A Statistical Approach for Reconstructing Continuous Series of Mortality by Cause of Death

Carlo G. Camarda
Max Planck Institute for Demographic Research
Laboratory of Statistical Demography
camarda@demogr.mpg.de

Draft of the paper for the
Population Association of America 2012
Annual Meeting

September, 2011

Abstract

Data on cause-specific mortality have been collected over several decades for different countries. Nevertheless this valuable source of information has not been sufficiently explored. This is mainly due to frequent revisions of the International Classification of Diseases which lead to discontinuities and disruptions in cause-specific mortality trends. This paper proposes a new methodology for reconstructing continuous series of mortality by cause of death. The model assumes death counts as realizations from a Poisson distribution. The expected values of this distribution is a composed mean which embodies the continuous series and the redistribution pattern among causes of death. An algorithm allows the simultaneous estimation of these two components. The known associations among causes is used as input and the only assumption that is made about the underlying continuous series is smoothness. We present a simulation study and an application for demonstrating the performance of our model and its practical characteristics.

Keywords: Cause-specific mortality, continuous series, statistical model, composite link model, smoothing.

1 Introduction

Unlike medical and epidemiological research, demographic studies often use data on whole populations. Sex, age and cause of death (CoD) for such populations are the fundamental (and often the only) information available. While deaths are easily categorized by sex and age, the classification by cause of death is not unique: it can change with time and it is peculiar to each population.

Nowadays, a widely used system for coding data on causes of death is the WHO International Classification of Diseases and Related Health Problems (ICD). Since its first appearance in 1893, this classification was continuously upgraded and revised to reflect progress of medical knowledge. Consequently, with every new revision, the time series of mortality by cause of death are interrupted. National statistical offices rarely produce a double classification (cross-tabulation of deaths by both the actual and the previous revision) that would make it possible to directly redistribute the deaths of the previous periods according to the new classification. Furthermore over time and among countries, coding practices were changed and deaths without or insufficient further diagnostics, i.e. uncertain classification, occurred. As a result, time series of mortality often exhibit disruptions, especially after changes in ICD, disruptions which are undue to actual changes in mortality trends.

In the light of such issues, a pressing problem in mortality studies concerns the correct reconstruction of continuous time-series of death counts by CoD. Important results have been obtained using a methodology which redistributes counts in an earlier period among causes from the newer classification. Among others, see Meslé and Vallin (1996), Janssen and Kunst (2004), Pechholdová (2009). The approach basically uses a "correspondence table" that provides information about which CoD correspond between the two periods. Using medical and clinical knowledge, and aiming to reasonable trends for each cause, "transition coefficients" for each of these correspondences are identified. Visual inspection of the trends by each CoD are crucial for defining such coefficients and for eventual posterior corrections which take into account also age and sex.

Such approach involves a lot of manual toil and often requires subjective adjustments. On the contrary in this paper we suggest a novel methodology which combines statistical analysis and demographic knowledge. Briefly, we assume that deaths by each age, sex and CoD are realizations from a Poisson distribution (Keiding, 1990) and that the expected values of this distribution is a composed mean. Such composition embodies the two parts of the model: the continuous series and the redistribution pattern among CoDs. In this way the two components can be simultaneously estimated.

The composite link model (CLM), proposed by Thompson and Baker (1981), provides an elegant framework for modelling data realizations of composed means. It is an extension of the generalized linear model (GLM) (McCullagh and Nelder, 1989) and can itself be extended in order to estimate "transition coefficients" too.

The paper is structured as follows. A typical example of disruption due to change in ICD will set the stage in the following section, after which we will shortly introduce the proposed model and a suitable estimation procedure. In Section 4, we illustrate the approach via simulated data and present some applications. A critical discussion of the method concludes the paper.

2 An example of disruption

As an example for disruption due to change in the coding we present the case of Belorussian deaths by specific heart diseases between two "Soviet Nomenclatures": 1965-1969 and 1970-1980[1]. Table 1 lists the CoDs in the data and Figure 1 gives a clear picture of the time series and their disruptions occurred in 1970.

	Code	**Cause of Death**
1965-1969	[1]	Cerebrovascular disorders with hypertension (88)
	[2]	Cerebrovascular disorders with hypertension and cerebral arteriosclerosis (90)
	[3]	Hypertensive heart disease (110)
	[4]	Cerebral hypertensive disease except disorders of the central nervous system (111)
	[5]	Heart and cerebral hypertensive disease (113)
1970-1980	[6]	Hypertensive heart disease (86)
	[7]	Hypertensive heart and renal disease (90)
	[8]	Atherosclerotic cardiosclerosis with hypertensive disease (92)
	[9]	Cerebrovascular disorders with hypertensive disease (98)

Table 1: Specific heart causes of death in Belarus from 1965 to 1980. In parenthesis the original codes used in the "Soviet Nomenclatures".

We have five and four CoDs in the first and second period, respectively. Both causes 110 and 86 collect deaths due to hypertensive heart diseases, but they are differently coded in the original documents, we will thus consider them as different CoDs. This choice is also supported by the clear break in the time series of these CoDs (see Fig. 1).

The CoDs in this dataset belongs to the same group, therefore we can assume that all deaths due to old CoDs would have been classified within the new CoDs, if they existed during the first period. The aim is to reconstruct mortality trends in the first period as (back-)projection of the time series of the new CoDs. This means that we need to redistribute deaths occurred in the first period among new CoDs.

Changing in medical practices as well as development of new and unknown CoDs can lead to specific irregularities in a mortality time series by CoD. However, the specific reasons for such irregularities are usually well understood from the medical and epidemiological perspectives. In the absence of such knowledge, actual cause-specific series of deaths are mainly the result of mortality development over time and thus the assumption of a smooth change in mortality is the most reasonable and flexible. This implies that disruptions in specific points in time are

[1] The "Soviet Nomenclatures" was based on the ICD system provided by the WHO. Data for this paper were kindly provided by Pavel Grigoriev and they will be soon published in the MPIDR Working Papers series.

only the outcomes of coding revision. If irregularities in the series are due to known events, rather than change in classification, the associated years will be excluded from the smoothing procedure.

Furthermore, smooth curves can be considered continuous function and therefore, from the estimated series, we could compute derivatives over time avoiding erratic behavior due to random fluctuations. This is surely an additional advantage of using smoothness assumption for the latent cause-specific mortality series.

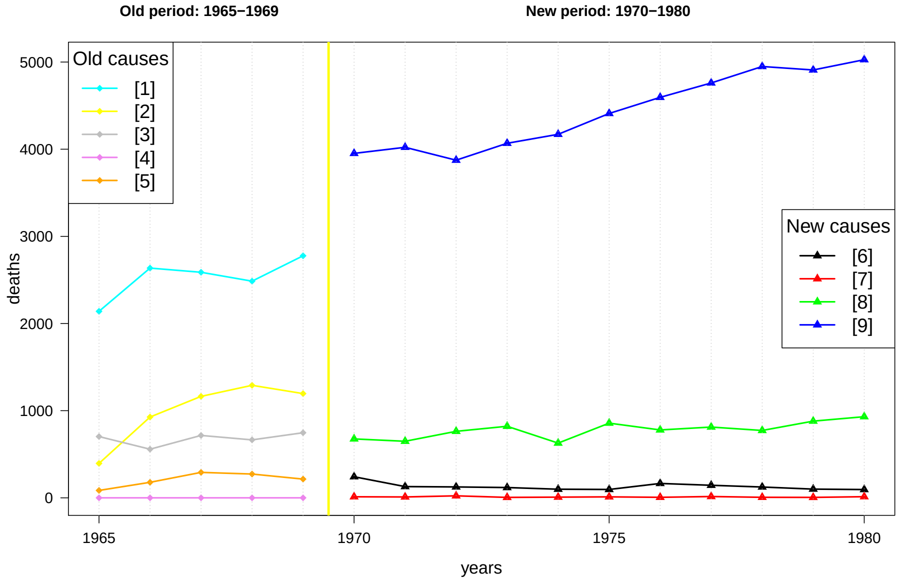


Figure 1: Death counts by specific heart diseases in Belarus, all ages and both sex combined.

The observed death counts can be viewed as the outcome of a classification revision that transforms a smooth, but latent time series into observed data. In other words, data before the coding revision are assumed to be deaths due to the new (but still not in use) causes redistributed within the old causes. Hence the redistribution pattern needs to be estimated together with the smooth time series. Alternatively known mortality functions could be incorporated to describe trends over time. However, they enforce rigid patterns and they are thus not considered in the following.

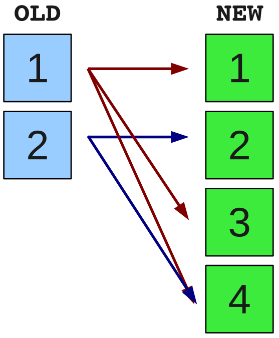
3 The model

In the current version, the model considers both sexes and all ages combined, and only two classification periods. Moreover, as mentioned, at this stage, we will assume that disruptions in the mortality trends are only due to switch in the official classification, without considering variations in coding practices.

We define as "correspondence table" a matrix which provides information about possible exchange among CoDs in the two revisions. It can be succinctly written as a logical or $(0, 1)$ matrix in which rows are indexed by cause in the first period and columns are indexed by cause in the second period. Medical knowledge and a deep understanding of the changes between two ICD revisions are necessary for setting up a "correspondence table". Such matrix can thus be considered as an input in our model and, given all eventual exchanges among causes between two revisions, a "correspondence table" has a unique definition.

Let's present two examples of changes in ICD and the associated "correspondence table". On the left we will schematize the transitions among causes, and on the right side we will summarize such transitions into the related "correspondence table". In Section 6 we will present the "correspondence table" associated with the Belorussian data.

Example 1 In this first example, we have two causes in the old period, namely $[1, 2]$, and four causes in the new period $[1, 2, 3, 4]$.

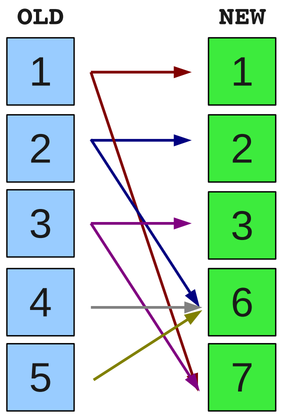


$$\boldsymbol{T} = \begin{array}{c c} & \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{array}{c} 1 \\ 2 \end{array} & \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix} \end{array} .$$

Each CoD in the first period has a correspondence in the second period (see the ones on the diagonal of the first 2 columns). New causes [3] and [4] collect part of the death counts which were previously classified only in cause [1] (see the ones in the first row at the third and fourth column). Meantime cause [4] also collects part of the deaths which were belonging to cause [2] in the old revision (see the one at the lower-right corner).

Otherwise stated, all the death counts classified in CoD [1] in the old period will be split in the new CoDs $[1, 3, 4]$. And those deaths previously classified in CoD [2] are now divided between new CoDs $[2, 4]$.

Example 2 Here we have a more general case: five CoDs in both periods, namely causes $[1, 2, 3, 4, 5]$ and $[1, 2, 3, 6, 7]$. The new ICD revision adds two new causes to the five causes, but it does not consider causes $[4, 5]$ anymore. Of course people will continue dieing from such causes, but they will be classified within one or more new causes. A possible diagram and the corresponding "correspondence table" can be written as follows:



$$\boldsymbol{T} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 6 & 7 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array} .$$

Also in this second example, CoDs in the first period has a correspondence in the second period (see the ones on the diagonal of the first 3 columns). Whereas new CoD [6] takes a proportions of death counts from old CoDs $[2, 4, 5]$ (see ones in the fourth column), new CoD [7] includes part of the old CoDs $[1, 3]$ (ones in the last column).

We observe a matrix of death counts: $\boldsymbol{D} = (d_{ij})$, where i and j index years and CoD, respectively. We divide the whole time-window in two periods of m_O and m_N years with associated n_O and n_N causes of deaths. The total number of CoDs denoted by n is given by the unique CoDs presented in both periods.

We arrange the matrix of deaths by column order into a vector $\boldsymbol{d}$. The actually observed death counts are assumed to be realizations from a Poisson distribution, $\boldsymbol{d} \sim \mathcal{P}(\boldsymbol{\mu})$. As mentioned, the expected values $\boldsymbol{\mu}$ are equal to a composed mean:

$$\boldsymbol{d} \sim \mathcal{P}(\boldsymbol{\mu} = \boldsymbol{C} \cdot \boldsymbol{\gamma}) \,, \qquad (1)$$

where $\boldsymbol{\gamma}$ is a matrix in column order: $\boldsymbol{\Gamma} = (\gamma_{ij})$ over years $i = 1, \ldots, m$ and new CoDs $j = 1, \ldots, n_N$. The matrix $\boldsymbol{C}$ which is called composition matrix has $(m_O\, n_O + m_N\, n_N)$ rows and $(m_O + m_N)\, n_N$ column, and it embodies the mechanism which occurs in ICD revision (see Section 3.1).

The vector $\boldsymbol{\gamma}$ is defined for new CoDs over both old and new classification periods. In mortality research we define such series as continuous series of death counts and we aim to estimate them. In case the old revision was identical to the new, i.e. $\boldsymbol{T}$ equal to an identity matrix, we could directly estimate $\boldsymbol{\gamma}$ from the data. This is not possible because $\boldsymbol{\gamma}$ is unknown in the old period due to changes in the classification, therefore it is crucial to simultaneously estimate both $\boldsymbol{\gamma}$ and the elements in matrix $\boldsymbol{C}$.

3.1 The composition matrix $\boldsymbol{C}$

The composition matrix $\boldsymbol{C}$ describes how the continuous series $\boldsymbol{\gamma}$ were mixed before generating the data, and it is characteristic for the CoD exchange pattern. In $\boldsymbol{C}$ one can see how theoretical death counts of $\boldsymbol{\gamma}$ are totally or partially redistributed among causes and according to the matrix

$\boldsymbol{T}$.

For modelling an ICD revision, we have thus to define the matrix $\boldsymbol{C}$ according to our assumptions in the "correspondence table". In this paper we will assume that all exchanges happens due to an ICD revision and that such exchanges hold during the complete old period.

Here it is important to note that the composition matrix is uniquely defined by $\boldsymbol{T}$ and it contains the "transition coefficients", here denoted as p_{ik}, proportion of deaths belonging to CoD i that get redistributed into CoD k.

For the sake of brevity we present only the compositional matrix for the mentioned example 1 on page 5. The matrices $\boldsymbol{C}$ associated to the second example and to the Belorussian data are given in the appendix and Section 6, respectively. Moreover, we will consider only 2 years in both the old and new ICD period, i.e. $i = [1, 2, 3, 4]$. Extension to longer period follows easily. Moreover for clarity we will add the unknown distributions γ_{ij} and the observed deaths d_{ij} as column and row names, respectively.

$$
\boldsymbol{C} = \begin{array}{c|cccc|cccc|cccc|cccc|}
 & \gamma_{11} & \gamma_{21} & \gamma_{31} & \gamma_{41} & \gamma_{12} & \gamma_{22} & \gamma_{32} & \gamma_{34} & \gamma_{13} & \gamma_{23} & \gamma_{33} & \gamma_{43} & \gamma_{14} & \gamma_{24} & \gamma_{34} & \gamma_{44} \\
d_{11} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{3,1} & 0 & 0 & 0 & p_{4,1} & 0 & 0 & 0 \\
d_{21} & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{3,1} & 0 & 0 & 0 & p_{4,1} & 0 & 0 \\
d_{31} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
d_{41} & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
d_{12} & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{4,2} & 0 & 0 & 0 \\
d_{22} & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{4,2} & 0 & 0 \\
d_{32} & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
d_{42} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
d_{33} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
d_{43} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
\hline
d_{34} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
d_{44} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array} .
$$

Note that each row of $\boldsymbol{C}$ (which corresponds to observed death counts) presents the contribution that each unknown latent γ_{ij} gives for the realization of the actual data. For instance, the second row reports that d_{21} (deaths in the second year and first CoD) is the realization of a Poisson distribution with mean equal to

$$
\begin{aligned}
E(d_{21}) = \mu_{2,1} &= (0 \cdot \gamma_{11}) + (1 \cdot \gamma_{21}) + (0 \cdot \gamma_{31}) + \ldots + (0 \cdot \gamma_{13}) + (p_{3,1} \cdot \gamma_{23}) + (0 \cdot \gamma_{33}) + \ldots \\
&= \gamma_{21} + (p_{3,1} \cdot \gamma_{23}) \, ,
\end{aligned}
$$

which means that the Poisson mean for d_{21} is equal to the latent distribution in the same year and CoD, γ_{21}, plus a partial contribution of the latent distribution in the same year and CoD [3], $p_{3,1} \cdot \gamma_{23}$.

Since we aim to redistribute only death counts in the old period, the compositional matrix

$\boldsymbol{C}$ has a single entry equal to 1 at the rows associated with deaths in the new period, i.e. data observed in the new period are realizations of a Poisson distribution without any composition.

It is worth mentioning that all columns in $\boldsymbol{C}$ must sum up to 1, i.e. all death counts must be redistributed. This reduces the number of "transition coefficients" to estimate. In the compositional matrix associated with the first example, we would have $p_{1,3} \equiv 1$ and $p_{4,2} \equiv 1 - p_{4,1}$. This means that we would need to estimate only $p_{4,1}$, the proportion of counts classified in CoD [1] that should move to [4].

Obviously more complex and larger examples can be accounted without changing the basic structure of the model in equation (1). However it is obvious that a large number of eventual exchanges among causes between two revisions leads to a large number of p_{jk}.

One can interpret such way of arranging changes due to ICD revision as a more complex and elegant way of looking at the problem than the traditional approach, but surely it cuts deeper. Whereas the well-established method for reconstructing mortality series by CoD can be considered as a deterministic procedure, the proposed model embeds the redistribution of counts among CoD in a stochastic process. This allows us to estimate simultaneously the optimal combination of the two model components: (1) "transition coefficients", $p_{i,k}$ and (2) the continuous series of death counts, $\boldsymbol{\gamma}$. Moreover such approach is purely based on data, though additional assumptions and constraints can be made. Another important advantage lays in the fact that uncertainty of the estimates could be assesses.

3.2 A Composite Link Model Approach

The model assumes data as realization of composed mean and it presents latent distributions (γ_{ij}). The composite link model (CLM, Thompson and Baker, 1981) is a suitable framework for such setting.

Within this frame and defining $\breve{x}_{ik} = \sum_j c_{ij}\gamma_j/\mu_i$, we propose an adjusted iterative reweighted least-squares (IRWLS) for estimating $\boldsymbol{\gamma}$:

$$(\breve{\boldsymbol{X}}'\tilde{\boldsymbol{W}}\breve{\boldsymbol{X}} + \lambda \boldsymbol{P})\tilde{\boldsymbol{\gamma}} = \breve{\boldsymbol{X}}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}}\,, \qquad (2)$$

where $\tilde{\boldsymbol{W}} = \mathtt{diag}(\tilde{\boldsymbol{\mu}})$, $\tilde{\boldsymbol{z}} = \tilde{\boldsymbol{W}}^{-1}(\boldsymbol{d} - \tilde{\boldsymbol{\mu}}) + \breve{\boldsymbol{X}} \ln \tilde{\boldsymbol{\gamma}}$, and $\boldsymbol{P}$ measures the roughness of the vector $\boldsymbol{\gamma}$ which combines in a vector the series of mortality by new CoDs over the both old and new period. The amount of smoothness will be weighted by a positive regularization parameter λ common for each series (Camarda et al., 2008; Eilers, 2007).

Once we identified the system for estimating the latent distributions, we need to write down an algorithm for the "transition coefficients". From the structure of $\boldsymbol{C}$ we can write:

$$\boldsymbol{\mu} = \boldsymbol{C}\boldsymbol{\gamma} = \breve{\boldsymbol{\gamma}} + \boldsymbol{\Psi}\boldsymbol{p}\,,$$

where $\breve{\boldsymbol{\gamma}}$ have the same elements of $\boldsymbol{\gamma}$ rearranged in order to match the elements of $\boldsymbol{\mu}$. Moreover $\boldsymbol{p}$ are all the "transition coefficients" concatenated into a vector and $\boldsymbol{\Psi}$ is the associated model matrix. Since $\boldsymbol{d} \sim \mathcal{P}(\boldsymbol{\mu})$, we therefore approximate $(\boldsymbol{d} - \breve{\boldsymbol{\gamma}})$ as

$$(\boldsymbol{d} - \breve{\boldsymbol{\gamma}}) \approx N(\boldsymbol{\Psi p}, \texttt{diag}(\boldsymbol{\mu}))\,.$$

Consequently, we can estimate the “transition coefficients” using a plain weighted least-squares:

$$(\boldsymbol{\Psi}'\tilde{\boldsymbol{V}}\boldsymbol{\Psi})\tilde{\boldsymbol{p}} = \boldsymbol{\Psi}'\tilde{\boldsymbol{V}}(\boldsymbol{d} - \breve{\boldsymbol{\gamma}})\,, \tag{3}$$

where $\tilde{\boldsymbol{V}} = \texttt{diag}(1/\tilde{\boldsymbol{\mu}})$.

An additional penalty within the weighted least-squares system will enforce positiveness in the estimated “transition coefficients” (Bollaerts et al., 2006).

Once λ is selected, the systems of equations described in (2) and (3) have a unique solution. A subjective choice of the smoothing parameter is used in the following and it could be used to test prior knowledge of the mortality trends. Routines for building each model components were implemented in `R` (R Development Core Team, 2011) and they are available from the author upon request.

4 Simulation and Applications

5 Simulation Study

To demonstrate the performance of our approach we applied it to a simulated scenario using the “correspondence table” presented for the second example.

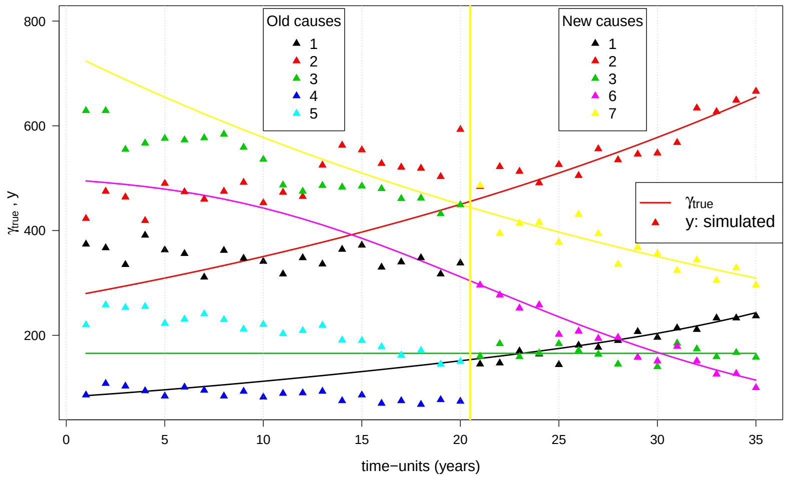


Figure 2: True latent series and simulated death counts by causes of death over time. Second example.

Figure 2 shows the simulated mortality series by CoD in two periods of 20 and 15 time-units (years). The solid lines represent the true latent distributions γ_{ij}, i.e. the expected values of what we would have observed if there were not change at year 20, and new CoDs $[1,2,3,6,7]$ were present during the whole period. The dots are the realizations of these latent series μ_{ij} after being redistributed via the matrix $\boldsymbol{C}$, namely our observed death counts d_{ij}. In the second period deaths closely follow the true latent series because no exchange have appeared in the last 15 years. In the first period the expected values are mixed before producing the deaths. This redistribution follows the associated "correspondence table" (page 5) and we use the following true "transition coefficients":

$$\begin{array}{lllll} p_{6,2} = 0.3 & \text{and} & p_{6,4} = 0.2 & \Rightarrow & p_{6,5} = 0.5 \\ & & p_{7,1} = 0.4 & \Rightarrow & p_{7,3} = 0.6\,. \end{array}$$

In other words, on average, deaths coming for the new CoD [6] are, in the old period, redistributed to old CoDs $[2,4,5]$ with "transition coefficients" 0.3, 0.2 and 0.5, respectively. Moreover, 40% of the deaths which theoretically would have belonged to CoD [7] goes to the old CoD [1]. Meanwhile the remaining 60% are included in old CoD [3].

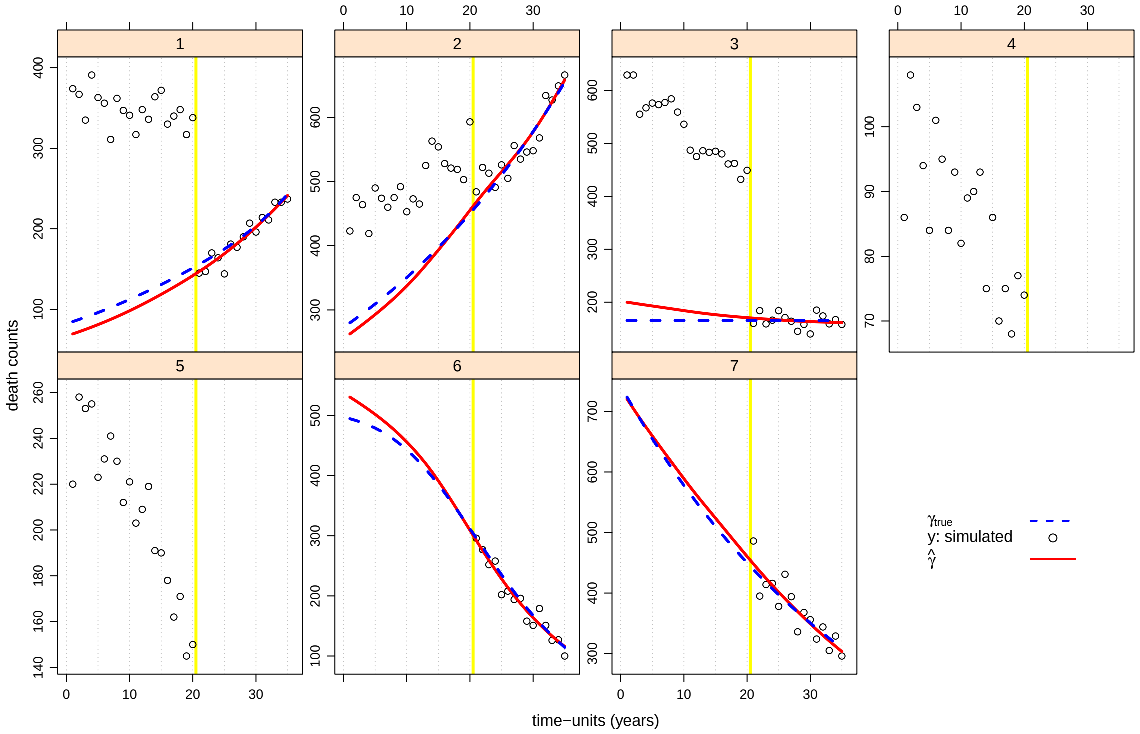


Figure 3: True latent series, simulated death counts and estimated continous series by causes of death over time. Second example.

Figure 3 shows seven panels with the true, simulated and fitted values from the proposed model. Each panel presents a single CoD over both old and new period. Obviously only CoDs $[1,2,3]$ show simulated deaths over the whole time-window. We simultaneously estimate the latent continuous mortality series $\hat{\boldsymbol{\gamma}}$ and the "transition coefficients" embedded in the compositional matrix $\boldsymbol{C}$. The selected λ is equal to 10^5.

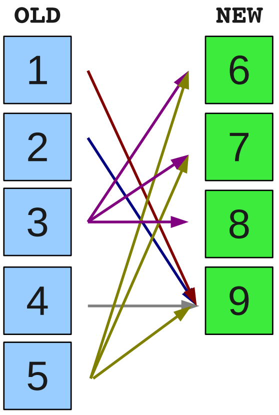
Estimated series are depicted in solid red lines and they closely follow the true expected values showing a rather good fit of the model. The three estimated $\hat{p}_{ij}$ are also similar to the true ones:

$$\begin{aligned} \hat{p}_{6,2} = 0.329 \quad \text{and} \quad \hat{p}_{6,4} &= 0.197 \quad \Rightarrow \quad \hat{p}_{6,5} = 0.474 \\ \hat{p}_{7,1} &= 0.419 \quad \Rightarrow \quad \hat{p}_{7,3} = 0.581 \,. \end{aligned}$$

It is worth pointing out how the suggested model does not assume any specific trend in the mortality series leading to a flexible model which uniquely relies on the data: observed death counts and known "correspondence table". Put differently, the presented approach searches for the best redistribution of counts in the old period, following all possible combination of exchanges and knowing that mortality in the new period is a smooth continuation of past mortality. As mentioned, prior knowledge on "transition coefficients" and mortality patterns could be incorporated into the model leading to more efficiency (less parameters to estimate), but at the price of flexibility.

6 Belorussian data

In order to apply the proposed model to the Belorussian mortality series by CoD introduced in Section 2, we need to set up the associated "correspondence table". Meslé et al. (1992) reconstructed mortality series by causes of deaths for the USSR and here we use their suggestions to create the associations among CoDs in Belarus, a former USSR republic. On one side all potential transfers among heart diseases are considered into the model, on the other side the model will automatically decide the proportions of death counts to redistribute, namely the "transition coefficients". The "correspondence table" $\boldsymbol{T}$ for Belarus can be written as follows:



$$\boldsymbol{T} = \begin{array}{c c} & \begin{array}{cccc} 6 & 7 & 8 & 9 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix} \end{array} .$$

Death counts coded with [3] and [5] in the old period are partially redistributed in new code [6] and [7]. Such associations are picked up by the ones in first and second column of $\boldsymbol{T}$. Meantime all deaths coded in the second period with [8] goes to the CoD [3] (see the single one in the third column). Finally death counts due to CoD [9] are redistributed among CoDs $[1, 2, 4, 5]$ (last column of $\boldsymbol{T}$).

Once the “correspondence table” is defined we can uniquely define the compositional matrix $\boldsymbol{C}$. For the sake of space, let reduce the Belorussian data supposing only two years in both periods: 1,2 and 3,4. The matrix $\boldsymbol{C}$ associated with $\boldsymbol{T}$ is then given by:

$$\boldsymbol{C} = \begin{array}{c|cccc|cccc|cccc|cccc|}
 & \gamma_{16} & \gamma_{26} & \gamma_{36} & \gamma_{46} & \gamma_{17} & \gamma_{27} & \gamma_{37} & \gamma_{37} & \gamma_{18} & \gamma_{28} & \gamma_{38} & \gamma_{48} & \gamma_{19} & \gamma_{29} & \gamma_{39} & \gamma_{49} \\
d_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{9,1} & 0 & 0 & 0 \\
d_{21} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{9,1} & 0 & 0 \\
\hline
d_{12} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{9,2} & 0 & 0 & 0 \\
d_{22} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{9,2} & 0 & 0 \\
\hline
d_{13} & p_{6,3} & 0 & 0 & 0 & p_{7,3} & 0 & 0 & 0 & p_{8,3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
d_{23} & 0 & p_{6,3} & 0 & 0 & 0 & p_{7,3} & 0 & 0 & 0 & p_{8,3} & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
d_{14} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{9,4} & 0 & 0 & 0 \\
d_{24} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{9,4} & 0 & 0 \\
\hline
d_{15} & p_{6,5} & 0 & 0 & 0 & p_{7,5} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{9,5} & 0 & 0 & 0 \\
d_{25} & 0 & p_{6,5} & 0 & 0 & 0 & p_{7,5} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{9,5} & 0 & 0 \\
\hline
d_{36} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
d_{46} & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
d_{37} & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
d_{47} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
d_{38} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
d_{48} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
\hline
d_{39} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
d_{49} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array} .$$

With this compositional matrix we summarize all possible CoD redistribution in the old period. Since the sum of the element of $\boldsymbol{C}$ column-wise must be equal to 1, the number of “transition coefficients” to estimate reduce to five: $p_{6,3}$, $p_{7,3}$, $p_{9,1}$, $p_{9,2}$ and $p_{9,4}$. The remaining coefficients in $\boldsymbol{C}$ can be computed as follows:

$$\begin{aligned}
p_{6,5} &\equiv 1 - p_{6,3} \\
p_{7,5} &\equiv 1 - p_{7,3} \\
p_{8,3} &\equiv 1 \\
p_{9,5} &\equiv 1 - p_{9,1} - p_{9,2} - p_{9,4}
\end{aligned}$$

As example we explicitly write down the expected value for the deaths observed in the second year and CoD [3]:

$$\begin{aligned}
E(d_{23}) = \mu_{2,3} &= (p_{6,3} \cdot \gamma_{26}) + (p_{7,3} \cdot \gamma_{27}) + (p_{8,3} \cdot \gamma_{28}) \\
&= (p_{6,3} \cdot \gamma_{26}) + (p_{7,3} \cdot \gamma_{27}) + \gamma_{28} \,.
\end{aligned}$$

Another advantage of using the proposed approach lies in the fact that the expected values for all years and CoD could be easily computed as a single matrix multiplication, i.e. $E(\boldsymbol{d}) = \boldsymbol{\mu} = \boldsymbol{C} \cdot \boldsymbol{\gamma}$.

Figure 4 shows the outcomes of our model applied to the Belorussian data. Each panel represents a specific CoD and fitted values are obviously given for the last 4 new CoDs. The model is able to capture mortality pattern in the second period following the death counts from 1970 onwards. Meantime in the first period we redistribute all counts due to old $[1, 2, 3, 4, 5]$ into the new CoDs keeping the same redistribution over the whole old period and assuming a smooth trend in mortality.

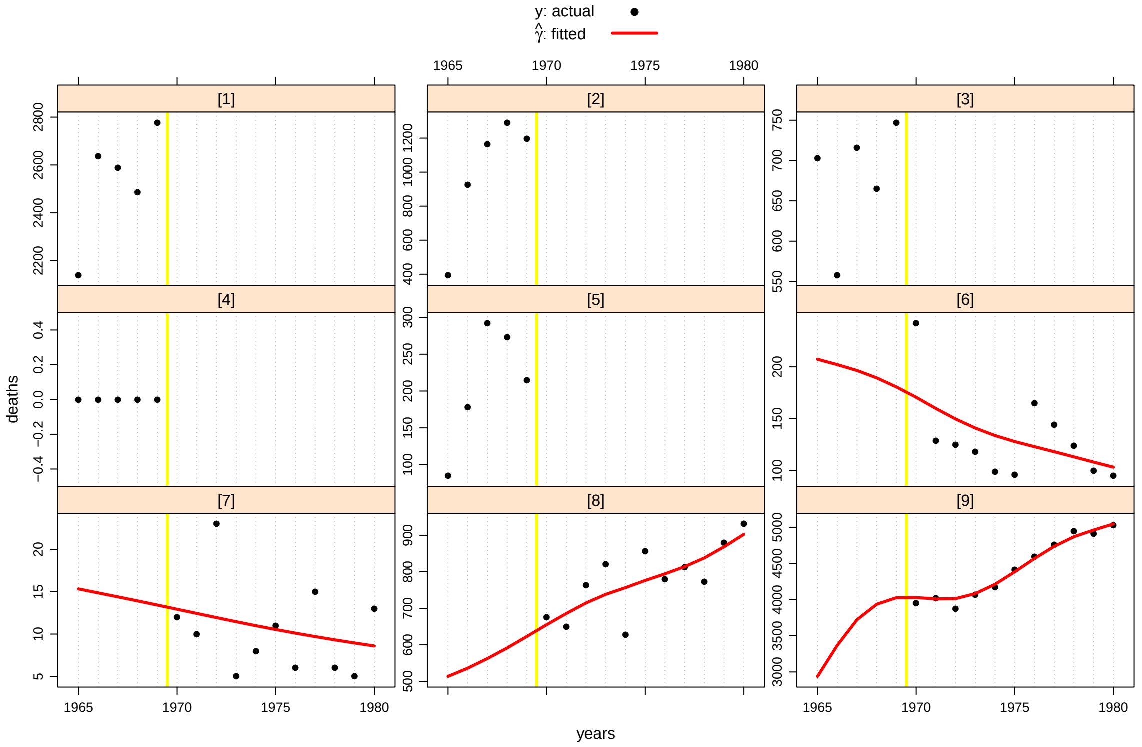


Figure 4: True latent series, simulated death counts and estimated continous series by causes of death over time. Second example.

The seemingly different goodness of fit in each CoD can be attributed to the large differences in number of deaths in each CoD. For instance, whereas the estimated continuous series follows closely the observed deaths for CoD [9], the fitted values shows only the main downward trend for [7]. Such behavior is correct when deaths are assumed as realization from a Poisson process: the model tends to reduce errors in years with large sample size.

We arrange the fitted "transition coefficients" like in the "correspondence table", but additionally we compute the relative frequencies as proportions of the row totals:

$$\hat{p}_{ij} = \begin{array}{cc} & \begin{matrix} 6 & 7 & 8 & 9 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{pmatrix} 0 & 0 & 0 & 0.696 \\ 0 & 0 & 0 & 0.275 \\ 0.605 & 0.008 & 1 & 0 \\ 0 & 0 & 0 & 10^{-9} \\ 0.395 & 0.992 & 0 & 0.029 \end{pmatrix} \end{array} \Rightarrow \quad \mathrm{rf}_r(\hat{p}_{ij}) = \begin{array}{cc} & \begin{matrix} 6 & 7 & 8 & 9 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{pmatrix} 0 & 0 & 0 & 1.000 \\ 0 & 0 & 0 & 1.000 \\ 0.375 & 0.005 & 0.620 & 0 \\ 0 & 0 & 0 & 1.000 \\ 0.279 & 0.701 & 0 & 0.021 \end{pmatrix} \end{array} .$$

From the matrix of $\mathrm{rf}_r(\hat{p}_{ij})$ we can identify how deaths counts have been redistributed according to the model. For instance, the total amount of deaths belonging to $[1, 2, 4]$ have been classified in new CoD [9]. Namely, deaths due to cerebrovascular disorders with hypertension, [1], cerebrovascular disorders with hypertension and cerebral arteriosclerosis, [2], and cerebral hypertensive disease except disorders of the central nervous system, [4], are all classified within the new "cerebrovascular disorders with hypertensive disease" group. Moreover in this last CoD 2.1% of deaths due to heart and cerebral hypertensive diseases [5] are also included.

Furthermore 37.5%, 0.5% and 62% of deaths that in the old period are due to CoD [3] are redistributed among CoD [6], [7] and [8], respectively. Put differently, deaths due to hypertensive heart diseases in the old period are now distributed among hypertensive heart diseases, hypertensive heart and renal disease and atherosclerotic cardiosclerosis with hypertensive disease, with the mentioned percentages.

7 Concluding remarks

Cause-specific mortality data often present breaks due to revision of the International Classification of Diseases (ICD). The lack of data continuity has led to an underrate of this important source of information for understanding mortality developments. In this paper, we present a novel approach for reconstructing continuous series of mortality by cause of death.

The established procedure manually redistributes counts occurred during a first period among causes present in the new period. This approach employs visual inspection and subjective assessment for correcting eventual irregular trends. Alternatively, the proposed model is uniquely based on data and knowledge on the potential associations among causes of deaths.

An understanding of exchanges among old and new causes of deaths is used in constructing the so-called "correspondence table": a logical matrix in which possible associations between two causes of death are depicted by one. Once such matrix is obtained, the model will search for the best redistribution of counts among possible associations, assuming a smooth trend in mortality over time.

Deaths are assumed as realizations from a Poisson distribution. The unknown continuous series are considered smooth latent distributions, which are mixed during the old period. In other words, the expected values of the Poisson distribution compose the continuous series using a compositional matrix which embodies the associations from the "correspondence table".

Using such framework, the composite link model becomes a suitable approach for estimating both cause-specific mortality trends and, simultaneously, the exchange pattern between old and new causes of death. The paper briefly present the model and its estimation algorithm with emphasis on the specific problem in hand.

As was demonstrated by the simulation study, the model allows the description of the true underlying mortality series and a good estimation of the exchanges among causes of death. Moreover the application on the actual Belorussian data produces reasonable outcomes in both cause-specific mortality trends and proportion of death counts redistributed among causes.

The assumption of smooth latent distributions can be modified rather easily if more specific hypotheses are reasonable. For instance, as commonly done by demographers, we forced the sum of fitted and actual death counts to be equal each year. In this way irregularities in the mortality developments over time are picked by the latent distribution too.

In many datasets on mortality by cause of death, it is unclear the structure of the "correspondence table", i.e. we cannot be sure that certain cause exchange with others. In this cases, a generalization of the suggested model which, from all possible exchanges, is able to both estimate and select the most important "transition coefficients" is enviable. Regularization techniques such as ridge and "lasso" regression (Hoerl and Kennard, 1988; Tibshirani, 1996) are already available and will be explored within our setting.

The model assumes both sexes and all ages combined and only two classification periods. We currently are developing an extension of the presented model to include more ICD revisions. Likely a multiple-steps approach could be used to back project cause-specific mortality throughout several revisions with different "correspondence tables".

On one hand it is reasonable to assume that mortality dynamics develop in a smooth fashion over both period and ages. On the other hand, causes of death and possible associations could be different between ages. We plan to generalize the suggested model over the age domain specifying a compositional matrix in three dimensions: observed counts, latent distributions and ages. In this new version "transition coefficients" can (smoothly) change over the age range, following an eventual age-specific "correspondence table".

Acknowledgements

The author wishes to express his thanks to Pavel Grigoriev who provided the Belorussian data. Helpful discussion with France Meslé and Jacques Vallin are gratefully acknowledged.

References

Bollaerts, K., P. H. C. Eilers, and I. van Mechelen (2006). Simple and multiple P-splines regression with shape constraints. *British Journal of Mathematical and Statistical Psychology 59*, 451–469.

Camarda, C. G., P. H. C. Eilers, and J. Gampe (2008). Modelling General Patterns of Digit Preference. *Statistical Modelling 8*, 385–401.

Eilers, P. H. C. (2007). Ill-posed Problems with Counts, the Composite Link Model and Penalized Likelihood. *Statistical Modelling 7*, 239–254.

Hoerl, A. and R. Kennard (1988). Ridge Regression. In *Encyclopedia of Statistical Sciences*, Volume 8, pp. 129–136. New York: Wiley.

Janssen, F. and A. E. Kunst (2004). ICD coding changes and discontinuities in trends in cause-specific mortality in six European countries, 1950-99. *Bulletin of the World Health Organization 82*, 904–913.

Keiding, N. (1990). Statistical Inference in the Lexis Diagram. *Philosophical Transactions: Physical Sciences and Engineering 332*, 487–509.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Model* (2nd ed.). Monographs on Statistics Applied Probability. London: Chapman & Hall.

Meslé, F., V. M. Shkolnikov, and J. Vallin (1992). Mortality by cause in the USSR in 1970-1987: the reconstruction of time series. *European Journal of Population 8*, 281–308.

Meslé, F. and J. Vallin (1996). Reconstructing long-term series of causes of death: The case of France. *Historical Methods 29*, 72–87.

Pechholdová, M. (2009). Results and observations from the reconstruction of continuous time series of mortality by cause of death: Case of West Germany, 1968-1997. *Demographic Research 21*, 535–568.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Thompson, R. and R. J. Baker (1981). Composite Link Functions in Generalized Linear Models. *Applied Statistics 30*, 125–131.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society 58*, 267–288.

Appendix: the compositional matrix for the second example

In this appendix, we present the composite matrix $\boldsymbol{C}$ from the second example. The "correspondence table" $\boldsymbol{T}$ is given on page 5, and we assume only 2 years in both the old and new period, i.e. $i = [1, 2, 3, 4]$.

$$
\boldsymbol{C} =
\begin{array}{c|cccc|cccc|cccc|cccc|cccc}
 & \gamma_{11} & \gamma_{21} & \gamma_{31} & \gamma_{41} & \gamma_{12} & \gamma_{22} & \gamma_{32} & \gamma_{34} & \gamma_{13} & \gamma_{23} & \gamma_{33} & \gamma_{43} & \gamma_{16} & \gamma_{26} & \gamma_{36} & \gamma_{46} & \gamma_{17} & \gamma_{27} & \gamma_{37} & \gamma_{47} \\
\hline
d_{11} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{7,1} & 0 & 0 & 0 \\
d_{21} & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{7,1} & 0 & 0 \\
d_{31} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
d_{41} & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
d_{12} & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{6,2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
d_{22} & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{6,2} & 0 & 0 & 0 & 0 & 0 & 0 \\
d_{32} & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
d_{42} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
d_{13} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{7,3} & 0 & 0 & 0 \\
d_{23} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{7,3} & 0 & 0 \\
d_{33} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
d_{43} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
d_{14} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{6,4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
d_{24} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{6,4} & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
d_{15} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{6,5} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
d_{25} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{6,5} & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
d_{36} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
d_{46} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
\hline
d_{37} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
d_{47} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}
$$