LONDON ELECTORAL HISTORY – STEPS TOWARDS DEMOCRACY

7.14 PHONETIC CODING FOR THE LED – SOUNDEX CLASSIFICATION

Some imputed categories have also been applied to the data in the course of creating the LED.[1] Phonetic coding of surnames is an example of rule-based classification. Many English surnames are homophones that are spelled differently: examples include Lee, Lea, and Leigh. A phonetic algorithm aims to remove the least stable elements of a surname string, and to give a common value to the most readily confused characters. Thus surnames with the same pronunciation should be encoded with the same string, so that matching can occur despite minor differences of spelling.

It should be noted that phonetic coding is imperfect. It will place in the same category some names that were not homophones. Meanwhile, it may fail to unite some character strings that undoubtedly were homophones, for example, if the initial letter of the surname string is different. There is potential for confusion if the 'soft G' initial letter of a surname string (as in 'soft' George rather than 'hard' Gordon) was rendered as 'J'. Other ambiguities might arise between 'hard C' (as in Castle) and 'soft C' (as in Citadel) in a surname string, if rendered as 'K'. Or, for those challenged by aspirates, if the surname string began with an 'H'.

Hence, like the use of DNA evidence in courts of law, phonetic coding should therefore be treated with caution; and supporting evidence should be sought wherever possible.

The parliamentary committee, which adjudicated the petition following the Middlesex election of 1802, had to consider differences between the spelling of a voter's name in the poll book and in the Land Tax Assessments. They accepted that if the name sounded the same, then the variation could be ascribed to the error of the poll clerk and the poll was not thereby invalidated. 'Colten' was taken to stand for Coulton, the real name of the voter, or his occupier, and the name on the assessment;

'Hoskins was accepted for Hoskinson; Myers for Swires; Ridgeway for Ridgley; Tacker for Decka'.[2]

In adopting a rule-based phonetic coding system, the LED thus replicates the spirit of the parliamentary committee's pragmatic approach.

A variety of phonetic coding systems are available to deal with English-language homophones. Those used by the security services remain secret.[3] Moreover, the strategy of pair comparison, which used in probabilistic record linkage, cannot be embedded in a systematic database. So the Guth algorithm,[4] which quantifies the degree of similarity between two character strings and has been used to analyse free-standing historical records, could not be used for this study.

After reviewing the options, the LED adopted the freely available Russell Soundex algorithm. One strong reason for that was because its wide availability makes it a *de facto* standard. For example, it is used by the U.S. National Archives and Record Administration. Russell Soundex codes thus facilitate comparison of results from the LED with those from other datasets.

The system works with the retention of surname character strings, which allows users to evaluate and to adopt other algorithms, if required.

To facilitate all exercises of data enquiry, recent advances in data mining and de-duplication technologies are available in commercial software. There are also websites which evaluate the efficacy of de-duplication systems.[5]

Within the LED, the field `Scode` contains the Russell Soundex code for each surname string.[6] Russell Soundex coding retains the initial letter of each surname string, and removes any terminal 'S' from that string. The letters 'W' and 'H' are ignored, as are non-alphabetical characters such as apostrophes. The letters 'A', 'E', 'I', 'O', 'U', and 'Y' are treated as separators. They are never coded (except internally), and they are only output as the initial letter of the surname. When two characters that would receive the same coded value are not separated by a separator, a single code is produced.[7] The coding of the remaining letters is given in Table 97.

The outcome, in common with procedures throughout the creation of the LED, renders the data assessable by historians, without interposing the needs of the users ahead of the integrity of the original source data.

Table 97 Russell Soundex coding of surname characters

Character	Code
B,F,P,V	2
C,G,J,K,Q,S,X,Z	3
D,T	4
L	5
M,N	6
R	7

Source: LED.

Notes

[1] See section 4.1.7 and Table 23.

[2] Peckwell, *Cases of controverted elections*, ii, p. 68.

[3] The algorithm of the New York State Intelligence Information System (NYSIIS), no longer used by the New York State authorities, was described in H.B. Newcombe, *Handbook of record linkage: methods for health and statistical studies, administration, and business* (Oxford, 1988), pp. 182-3; and in J. Atack, F. Bateman, and M.E. Gregson, '"Matchmaker, matchmaker, make me a match": a general personal computer based matching program for historical research', *Historical Methods*, 25 (1992), pp. 63-4.

[4] G. Guth, 'Surname spellings and computerised record linkage', *Historical Methods*, 10 (1976), pp. 10-19.

[5] See, for example, www.dedupesoftware.com.

[6] With warm thanks to Matthew Woollard of the History Data Service at the University of Essex, who furnished the Soundex codes from the KLEIO program – a software program which striver to stay as close as possible to the original source material.

[7] The procedure is fully explained in M. Thaller, *Kleio: a database system* (Halbgraue Reihe zur Historischen Fachinformatik B11, St Katharinen, 1993), pp. 126-8. It is similar to that described by I. Winchester, 'The linkage of historical records by man and computer: techniques and problems', *Journal of Interdisciplinary History* (1970), pp. 107-24.