Journal of Official Statistics, Vol. 14, No. 3, 1998, pp. 255–266

Linking of Classifications by Linear Mappings

Beat Hulliger[1]

Statistical classifications must be revised from time to time. In order to compare different time points the links between the versions of a classification must be established. Correspondence tables describe these links. This article formulates the links as linear mappings and the correspondence tables as matrices. The evolution of a classification or the linking of a chain of several classifications may then be described by matrix products. The aggregation to higher levels of a classification is again a linear mapping and therefore any correspondence table between any two levels of two classifications may be derived by a product of basic matrices.

Key words: Nomenclature; correspondence table; revision of classification; time series.

1. Introduction

A statistical classification or nomenclature is a systematic set of mutually disjoint categories, called items. The names of the categories are unique and are called codes. The description of the content of a category is contained in a title and possibly in explanatory notes. The application of a classification to the units of a population, i.e., the determination of the category into which a unit falls, is called coding. Every unit of the population should be classifiable.

Often classifications are hierarchical, i.e., the items of a first level are subdivided into items of a second level and so on until a basic level is reached. A well known example of such a tree-like classification is the regional subdivision of a country into provinces, districts and communes.

A classification defines which elements of a population are similar enough to be counted as the same (Kotz, Johnson, and Read 1982, Vol. 2, p. 1) and therefore classifications are at the basis of statistics. Hierarchical classifications are used extensively to aggregate data. Standard classifications are a key element to make statistical information comparable over time and space.

However, classifications must be revised from time to time either because the population changes or because the use of the classification changes. Revisions may concern the titles or explanatory notes only or they may concern the structure of a classification. Here we treat only structural changes. In hierarchical classifications structural changes may concern only one level of a classification or they may involve two or more levels.

[1] Swiss Federal Statistical Office, CH-3003 Berne, Switzerland.

Acknowledgment: The author thanks his colleagues, in particular Philippe Eichenberger and Paul-André Salamin, the Associate Editor and two anonymous referees for their helpful comments.

There are four types of ‘‘one-level’’ changes:

1. Birth of an item.
2. Death of an item.
3. Split of one item into many items.
4. Fusion of many items into one item.

Of course combinations of these four changes may occur. Splits and fusions may be mixed such that changes from many items to many items result. A birth of an item means that there is no item in the old version of a classification which from a theoretical point of view can be seen as a predecessor. In practice, any population unit which is classified into a newborn item of the new version should be classifiable under the old version, too. Thus neither births nor deaths of items can be observed.

For example Hoffmann and Scott (1990) report on the changes of the International Standard Classification of Occupations (ISCO) from the 1968 to the 1988 version; 157 of the basic level items of ISCO-68 were not changed, 31 fused into 14 new items, 96 split into 174 new items and there were 32 births in ISCO-88. Thus there were 32 new items in ISCO-88 where no reference to an item of ISCO-68 could be assigned.

A change may affect two contiguous levels of a classification. Thus an item may be promoted to an upper level or be relegated to a lower level. For example the city of Zürich was promoted to form an own district. Or, which is usually more frequent, an item may change its parent item. For example the communes of the district of Laufen changed from the canton of Berne to the canton of Basel-Land. All other changes may be interpreted as combinations of one- and two-level changes.

In practice a complete record of all structural changes is necessary to establish time series for periods extending over several versions of a classification. The record of one revision is usually called a correspondence table (see e.g., EDI Expert Group 6 1996 or Lestang 1983). The more complex and frequent revisions occur the more difficult becomes the coherent recording and tracking of revisions.

This article shows how simple matrix algebra may be used to keep track of the revisions of a classification. It is shown that it is only necessary to keep track of the ‘‘one-level’’ changes at the basic level of a nomenclature while the ‘‘one-level’’ changes of aggregate levels can be derived by matrix algebra from the changes at the basic level and from the aggregations of the different versions.

In Section 2 the aggregation of items to their parent items is described as a linear mapping. The evolution of the items of a level, usually described by a correspondence table, is a linear mapping again. For example when a split occurs an item is mapped to its successors. The matrix of this mapping and its suitable standardisation are described in Section 3. In Section 4 the derivation of weighted correspondence matrices from observations is explained. In Section 5 the correspondence table resulting from several revisions is derived as a matrix product. The extension to correspondence tables of higher levels and to whole nets of classifications is shown in Sections 6 and 8. An application to the cantons of Switzerland is presented in Section 7. Section 8 concludes with some limitations of the approach and indications for future research.

2. Aggregation

A hierarchical relation between a lower and an upper level of a classification may be described as a linear mapping. Suppose there are n lower level items which are grouped into a upper level items. Suppose that the items of each level are ordered and each item has its order number within its level.

To each lower level item with order number $i(i = 1, \ldots, n)$ belongs an $n \times 1$ identification vector y_i which has all entries 0 except the i-th entry which is 1. The set of vectors $\{y_1, \ldots, y_n\}$ forms the basis of a vector space $\mathcal{Y}$. Similarly, to each upper level item $j(j = 1, \ldots, a)$ belongs an $a \times 1$ identification vector u_j. The set of vectors $\{u_1, \ldots, u_a\}$ forms the basis of a vector space $\mathcal{U}$. The $n \times a$ matrix A, with entries $A_{ij} = 1$ if lower level item i belongs to upper level item j and 0 otherwise, describes the grouping or aggregation completely.

When $A^\top$, the transpose of A, is applied to an identification vector of a lower level, y_i say, then the identification vector of its parent item results. The matrix $A^\top$ corresponds to a linear mapping of the space $\mathcal{Y}$ to the space $\mathcal{U}$, i.e., $A^\top : \mathcal{Y} \rightarrow \mathcal{U}$ (The same notation is used here for the matrix and the linear mapping that corresponds to it.) The mapping is not one-to-one since the number of items is usually greatly reduced by an aggregation. Conversely when A is applied to an identification vector of the upper level, then the result is the sum of the identification vectors of the corresponding lower level items. For an introduction to matrix algebra see e.g., Rao (1973). In the following examples the items are denoted by their identification vector.

Example 1. Suppose there are five lower level items, the first three items belong to a first parent and the fourth and fifth belong to a second parent. The aggregation matrix is

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \tag{1}$$

The first row of A means that y_1 belongs to u_1, but not to u_2. Thus the first column of A has an entry of 1 for the lower level items that belong to u_1. When $A^\top$ is applied to the identification vector of item y_4, one gets

$$A^\top y_4 = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} = u_2 \tag{2}$$

which shows that y_4 belongs to u_2. Conversely

$$Au_2 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} = y_4 + y_5 \tag{3}$$

which shows that u_2 disaggregates into y_4 and y_5.

Any further aggregation to b items v in a space $\mathcal{V}$, say, may be described by a further $a \times b$ aggregation matrix, say B. Then the direct aggregation from y to v is described by the matrix product of transposes $B^\top A^\top$. This holds because the composition of linear mappings is equivalent to the product of the corresponding matrices. In other words the linear mappings $A^\top : \mathcal{Y} \rightarrow \mathcal{U}$ and $B^\top : \mathcal{U} \rightarrow \mathcal{V}$ may be combined to $B^\top A^\top : \mathcal{Y} \rightarrow \mathcal{V}$.

3. Changes at One Level and Standardisation

The changes involving only one level of a classification (cf. Section 1) may be described as linear mappings, i.e., by matrices. Suppose the identification vectors for Versions 1 and 2 are $x_i (i = 1, \ldots, m)$ and $y_j (j = 1, \ldots, n)$, respectively. The $m \times n$ matrix C, with elements $C_{ij} = 1$ if item j is a successor of i and 0 otherwise, describes all possible changes. We call C a correspondence matrix. Often a correspondence table between two versions of a classification or between two different classifications is written as a list with two columns (e.g., Annex II of the Industrial Commodity Statistics Yearbook, 1994 (1996)). Each row in such a list corresponds to one of the cells of the matrix C except that the cells with a 0 entry are left out. Thus C is nothing else than a correspondence table in its full extension.

The product $C^\top x_i$ gives for any identification vector x_i for Version 1 the sum of the identification vectors of its successors in Version 2. In fact multiplying x_i by $C^\top$ amounts to picking out of C the row which corresponds to x_i. Of course $C^\top x_i$ may have several entries equal to 1 if item i has split or it may have all entries equal to 0 if item i has died. Conversely Cy_j gives the predecessors of item j. If Cy_j has several entries equal to 1 then item j is the result of a fusion. If it has only 0's then item j is a birth. The matrix $C^\top$ corresponds to a linear mapping of the vector space $\mathcal{X}$ of Version 1, which is spanned by the basis $\{x_1, \ldots, x_m\}$, to the vector space $\mathcal{Y}$ of Version 2, which is spanned by $\{y_1, \ldots, y_n\}$, i.e., $C^\top : \mathcal{X} \rightarrow \mathcal{Y}$ and $C : \mathcal{Y} \rightarrow \mathcal{X}$.

It is convenient to standardise the column-sums of $C^\top$ to 1, except if all entries in a column are 0. If a column of $C^\top$ has several non-zero entries it describes a split and the entries in the standardised column are the weights of the split. The choice of the weights in a split depends on the particular analysis to be performed (cf. Section 4). A simple solution, which is applied in the following examples, is to give equal weight to all the successors of an item. We denote the column-sum standardisation of matrices with a star like in C^*. The column-sum standardised transpose of C is $C^{\top *}$. Note that the order of standardisation and transposition is important, i.e., $C^{\top *} \neq C^{* \top}$.

The four basic changes of Section 1 are now represented in C^* as follows: A birth is a column of 0's. A death is a row of 0's. A split is a row, where several entries are larger than 0. A fusion is a column where several entries are larger than 0 and since C^* is column-sum standardised the non-zero entries sum to 1. Usually the non-zero entries of a ''split-row'' are 1 except if part of the split is fused with another item (''many-to-many'' changes). Items that do not change have exactly one 1 in their row and column, respectively, and the remaining entries are 0.

Example 2. Let the old version of a classification contain four items and the new Version 3. Suppose the new third item is created from parts of the old first and second items, that the other part of the old second item is fused with the old third item, and that the fourth old item dies. The resulting correspondence matrix is

$$C = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{4}$$

The successors of x_1 are

$$C^{\top *} x_1 = \begin{bmatrix} 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 1 & 0 \\ 1/2 & 1/2 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 0 \\ 1/2 \end{bmatrix} = \frac{1}{2} y_1 + \frac{1}{2} y_3 \tag{5}$$

i.e., the successors of x_1 are y_1 and y_3 and both receive 50% of x_1. The successors of x_3 are

$$C^{\top *} x_3 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \tag{6}$$

i.e., the only successor of x_3 is y_2. The predecessors of y_2 are

$$C^* y_2 = \begin{bmatrix} 1 & 0 & 1/2 \\ 0 & 1/2 & 1/2 \\ 0 & 1/2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1/2 \\ 1/2 \\ 0 \end{bmatrix} = \frac{1}{2} x_2 + \frac{1}{2} x_3 \tag{7}$$

both contributing 50% to y_2. The predecessors of y_3 are

$$C^* \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 1/2 \\ 0 \\ 0 \end{bmatrix} = \frac{1}{2} x_1 + \frac{1}{2} x_2 \tag{8}$$

both contributing 50%.

An aggregation matrix is a special case of a correspondence matrix: there are only fusions in an aggregation, as the word suggests. Therefore, aggregation matrices are already row-sum standardised, i.e., $A^{\top *} = A^{\top}$. However, when used to disaggregate, A should be column-sum standardised and Equation (3) becomes

$$A^* u_2 = \begin{bmatrix} 1/3 & 0 \\ 1/3 & 0 \\ 1/3 & 0 \\ 0 & 1/2 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \frac{1}{2} y_4 + \frac{1}{2} y_5 \tag{9}$$

i.e., item y_4 and y_5 contribute 50% each to its parent item. The equal weights of A^* may, of course, not be appropriate. In Section 4 a weighted aggregation matrix is derived from the observations of a variable, thus leading to possibly unequal weights.

4. Weighted Correspondence Matrices

A correspondence matrix may be established from the data of a census or a sample, i.e., from the units which are classified by the considered classifications. For this purpose the data for at least one reference period must be coded according to the two classifications which should be linked (MacDonald 1995). Suppose that for a certain point in time there are three variables available for a population of size N : c_1, c_2 and α. The two vectors c_1 and c_2 give the position of the units of the population in two classifications, i.e., they are equivalent to the codes of the two classifications assigned to the individual units. The vector α is a positive variable, measured for every unit, which shall be used to establish weights. For example α may contain the number of employees of enterprises, c_1 the codes of an old classification of economic activities and c_2 the codes of a new classification of economic activities.

Let m be the number of items in Classification 1 (cf. Section 3). Let C_1 be the $N \times m$ matrix with elements $(C_1)_{ki} = 1$ if $c_{1k} = i$ and 0 otherwise. Similarly define C_2. Then C_1 is the matrix which corresponds to a linear mapping of the units of the population into the items of the first classification and C_2 maps the units into the second classification. Both matrices may contain 0-columns corresponding to items that have not been observed. The aggregates a_1 of α for Classification 1, which are simple subtotals over the units belonging to the same items of this classification, may also be obtained as $a_1 = C_1^\top \alpha$, while the aggregates for Classification 2 may be obtained as $a_2 = C_2^\top \alpha$. The correspondence matrix between the two classifications weighted with the variable α is then

$$C_{12\alpha} = C_1^\top \mathrm{diag}(\alpha) C_2 \tag{10}$$

where diag(α) is the $N \times N$ matrix with diagonal α and 0 elsewhere. For $\alpha_k = 1$, $(k = 1, \ldots, N)$ the weighted correspondence matrix $C_{12\alpha}$ is just the contingency table of the two classifications. Thus for general α we may obtain $C_{12\alpha}$ easily as an α-weighted contingency table.

The weighted correspondence matrix $C_{12\alpha}$ allows the direct mapping from a_1 to a_2 and back:

$$a_1 = C_{12\alpha}^{*} a_2 \tag{11}$$

$$a_2 = C_{12\alpha}^{\top *} a_1 \tag{12}$$

Suppose we know the aggregates b_1 of another variable β according to Classification 1 but we are not able to aggregate β according to Classification 2. The correspondence matrix $C_{12\alpha}^{\top *}$ may be applied to b_1 to obtain an estimate $\hat{b}_2$ of the aggregates according to Classification 2, namely

$$\hat{b}_2 = C_{12\alpha}^{\top *} b_1 \tag{13}$$

The vector β may stem from another source than c_1, c_2 and α. The quality of the estimate $\hat{b}_2$ depends on the correlation between α and β. If $C_{12\alpha}^{\top *}$ contains 0-columns because no observation was coded in the corresponding category of Classification 1 then the sum over $\hat{b}_2$ may not correspond to the sum over b_1. Nevertheless, this technique may help to establish time series of aggregates when classifications change.

Higher level aggregates of α may be written as matrix products, too. For example let A_1 denote an aggregation matrix for Classification 1. Then

$$A_1^\top a_1 = A_1^\top C_1^\top \alpha \tag{14}$$

yields the aggregates of α according to a higher level of Classification 1.

A weighted aggregation matrix $A_{1\alpha}$ may be derived from A_1 and a variable α by

$$A_{1\alpha} = \operatorname{diag}(C_1^{\top}\alpha)A_1 \tag{15}$$

The column-sum standardised matrix $A_{1\alpha}^{*}$ may then be used for disaggregation.

5. Concatenation

Denote the correspondence matrix for the changes between Version 1 with m items and Version 2 with n items by C_{12}. A further change to Version 3 has its correspondence matrix C_{23}. Let $z_k (k = 1, \ldots, l)$ denote the identification vectors for Version 3 and $\mathcal{Z}$ the space spanned by $\{z_1, \ldots, z_l\}$. Then C_{23} has dimension $n \times l$. The correspondence between Version 1 and Version 3 is described by the matrix product of C_{12} and C_{23}, i.e., by

$$C_{13} = C_{12}C_{23} \tag{16}$$

This holds because the composition of two linear mappings is expressed by the product of their matrices: The mappings $C_{12}^{\top} : \mathcal{X} \rightarrow \mathcal{Y}$ and $C_{23}^{\top} : \mathcal{Y} \rightarrow \mathcal{Z}$ are combined to $C_{23}^{\top}C_{12}^{\top} : \mathcal{X} \rightarrow \mathcal{Z}$.

Using standardised correspondence matrices in (16) we obtain a standardised product directly, because for every $i (i = 1, \ldots, m)$

$$\sum_{k=1}^{l} \left(C_{23}^{\top *} C_{12}^{\top *}\right)_{ki} = \sum_{k=1}^{l} \left[\sum_{j=1}^{n} \left(C_{23}^{\top *}\right)_{kj} \left(C_{12}^{\top *}\right)_{ji}\right] = \sum_{j=1}^{n} \left[\sum_{k=1}^{l} \left(C_{23}^{\top *}\right)_{kj}\right] \left(C_{12}^{\top *}\right)_{ji} \tag{17}$$

and since $\sum_{k=1}^{l}(C_{23}^{\top *})_{kj}$ is 0 or 1 due to the column-sum standardisation of $C_{23}^{\top *}$ and $\sum_{j=1}^{n}(C_{12}^{\top *})_{ji}$ is 0 or 1 for the same reason, the expression becomes either 0 or 1.

Example 3. Suppose C_{12} is the correspondence matrix in (4) and

$$C_{23} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \tag{18}$$

i.e., y_1 becomes z_1, y_2 becomes z_4 and y_3 splits into z_2 and z_3. Then

$$C_{13}^{*} = C_{12}^{*}C_{23}^{*} = \begin{bmatrix} 1 & 0 & 1/2 \\ 0 & 1/2 & 1/2 \\ 0 & 1/2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \tag{19}$$

$$= \begin{bmatrix} 1 & 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{20}$$

Thus the changes from Version 1 to Version 3 are as follows: Two parts of the split of item x_1 are fused to z_2 and z_3; The parts of the split of item x_2 are fused to z_2, z_3 and z_4 and item x_3 is fused to z_4, too. Item x_4 dies. Note that the column-sums of C_{13}^{*} are automatically standardised.

Suppose we have registered all revisions of a classification and their correspondence matrices C_t, where t indicates the time of the revision. If we are interested in the correspondence matrix C_{t_1,t_2} between the versions of the nomenclature for time t_1 and t_2 then we just have to calculate the product of the correspondence matrices C_t with $t \in [t_1, t_2]$:

$$C_{t_1,t_2} = \prod_{t \in [t_1,t_2]} C_t$$

6. Changes Involving Two Levels

Changes of upper level items do not necessarily affect the lower level items as such but the hierarchical links certainly change whenever a change in an upper level occurs. Suppose the hierarchical links for Version 1 are described by the $m \times a$ aggregation matrix A_1 and for Version 2 by the $n \times b$ aggregation matrix A_2 and the vector spaces of the lower levels are $\mathcal{X}$ and $\mathcal{Y}$ while the vector spaces of the upper levels are $\mathcal{U}$ and $\mathcal{V}$. Let C_{12} describe the evolution of the lower level. Then

$$A_2^\top C_{12}^{\top *} x_i \tag{21}$$

describes the correct aggregation groups for Version 2 of an item with identification vector x_i in Version 1. Thus the mappings $C_{12}^{\top *} : \mathcal{X} \rightarrow \mathcal{Y}$ and $A_2^\top : \mathcal{Y} \rightarrow \mathcal{V}$ are combined to $A_2^\top C_{12}^{\top *} : \mathcal{X} \rightarrow \mathcal{V}$.

Conversely

$$A_1^\top C_{12}^{*} y_j \tag{22}$$

gives the aggregations according to Version 1 of the predecessors of item j, i.e., $A_1^\top C_{12}^{*} : \mathcal{Y} \rightarrow \mathcal{U}$.

Example 4. Let

$$A_1 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \quad C_{12} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad A_2 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{23}$$

There are five basic items in Version 1 and four in Version 2. The first two basic items fuse and part of the fourth item is fused with the fifth item. Now $A_2^\top C_{12}^{\top *} x_3 = [0, 1/2, 1/2]^\top$, i.e., the new aggregates of the old Item 3 are the second and third in the list of upper level items of Version 2, and the contribution is 50% to each new aggregate. Conversely $A_1^\top C_{12}^{*} y_1 = [1, 0]^\top$. Thus the old aggregation group of new basic Item 1 is the first in the list of upper level items of Version 1.

Promotion, relegation and change of the parent item are easily reflected by changes in corresponding aggregation matrices. A condition is that the classification is built in such a way that every item which is not at the basic level has at least one lower level item as a child.

It is possible to describe the direct change of any upper level items by the concatenation of the hierarchical mappings with the correspondence mapping of the basic level. Let C_{12} again be the correspondence matrix of the basic level and A_1, A_2 the corresponding aggregation matrices. The matrix product

$$D_{12}^{\top} = A_2^{\top} C_{12}^{\top *} A_1^{*} \tag{24}$$

describes the direct correspondence matrix between the upper levels in Versions 1 and 2. In other words $D_{12}^{\top} : \mathcal{U} \to \mathcal{V}$. Note that $D_{12}^{\top}$ is automatically column-sum standardised. Whether the weights in $D_{12}^{\top}$, which stem from the lower levels, are appropriate must be decided in view of the application of the correspondence matrix.

Example 5. With the matrices of the last example we get

$$D_{12}^{\top} = \begin{bmatrix} 1 & 0 \\ 0 & 1/4 \\ 0 & 3/4 \end{bmatrix} \tag{25}$$

Thus the second upper level item of Version 1 is split into the second and third upper level item of Version 2 with weights 25% and 75%, respectively.

7. Application to the Cantons of Switzerland

The division of Switzerland into cantons is a very stable nomenclature since the last century. Nevertheless, on January 1, 1979 the canton of Jura (JU) was created as a split from the canton of Berne (BE), and on January 1, 1994 the district of Laufen was split from the canton of Berne and joined to the canton of Basel-Land (BL). Thus we have three versions of the nomenclature of cantons. Version 1 is valid up to December 31, 1978, Version 2 is valid from January 1, 1979 to December 31, 1993 and Version 3 is valid from January 1, 1994 onwards. Table 1 shows the number of permanent residents s per canton at the end of the years 1970, 1978, 1993, and 1995.

At the beginning of 1979 64,800 of the residents of Berne established the new canton of Jura and at the beginning of 1994 15,470 persons, i.e., the district of Laufen, changed from Berne to Basel-Land. We use these figures for the weighting of the correspondence matrices. The first weighted correspondence matrix C_{12} has dimension 25×26 and contains the number of resident persons of 1978 (in thousands) in the diagonal and zero elsewhere, except for Row 2, corresponding to Berne, where the second entry is 908.3 and the last entry is 64.8. The second weighted correspondence matrix C_{23} has dimension 26×26 and contains the number of resident persons of 1993 in the diagonal and zero elsewhere, except for Row 2, where the second entry is 941,147 and the 13th entry, corresponding to Basel-Land, is 15,470.

Suppose we would like to compare the number of resident persons according to the 1970 census with the number of resident persons at the end of 1995 for regions which are aggregates of cantons (see column headed ''region'' in Table 1). We may use Formula (16) to derive C_{13s}, the correspondence matrix weighted by the number of resident persons s. Then we apply (13) to get an estimate at cantonal level and (14) to obtain the appropriate aggregates. In this simple example it is easy to see explicitly what happens. For example, to derive $\hat{s}_{95,1}$ from $s_{95,3}$ we first join the figure for Jura back to Berne (Region 2 gets 69,188 from Region 1). Then we calculate an estimate for the population of 1995 of the district of Laufen based on its relative weight in Basel-Land on January 1, 1994

Table 1. Number of permanent residents in cantons

canton[a]	region[b]	s_{70}	s_{78}[c]	s_{93}	s_{95}
ZH	3	1107,788	1,109.8	1162,083	1175,457
BE	2	983,296	973.1	956,617	941,952
LU	4	289,641	292.6	335,385	340,536
UR	4	34,091	33.7	35,727	35,876
SZ	4	92,072	95.1	118,528	122,409
OW	4	24,509	25.4	30,837	31,310
NW	4	25,634	27.7	35,393	36,466
GL	3	38,155	36.1	39,138	39,410
ZG	4	67,996	73.9	88,583	92,392
FR	1	180,309	183.2	218,704	224,552
SO	2	224,133	216.3	236,389	239,264
BS	3	234,945	208.3	197,403	195,759
BL	3	204,889	215.4	234,910	252,331
SH	3	72,854	69.0	73,588	74,035
AR	3	49,023	46.9	54,087	54,104
AI	3	13,124	12.8	14,680	14,750
SG	3	384,475	385.1	436,967	442,350
GR	5	162,086	159.2	181,957	185,063
AG	2	433,284	443.7	518,945	528,887
TG	3	182,835	181.1	217,129	223,372
TI	5	245,458	262.0	297,955	305,199
VD	1	511,851	516.9	596,736	605,677
VS	5	206,563	215.1	266,713	271,291
NE	1	169,173	158.6	163,884	165,258
GE	1	331,599	344.2	387,606	395,466
JU	1	NA	NA	68,626	69,188
CH		6269,783	6,285.2	6968,570	7062,354

s_{yy} is the number of permanent residents at the end of year 19*yy*.

[a]The cantons are given in the so-called historical order.

[b]The regions are not an official nomenclature.

[c]The figures of 1978 are given in thousands.

Table 2. Number of permanent residents in regions

Region	Version 1			Version 3	
	$s_{70,1}$	$\hat{s}_{95,1}$	$s_{95,1}$	$\hat{s}_{70,3}$	$s_{95,3}$
1	1192,932	1390,953	1390,953	1258,411	1460,141
2	1640,713	1794,882	1795,556	1560,391	1710,103
3	2288,088	2455,977	2455,303	2302,931	2471,568
4	533,943	658,989	658,989	533,943	658,989
5	614,107	761,553	761,553	614,107	761,553
CH	6269,783	7062,354	7062,354	6269,783	7062,354

Columns with $s_{yy,c}$ are counts, $\hat{s}_{yy,c}$ are estimates. The year is indicated by *yy*, the applied nomenclature by *c*.

$(252{,}331 \cdot 15{,}470/(15,470 + 234,910) = 15,591)$ and we join this population back to Berne (Region 2 gets 15,591 from Region 3).

Table 2 shows the results for Version 1 and Version 3 of the nomenclature of cantons. While the regional aggregates of s_{70} and s_{95} according to their respective classification Versions 1 and 3, i.e., $s_{70,1}$ and $s_{95,3}$, cannot be compared directly, the comparisons based on Version 1, i.e., $s_{70,1}$ and $\hat{s}_{95,1}$, and on Version 3, i.e., $\hat{s}_{70,3}$ and $s_{95,3}$, are valid because they are not confounded with the change of the nomenclature. In this example we may derive the true regional aggregates $s_{95,1}$ because we know from other sources that the district of Laufen had 16,265 resident persons in 1995. This would not be possible in a situation where the units (here the communes) cannot be reclassified according to Version 1. Table 2 shows that the difference between the estimate $\hat{s}_{95,1}$ and the true aggregate $s_{95,1}$ is noticeable but small for most purposes.

8. Extensions, Limits and Further Research

Two different classifications may be linked by a correspondence matrix in the same way as two versions of the same classification. Matrix products will show the correct correspondence to other versions of the two classifications and to further related classifications. Thus a whole net of classifications may be connected and matrix calculus may help to keep these connections consistent. An example of such a set of (unweighted) correspondence tables between commodity codes based on the Standard International Classification of All Economic Activities (ISIC Rev. 2), three versions of the Standard International Trade Classification (SITC) and the Harmonized Commodity Description and Coding System (HS) is described in Annex II of the Industrial Commodity Statistics Yearbook, 1994 (1996).

The concatenation of correspondence matrices has no memory, i.e., the transition from the current version to the next version may depend only on the current and next versions and not on past versions. This limitation shows up in the following example (Syvänperä 1995). Suppose two items a and b were fused into c, but after a while c splits again into its former parts a and b. The product of the two correspondence matrices involved yields not only a path from a to c to new a but also from a to c to new b, which of course is undesirable. The problem is that the split from c to new a and b uses past information. A possible solution to the problem is to create an intermediate aggregation level and to treat the temporary fusion of a and b as an aggregation to an intermediate level item c. Thus the fusion and split of a and b would be reflected only in the aggregation matrices while the basic level correspondence matrices preserve the units.

Correspondence matrices are sparse matrices with many 0's. The implementation of this matrix calculus for the linking of classifications should store the correspondence matrices in an efficient way. It is clear that all changes have a reason and therefore more information than just the changes must be recorded. At the same time attributes which give information about the non-structural changes may have to be stored, too.

This article gives a theoretical framework for the evolution of classifications and for weighted correspondence tables. However the establishment of weighted correspondence tables is further complicated by misclassifications, differences between data-derived and theoretical correspondence tables and possible time lags between the coding according to different versions of a classification. These problems and their effect on the quality of estimates based on weighted correspondence tables must be studied further.

9. References

EDI Expert Group 6 on Statistics/European Board for EDI Standardization (1996). Concepts and Descriptors related to Classifications, Document EEG6/WG3/96005. Luxembourg: Eurostat.

Hoffmann, E. and Scott, M. (1990). The Revised International Standard Classification of Occupations (ISCO-88): A Short Presentation. In Developments in International Labour Statistics, R. Turvey (ed.). London: Pinter Publishers.

Industrial Commodity Statistics Yearbook 1994 (1996). New York: United Nations Publication, Series P, No. 34.

Kotz, S., Johnson, N.L., and Read, C.B. (1982). Encyclopedia of Statistical Sciences. New York: Wiley.

Lestang, P. (1983). La règle du trépied. Paris: INSEE, internal note. (In French).

MacDonald, B. (1995). Implementing a Standard Industrial Classification (SIC) System Revision. In Business Survey Methods, B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M. Colledge, Ph. S. Kott (eds.). New York: Wiley.

Rao, C.R. (1973). Linear Statistical Inference and Its Applications (2 ed.). New York: Wiley.

Syvänperä, R. (1995). Personal Communication.

Received May 1996
Revised May 1997