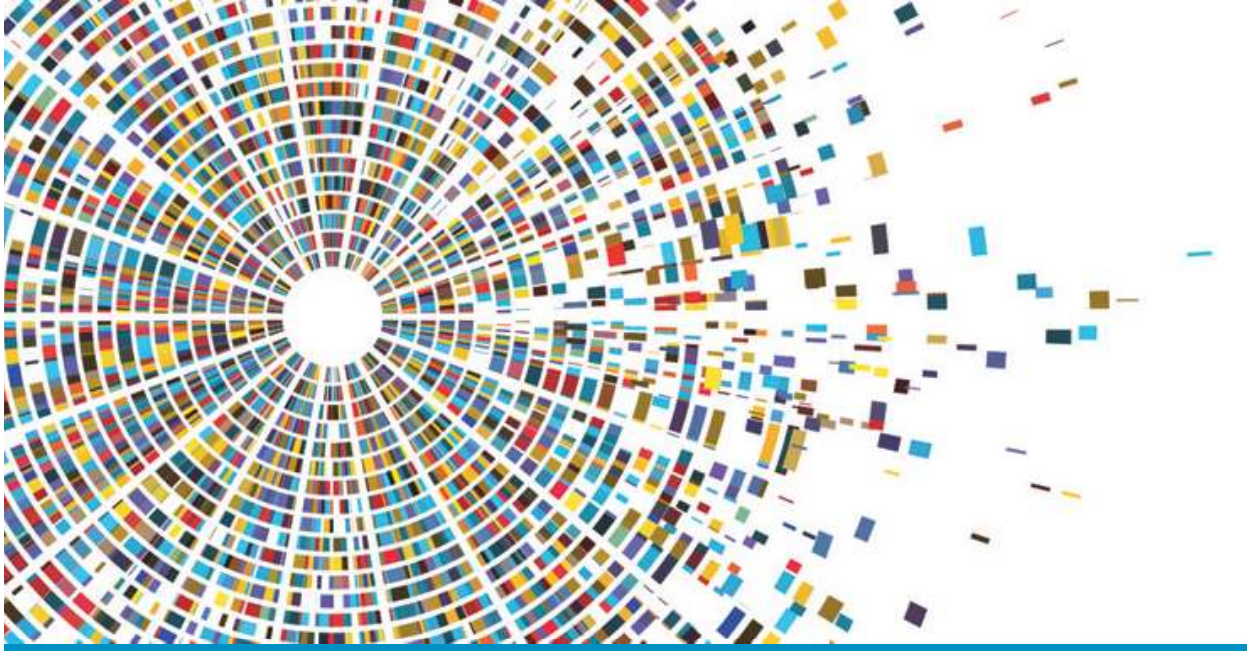

CLADAG 2023

BOOK OF ABSTRACTS AND SHORT PAPERS

14th Scientific Meeting of the Classification and Data Analysis Group
Salerno, September 11-13, 2023

edited by

Pietro Coretto
Giuseppe Giordano
Michele La Rocca
Maria Lucia Parrella
Carla Rampichini

Pearson

SCIENTIFIC PROGRAM COMMITTEE

Carla Rampichini (chair, University of Florence - Italy)

Claudio Agostinelli (University of Trento - Italy)
Michela Battauz (University of Udine - Italy)
Antonio Canale (University of Padua - Italy)
Carlo Cavicchia (Erasmus University Rotterdam - Netherlands)
Claudio Conversano (University of Cagliari - Italy)
Eustasio del Barrio (University of Valladolid - Spain)
Roberto Di Mari (University of Catania - Italy)
Stefania Fensore (University of "G. d'Annunzio" - Italy)
Nial Friel (University College Dublin - Ireland)
Maria Giovanna Ranalli (University of Perugia - Italy)
Leonardo Grilli (University of Firenze - Italy)
Luigi Grossi (University of Padua - Italy)
Christian Hennig (University of Bologna - Italy)
Mia Hubert (KU Leuven - Belgium)
Alfonso Iodice D'Enza (University of Naples "Federico II" - Italy)
Julien Jacques (University of Lyon - France)
José Joaquim Dias Curto (ISCTE-Instituto Universitário de Lisboa- Portugal)
Michele La Rocca (University of Salerno - Italy)
Silvia Montagna (University of Turin - Italy)
Barbara Pawelek (University of Cracow - Poland)
Fulvia Pennoni (University of Milano-Bicocca - Italy)
Mario Rosario Guarracino (University of Cassino - Italy)
Katrijn Van Deun (University of Tilburg - Netherlands)
Simone Vantini (Politecnico di Milano - Italy)
Donatella Vicari (Sapienza University of Rome - Italy)
Helga Wagner (Johannes Kepler University Linz - Austria)
Hiroshi Yadohisa (Doshisha University - Japan)

LOCAL PROGRAM COMMITTEE

Michele La Rocca (chair, University of Salerno - Italy)

Pietro Coretto (University of Salerno - Italy)
Giuseppe Giordano (University of Salerno - Italy)
Paolo Rocca Comite Mascambruno (University of Salerno - Italy)
Marcella Niglio (University of Salerno - Italy)
Maria Lucia Parrella (University of Salerno - Italy)
Marialuisa Restaino (University of Salerno - Italy)
Domenico Vistocco (University of Naples "Federico II" - Italy)
Maria Prosperina Vitale (University of Salerno - Italy)

CLADAG 2023 BOOK OF ABSTRACTS AND SHORT PAPERS:
14th Scientific Meeting of the Classification and Data Analysis Group, Salerno, September 11-13, 2023
edited by Carla Rampichini, Michele La Rocca, Pietro Coretto, Giuseppe Giordano, Maria Lucia Parrella

Front cover: Genome sequence map, chromosome architecture and genetic sequencing chart abstract data,
© Tartila / Shutterstock

Published by Pearson Education Resources, Italia
www.pearson.it
ISBN: 9788891935632

A COHORT STUDY ON THE GENDER GAP IN MORTALITY THROUGH THE TUCKER3 MODEL

Paolo Giordani [1], Susanna Levantesi [1], Andrea Nigri [2] and Virginia Zarulli [3]

[1] Department of Statistical Sciences, Sapienza University of Rome, (e-mail: paolo.giordani@uniroma1.it, susanna.levantesi@uniroma1.it)

[2] Department of Economics, Management and Territory, University of Foggia, (e-mail: andrea.nigri@unifg.it)

[3] Interdisciplinary Centre on Population Dynamics, University of Southern Denmark, (e-mail: vzarulli@sdu.dk)

ABSTRACT: In this manuscript, leveraging the Tucker3 model, we investigate the gender gap in mortality considering the ratio of male to female mortality rates, specific for age, cause of death, and cohort. The model is applied to a tensor containing gender gap data by causes, age classes, and non-extinct cohorts.

KEYWORDS: Mortality data, gender gap, cohort study, multi-way data, Tucker3

1 Introduction

Understanding mortality is of great importance for both private and public sectors to design appropriate pension or insurance plans. To this purpose, several interesting applications of multi-way models to mortality data are available in the literature (see, e.g., Cardillo *et al.*, 2023). Generally speaking, in these studies, data usually refer to mortality rates across demographic features such as causes of death, ages, countries, and years. This work represents a further step in mortality analysis by focusing on the gender gap (Zarulli *et al.*, 2021) in causes of death and its evolution by cohort. Limiting our attention to the three-way case, the Tucker3 model is applied to a tensor containing gender gap data in mortality distinguished by causes of death, age classes, and cohorts.

2 Three-way data and models

A three-way array or tensor $\underline{\mathbf{X}}$ of order $(I \times J \times K)$ can be seen as a box containing scores on a set of I observation units with respect to J variables in K different occasions. Observation units, variables and occasions are usually referred to as "modes". The generic element of $\underline{\mathbf{X}}$ is x_{ijk} giving the score

of observation unit i ($i = 1,\ldots,I$) on variable j ($j = 1,\ldots,J$) at occasion k ($k = 1,\ldots,K$). Thus, there are three ways or indices, one for each mode. The array $\underline{\mathbf{X}}$ can be seen as a collection of standard matrices of order $(I \times J)$, one for every occasion.

It is often convenient to summarize $\underline{\mathbf{X}}$ to unravel the relevant information hidden in the data. To this purpose, suitable extensions of Principal Component Analysis for arrays should be considered. One of the most famous models is the Tucker3 one (Tucker, 1966). The Tucker3 model synthesizes $\underline{\mathbf{X}}$ by extracting P ($< I$), Q ($< J$) and R ($< K$) components for the observation units, variables and occasions, respectively, thus allowing different levels of complexity for the three modes. Let $\mathbf{X}_a$ be the matrix of order $(I \times JK)$ obtained by juxtaposing next to each other the standard matrices pertaining to every occasion. The Tucker3 model can be formalized as

$$\mathbf{X}_a = \mathbf{A}\mathbf{G}_a\left(\mathbf{C} \otimes \mathbf{B}\right)^{\mathrm{T}} + \mathbf{E}_a, \tag{1}$$

where $\mathbf{A}$ of order $(I \times P)$, $\mathbf{B}$ of order $(J \times Q)$ and $\mathbf{C}$ of order $(K \times R)$ are the component score matrices for the observation units, the variables and the occasions, respectively. Therefore, each mode is summarized by the corresponding set of components. The triple interactions among such components are measured by the three-way array $\underline{\mathbf{G}}$ of order $(P \times Q \times R)$ called *core*. Finally, $\mathbf{E}_a$ is the error matrix of order $(I \times JK)$ and the symbol $\otimes$ denotes the Kronecker product. Estimation of the model parameters is carried out in the least square sense by

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C},\underline{\mathbf{G}}} ||\mathbf{E}_a||^2, \tag{2}$$

being $||\cdot||$ the Frobenius norm of matrices. An alternating least squares algorithm can be used. It can be shown that the obtained solution is not identifiable. In fact, all component matrices as well as the core array can be rotated. The non-identifiability can be exploited in order to rotate the solution to a simple structure. Given P, Q and R, we can assess the fit percentage of the Tucker3 model as

$$\left(1 - \frac{||\mathbf{E}_a||^2}{||\mathbf{X}_a||^2}\right) 100. \tag{3}$$

The closer to 100, the better the fit of the Tucker3 model. The optimal numbers of components P, Q and R can be found by balancing fit and parsimony, bearing in mind that interpretability is of relevant importance. For further details on the Tucker3 model and related multi-way models, the interested reader may refer to (Kroonenberg, 2008).

3 Results

The analyzed data come from the Human Cause-of-Death Database (HCD) and refer to the mortality rates distinguished by causes of death, age classes and cohorts registered in the United States of America. In particular, we consider the mortality rates of $I = 7$ causes of death (Infectious diseases, Neoplasms, Cardiovascular diseases, Respiratory diseases, Digestive diseases, External causes of death, Other causes of death) distinguished in $J = 7$ five-year age classes from 60 to 90 years for cohorts of people born in $K = 10$ years from 1919 to 1928. In order to deal with fully crossed data, i.e. all observation units have scores on all variables on all occasions, such mortality rates are collected for the years 1979–2018. Letting m^F_{ijk} and m^M_{ijk} be the mortality rates of the cause of death i at age class j for cohort k for females and males, respectively, the generic element of the three-way gender gap data array $\underline{\mathbf{X}}$ is

$$x_{ijk} = \frac{m^M_{ijk}}{m^F_{ijk}}, \tag{4}$$

expressing to what extent the mortality rate for a certain cause of death of a given age and belonging to a specific cohort of males differs from the corresponding rate for females.

To assess whether and how gender differences in mortality are related to causes of death, ages and cohorts, the Tucker3 model with $P = Q = 2$ and $R = 1$ components is used. To motivate this choice, we observe that the fit percentage is rather high (91.60%), despite the low total number of components ($P + Q + R = 5$), and the solution is well interpretable. In this respect, simplicity is achieved by transforming $\mathbf{G}_a$ to the identity matrix and applying the varimax (Kaiser, 1958) rotation to $\mathbf{B}$ compensating it in $\mathbf{A}$. In this way, the components for the causes of death and those for the age classes are related one-to-one.

The component matrix $\mathbf{A}$ for the causes of death is displayed in Figure 1. The component matrix $\mathbf{B}$ for the age classes (not reported here) distinguishes the younger ages (from 60 to 70) with large positive first component scores and the older ages (from 75 to 90) with large positive second component scores. Taking into account that the component scores for the cohorts are all positive and decreasing passing from cohort 1919 to cohort 1928, the main findings are that the gender gap for ages 60–70 increases in connection with Cardiovascular diseases and External causes and decreases with Infectious diseases and Other causes. This especially holds for the oldest cohorts. Conversely, for ages 75–90, the gender gap for Neoplasms and Respiratory diseases is high, whilst

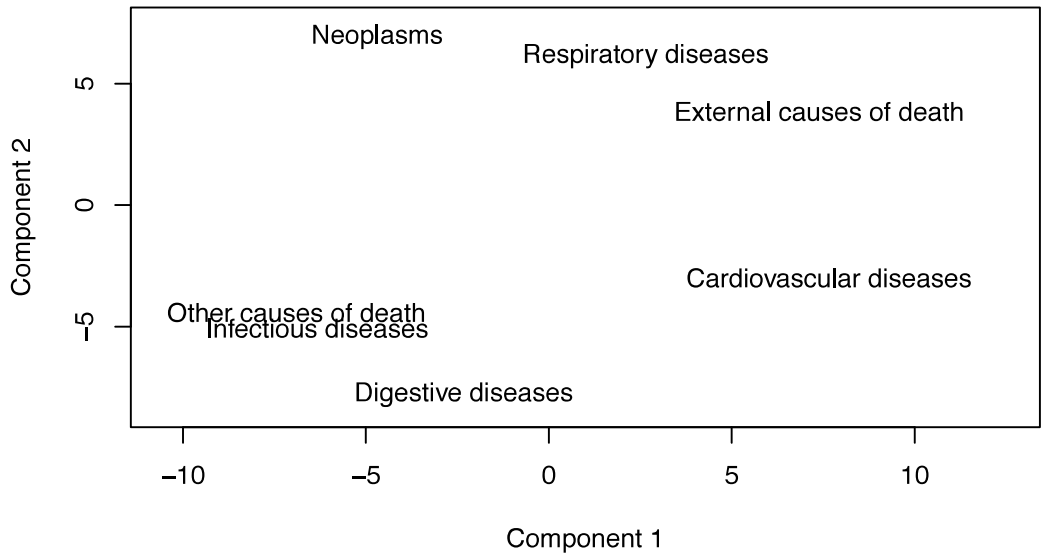

Figure 1. *Component scores for the causes of death.*

the opposite comment holds for Digestive diseases. Further results will be presented during the conference.

References

CARDILLO, G., GIORDANI, P., LEVANTESI, S., NIGRI, A., & SPELTA, A. 2023. Mortality forecasting using the four-way CANDECOMP/PARAFAC decomposition. *Scandinavian Actuarial Journal*, in press.

KAISER, H. F. 1958. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, **23**, 187–200.

KROONENBERG, P. M. 2008. *Applied Multiway Data Analysis*. Hoboken: Wiley.

TUCKER, L. R. 1966. Some mathematical notes on three-mode factor analysis. *Psychometrika*, **31**, 279–311.

ZARULLI, V., KASHNITSKY, I., & VAUPEL, J.W. 2021. Death rates at specific life stages mold the sex gap in life expectancy. *Proc. Natl. Acad. Sci.*