siècle, c'est-à-dire sur près de 250,000 mentions nominatives tirées de 30,000 actes d'état civil et des recensements nominatifs de 1666, 1667 et 1681. Une étude-pilote sur ces recensements a déjà permis de développer des méthodes de couplage aussi bien que de mettre au point des instruments pour la collecte des données. La qualité et l'abondance des documents ancien devraient permettre de renouveler nos connaissances sur les deux premiers tiers de l'histoire démographique du Canada et, par conséquent, sur le comportement des populations en milieu de colonisation.[17]

An interesting aspect of these Canadian historical studies has been the fact that they have run alongside major theoretical and technical development occuring in Canada in the field of record-linkage by Felligi,

[14] WRIGLEY, *op. cit.*, fn. 1. For England a useful study is E.A. WRIGLEY, *An Introduction to English Historical Demography* (London: Weidenfield & Nicolson, 1966). For Britain, Europe and the United States see D.V. GLASS and D.E.C. EVERSLEY, *Population in History* (London: Arnold 1965).

[15] IBID., *passim*. See also M.B. KATZ, *The Hamilton Project: An Interim Report*, No. 2 (Toronto: O.I.S.E., Nov. 1970).

[16] J. HENRIPIN, *La population canadienne au début du XVIII$^{e}$ siècle* (Paris: P.U.F., 1954).

[17] Hubert CHARBONNEAU, "Le programme de recherche en démographie historique du département de démographie de l'Univ. de Montréal," *Bull. de l'assoc. des démographes du Québec*, Vol. 2 (Oct. 1973): 51 (with bibliography).

Kennedy, Newcombe and others.[18] These developments have been in applied fields ranging from linking vital, hospital and morbidity statistics, to the couplage of information on criminals.

At the macro level of research a number of examples could be given. For example historical data of the first type (univariate) have been analysed for Canadian Indians and estimations made using Stable population methods.[19] In the second or multivariate case McInnis has underway a complex study of Ontario and Québec Counties during the 19th century, studying the association between demographic, economic and other variables.[20]

## VI

Obviously, the methodology described here is going to add to the historiographer's data bank. Yet once the social historian has completed her/his analysis of the data files and studied each rate, what will have been added to the sum of our knowledge of historical process? One response would be to echo Wrigley's call that this may be the century of other centuries common people. The accumulation of knowledge about the past may well be sufficient cause in itself, and the methods described here should permit an accumulation of data in areas of human endeavour inadequately studied. From record-linkage alone one could describe certain norms (normative meaning usual) of social behaviour: when people married, how many children they had, when they died, their probability of surviving infancy, whom they married (thus social class, residential, religious and other forms of exogamy and endogamy; monogamy and polygamy; the effect of laws and norms relating to divorce and the remarriage of widows and divorcees; age differences between spouses; illegitimate conception and thus accelerated marriage; bastardy and incest; etc. etc.) Certain parish registers or geneaological records even contain other data permitting fuller explanation by record-linkage of some patterns of behaviour. Furthermore, by using both micro data aggregated and macro sources directly many other critical themes of importance to economic historians as well as to social historians may be attempted—changes in settlement patterns, urbanisation, land use and employment, differentiation of the labour force, enclosure movements, scale and other indices of industrial production. Some results conceivably could even lead to the revision and revaluation of orthodox interpretations.

I have no wish to introduce a practical purpose to what has been so far a purely scientific voyage of discovery, yet who will complain if exploration also indicates that potential markets abound in newly found lands?

[18] WRIGLEY, *op. cit.* Select Bibliography.

[19] A. ROMANIUK and V. PICHÉ, "Natality Estimates for the Canadian Indians by Stable Population Models, 1900-1969," *Canadian Review of Sociology and Anthropology*, IX (Feb. 1972): 1-20.

[20] R. MARVEN McINNIS, "Birth Rates and Land Availability in Nineteenth Century Canada" a paper presented to the Population Assoc. of America meeting, Toronto, April 1972 (abstracted in *Population Index*, XXXVIII (July-Sept. 1972): 266-67.)

Here one could suggest two, both of which require the rigour of orthodox historiography to be brought to bear on their realisation. Firstly, there is an obvious need to chart the characteristics which are the continuities in a given culture. Secondly, and more specifically, social demographers, sociologists and economists urgently require data for long-run cycles (as against short-run) relating to the interface between economic and socio-demographic behaviour. This need is basic to our understanding of both modern industrialisation and the demographic transition, processes which certainly commenced in the eighteenth century. We still have relatively unconfirmed postulates about the early development of their relationship, and yet this information is essential if we are to meet and perhaps to overcome some of the crucial problems facing us 200 years later.