CONTENTS

Metodološki zvezki, Vol. 11, No. 1, 2014

Metodološki zvezki, Vol. 11, No. 1, 2014, 1-20

Symbolic Covariance Matrix for Interval-valued Variables and its Application to Principal Component Analysis: a Case Study

Katarina Košmelj[1], Jennifer Le-Rademacher[2] and Lynne Billard[3]

Abstract

In the last two decades, principal component analysis (PCA) was extended to interval-valued data; several adaptations of the classical approach are known from the literature. Our approach is based on the symbolic covariance matrix Cov for the interval-valued variables proposed by Billard (2008). Its crucial advantage, when compared to other approaches, is that it fully utilizes all the information in the data. The symbolic covariance matrix can be decomposed into a within part $CovW$ and a between part $CovB$. We propose a further insight into the PCA results: the proportion of variance explained due to the within information and the proportion of variance explained due to the between information can be calculated. Additionally, we suggest PCA on $CovB$ and $CovW$ to be done to obtain deeper insights into the data under study.

In the case study presented, the information gain when performing PCA on the intervals instead of the interval midpoints (conditionally the means) is about 45%. It turns out that, for these data, the uniformity assumption over intervals does not hold and so analysis of the data represented by histogram-valued variables is suggested.

1 Introduction

1.1 Principal component analysis for classical data

Principal component analysis (PCA) was first described by Pearson (1901) as an analogue of the principal axes theorem in mechanics; it was later independently developed and named by Harold Hotelling in the 1930s. It is a very popular exploratory tool in classical multivariate data (see e.g., Chatfield and Collins, 1980; Johnson and Wichern, 2002). Its major objective is to reduce the dimension of the variable space: the original p random variables $\mathbf{X} = (X_1, X_2, ..., X_p)$ are transformed into s random variables $\mathbf{Y} = (Y_1, Y_2, ..., Y_s)$, called Principal Components, where $s \ll p$, and the $\mathbf{Y}$ variables are uncorrelated. This transformation is defined in such a way that the first principal component (PC_1) accounts for as much of the variability, i.e., variance, in the data as

[1] Biotechnical Faculty, University of Ljubljana, Slovenia; katarina.kosmelj@bf.uni-lj.si

[2] Medical College of Wisconsin, Milwaukee, USA; jlerade@mcw.edu

[3] University of Georgia, Athens, USA; lynne@stat.uga.edu

possible, and each succeeding component in turn has the highest variance possible, under the constraint that it be orthogonal to (i.e., uncorrelated with) the preceding components.

The solution of the problem described above is given by the eigenvalues and eigenvectors of the covariance matrix of $X_1, X_2, ..., X_p$. Principal components are linear combinations of the original variables, defined by the eigenvectors of this covariance matrix. From the basic linear algebra it follows: there are p eigenvalues ordered: $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$; eigenvalues of the covariance matrix are the variances of the principal components. The eigenvalues add up to the sum of the diagonal elements, i.e., to the trace of the covariance matrix. This means that the sum of the variances of the principal components is equal to the sum of the variances of the original variables. The i-th principal component accounts for $\lambda_i / \sum_{j=1}^{p} \lambda_j$ of the total variance in the original data. When the decision on the reduced dimension s is taken, we calculate the proportion of variance accounted for by the first s principal components, $\sum_{j=1}^{s} \lambda_j / \sum_{j=1}^{p} \lambda_j$.

As the covariance on standardized variables equals the correlation, therefore, in this case, eigenvalues and eigenvectors of the correlation matrix are used. It is recommended to perform PCA on standardized variables when the original variables are measured on scales with different ranges.

1.2 Principal component analysis for symbolic data

In the second part of the 20^{th} century, the need to analyze massive datasets emerged. Symbolic data analysis started as a response to that demand; see Bock and Diday (2000), Billard and Diday (2003, 2006), among others. Symbolic analytical methods are often generalizations of their classical approach counterparts. A symbolic method should give the same results as its classical counterpart when applied to classical data (Billard, 2011, Le-Rademacher and Billard, 2012).

In the last two decades, PCA was adapted for symbolic data, first in the context of interval-valued data. A number of approaches were proposed. Le-Rademacher and Billard (2012) give a short overview of its historical development; let us review them briefly. Cazes et al. (1997) proposed the first adaptations of PCA known as the *centers method* and the *vertices method*, see also Douzal-Chouakria et al. (2011); Zuccolotto (2007) applied the vertices method to a dataset on job satisfaction; Lauro and Palumbo (2000) introduced a Boolean matrix to account for the interdependency of the vertices, Palumbo and Lauro (2003) and Lauro et al. (2008) proposed the *midpoint-radii* method treating interval midpoints and interval midranges as two separate variables; Gioia and Lauro (2006) proposed a PCA version based on an interval algebra approach.

Le-Rademacher and Billard (2012) describe these approaches in detail and discuss their characteristics in the context of symbolic data analysis: namely, these approaches fail in different ways to utilize the entire information included in the interval-valued data.

These deficiencies can be avoided when the symbolic covariance matrix Cov is used. Its calculation in the interval setting was first presented in Billard (2008). The crucial advantage of this symbolic covariance matrix is that it fully utilizes all the information in the data; also it is shown that the symbolic covariance matrix can be decomposed into a within part $CovW$ and a between part $CovB$. Two papers on this topic (Le-Rademacher and Billard, 2012 and Billard and Le-Rademacher, 2012) also provide a new approach to constructing the observations in PC space allowing for a better visualization of the results. Le-Rademacher and Billard (2013a) propose an approach to construct histogram values from the principal components of interval-valued observations. Le-Rademacher (2008) and and Le-Rademacher and Billard (2013b) extend these ideas to histogram-valued observations. In a different direction, Giordani and Kiers (2006) consider fuzzy data, which is a different domain from symbolic data and so is outside the purview of the present work. Likewise, a different domain is the PCA of time series data of Irpino (2006).

1.3 Objective of this study

We want to compare PCA results obtained on different data types. To enable the comparison of the results, the data were aggregated from the same dataset. For each observation and each variable, we aggregated the data in two different ways:

- the mean value;
- the $[min, max]$ interval which is based on the minimal and maximal value under observation.

The main objective of this study is to find out what is the information gain when analyzing the $[min, max]$ interval instead of the mean value.

In the next section, some well known characteristics of interval-valued data are summarized. Covariance in the interval setting will be illustrated and compared to the covariance in the classical setting. For PCA on interval-valued variables, a simple measure of the information gain will be introduced and additional PCA analyses will be suggested. These approaches allow for a deeper insight into the dataset under study.

The third section presents a simple case study. It consists of seven meteorological stations in Slovenia, they are described by seven variables, the data are from the 40 year-period 1971-2010. The results of different PCA analyses will be presented and compared. To facilitate the comparison of the results, the dataset is very small, however, the stations are chosen according to subject-matter knowledge. The last section gives some conclusions and suggestions for further work.

2 Interval-valued variables

Let us first note that an interval-valued random variable is just a standard random variable but its values are intervals. Let $\mathbf{X} = (X_1, X_2, ..., X_p)$ be a p-dimensional random variable taking values in R^p. Let X_j be an interval-valued random variable, its data exist for a random sample of size n and is in the form $X_{ij} = [a_{ij}, b_{ij}]$, $a_{ij} \leq b_{ij}$, $i = 1, ..., n$. In the case $a_{ij} = b_{ij}$, for any $i = 1, 2, ..., n$, X_{ij} has a classical value. Each observation described by a p-dimensional interval-valued variable can be visualized as a hypercube in R^p.

2.1 Mean and variance

The mean and the variance for an interval-valued variable are based on the assumption that the distribution of the values within each interval is uniform. They were first defined by Bertrand and Goupil (2000). The sample variance of X_j is:

$$S_j^2 = \frac{1}{3n}\sum_{i=1}^{n}(a_{ij}^2 + a_{ij}b_{ij} + b_{ij}^2) - \overline{X}_j^2, \tag{2.1}$$

where the sample mean $\overline{X}_j$ is the average of the interval midpoints

$$\overline{X}_j = \frac{1}{2n}\sum_{i=1}^{n}(a_{ij} + b_{ij}). \tag{2.2}$$

Billard (2008) showed that (2.1) can be rewritten as

$$S_j^2 = \frac{1}{3n}\sum_{i=1}^{n}[(a_{ij} - \overline{X}_j)^2 + (a_{ij} - \overline{X}_j)(b_{ij} - \overline{X}_j) + (b_{ij} - \overline{X}_j)^2], \tag{2.3}$$

and proved that the Total Sum of Squares SST can be decomposed into a within part SSW and a between part SSB :

$$nS_j^2 \ = SST_j \ = SSW_j \ + SSB_j. \tag{2.4}$$

The Within Sum of Squares SSW measures the internal variation and can be expressed as follows:

$$\begin{aligned} SSW_j = &\ \frac{1}{3}\sum_{i=1}^{n}[(a_{ij} - \frac{a_{ij}+b_{ij}}{2})^2 + (a_{ij} - \frac{a_{ij}+b_{ij}}{2})(b_{ij} - \frac{a_{ij}+b_{ij}}{2}) + (b_{ij} - \frac{a_{ij}+b_{ij}}{2})^2] \\ = &\ \sum_{i=1}^{n}\frac{(b_{ij}-a_{ij})^2}{12}. \end{aligned} \tag{2.5}$$

Thus, as expected, SSW is based on an implicit assumption that the distribution of values within each observed interval is uniform, X_{ij} ˜ $U(a_{ij}, b_{ij})$, $i = 1, 2, ..., n$. Other distributions are also relevant; e.g., Billard (2008) presents the formulae for SSW and SST when observations within each interval follow a triangular distribution.

The Between Sum of Squares SSB describes the between variation, i.e., the variation of the interval midpoints:

$$SSB_j = \sum_{i=1}^{n}(\frac{a_{ij}+b_{ij}}{2} - \overline{X}_j)^2, \tag{2.6}$$

and is independent of the distribution within the intervals.

2.2 Covariance

Let X_{j_1} and X_{j_2} be two interval-valued random variables with pairwise observations: $X_{j_1} = [a_{ij_1}, b_{ij_1}]$ and $X_{j_2} = [a_{ij_2}, b_{ij_2}]$ on a random sample of size n. The following holds: $a_{ij} \leq b_{ij}$, for $j = j_1, j_2$, and $i = 1, 2, ..., n$. Total Sum of Products SPT is decomposed into two components, the Sum of Products Within, SPW, and the Sum of Products Between, SPB; it is connected to the covariance Cov:

$$nCov_{j_1j_2} = SPT_{j_1j_2} = SPW_{j_1j_2} + SPB_{j_1j_2}. \qquad (2.7)$$

The Sum of Products Within SPW and Sum of Products Between SPB are related to $CovW$ and $CovB$, respectively, which are expressed as follows:

$$CovW_{j_1j_2} = \frac{SSW_{j_1j_2}}{n} = \frac{1}{n}\sum_{i=1}^{n}\frac{(b_{ij_1} - a_{ij_1})(b_{ij_2} - a_{ij_2})}{12}, \qquad (2.8)$$

$$CovB_{j_1j_2} = \frac{SSB_{j_1j_2}}{n} = \frac{1}{n}\sum_{i=1}^{n}(\frac{a_{ij_1} + b_{ij_1}}{2} - \overline{X}_{j_1})(\frac{a_{ij_2} + b_{ij_2}}{2} - \overline{X}_{j_2}). \qquad (2.9)$$

It may be interesting to notice that the entries of the $CovW$ matrix are always positive, their magnitudes depend on the ranges, $R_{ij} = b_{ij} - a_{ij}$, $j = j_1, j_2$; the greater the ranges of the two variables the greater is the entry of $CovW$. It should be pointed out that $CovW$ is not a true covariance matrix on the ranges; the terms for the true covariance matrix on the ranges would be $(R_{ij_1} - \overline{R}_{j_1})(R_{ij_2} - \overline{R}_{j_2})$. However, the $CovW$ matrix incorporates information on the size of the rectangles.

The entries of $CovB$ are classical covariances (divided by n not by $n - 1$) on the interval midpoints. When, instead of the intervals $[a, b]$, PCA is performed on the interval midpoints: $[(a + b)/2$, $(a + b)/2]$, $CovW$ is zero and $Cov = CovB$; in this case, the symbolic PCA results are the same as for a classical PCA on the interval midpoints.

Billard (2008) showed that the covariance between two interval-valued variables X_{j_1} and X_{j_2} can be calculated directly, using the following expression:

$$\begin{aligned} Cov_{j_1j_2} = \frac{1}{6n}\sum_{i=1}^{n}[2\,(a_{ij_1} - \overline{X}_{j_1})(a_{ij_2} - \overline{X}_{j_2}) + (a_{ij_1} - \overline{X}_{j_1})(b_{ij_2} - \overline{X}_{j_2}) \\ + (b_{ij_1} - \overline{X}_{j_1})(a_{ij_2} - \overline{X}_{j_2}) + 2\,(b_{ij_1} - \overline{X}_{j_1})(b_{ij_2} - \overline{X}_{j_2})] \qquad (2.10) \end{aligned}$$

Two special cases are easily checked: a) covariance of two identical variables equals its variance; b) covariance of two classical variables equals the well known classical covariance.

Figure 1 gives some insight into the calculation of the covariance in the classical and interval setting. Covariance in the classical setting is based on the position of the points, in the interval setting it is based on the rectangles: the location of the midpoints determines the between part, the size of the rectangles determines the within part, which

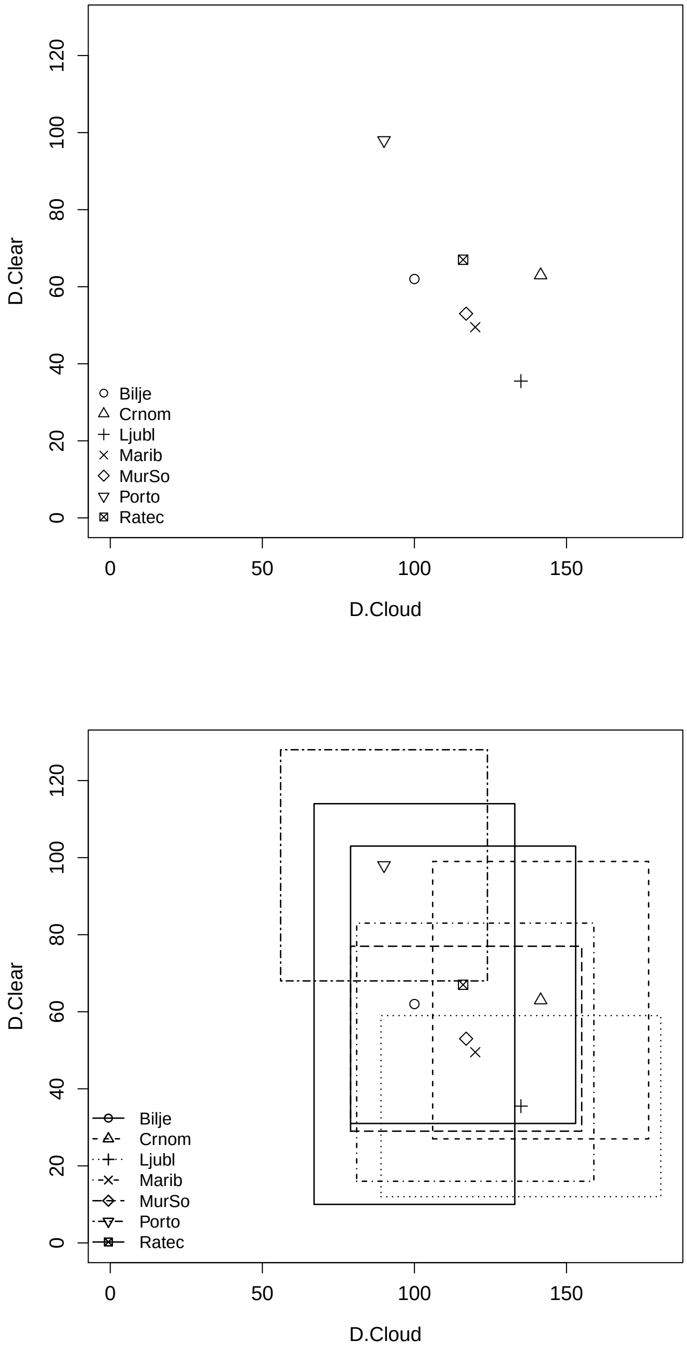

Figure 1: Calculation of the covariance in the classical setting (upper part) is based on the position of the points; in the interval setting (lower part) it is based on the rectangles: position of the midpoints and size of the rectangles determine its value.

is always positive. Covariance is calculated on the number of clear days (D.Clear) and the number of cloudy days (D.Cloud) for seven meteorological stations (for details, see next section). Figure 1 (upper part) illustrates the classical covariance, the position of the points suggests that the covariance is negative, the same would be expected from the subject-matter knowledge; the obtained value is $Cov(D.Cloud, D.Clear) = -212.0$. However, the covariance in the interval setting is positive, $Cov(D.Cloud, D.Clear) = +202.2$. This is due to a large within interval component $CovW(D.Cloud, D.Clear) = 411.7$, the between component is the same as classical covariance on the midpoints, $CovB(D.Cloud, D.Clear) = -212.0$; see Figure 1 (lower part).

2.3 Principal component analysis in the context of interval-valued data

A crucial advantage of the symbolic covariance matrix Cov is that it fully utilizes all the information in the data. It can be decomposed into a within part $CovW$ and a between part $CovB$. This decomposition allows for a deeper insight into the PCA results from the traces of these matrices. Since the trace of a matrix is a linear operator, the following holds:

$$tr(Cov) = tr(CovW) + tr(CovB). \tag{2.11}$$

Hence, we can assess the proportion of variance explained due to the within information and the proportion of variance explained due to the between information. The information gain when performing PCA on the intervals instead of the interval midpoints (conditionally the means) is due to the within information.

Additional PCA analysis can be done on $CovB$, these results are equivalent to the classical PCA results on the interval midpoints. A PCA analysis can also be performed on $CovW$; the interpretation of these results may enlighten some of the aspects of the within information.

3 A case study

We consider yearly data from the period 1971-2010 in Slovenia, data were collected by Slovenian Environment Agency (http://meteo.arso.gov.si/met/sl/archive/), and are shown in the Appendix. The following variables are taken into account: number of cold days (D.Cold), number of warm days (D.Warm), number of days with storms (D.Storm), number of days with precipitations (D.Prec), number of days with snow cover (D.SnCov), number of clear days (D.Clear), and number of cloudy days (D.Cloud). According to meteorological definitions, for a cold day the minimal daily air temperature is below 0 ^{0}C, for a warm day the maximal daily temperature is above 25 ^{0}C; a clear day has under 20% of cloudiness, a cloudy day has over 80%. Hence, D.Cold and D.Warm are based on the same variable, i.e., air temperature, the same holds for D.Clear and D.Cloud which are based on cloudiness.

For illustrative simplicity, only seven meteorological stations are chosen for this case study. They are: Bilje (Bilje), Črnomelj (Crnom), Ljubljana (Ljubl), Maribor (Marib),

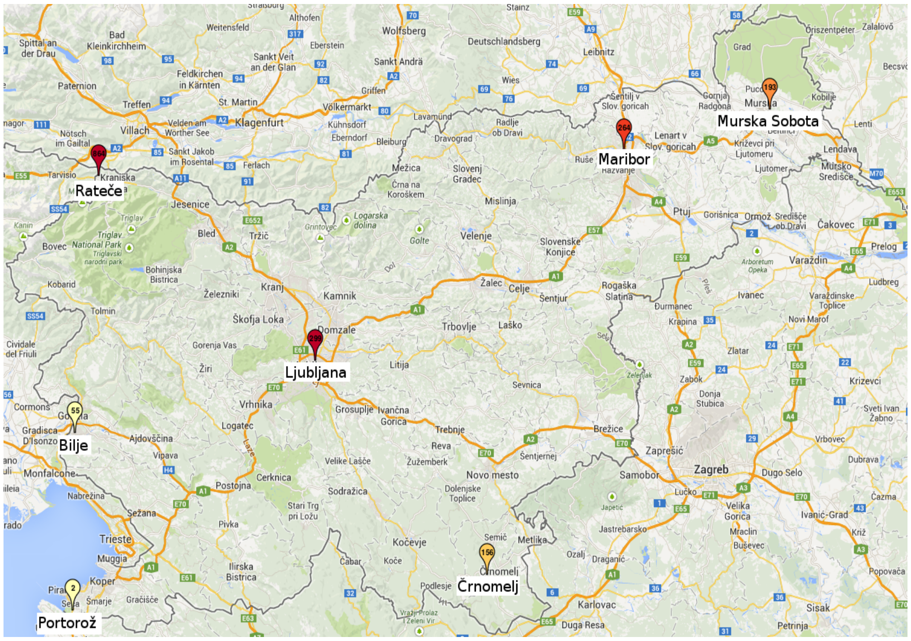

Figure 2: Geographical position of seven meteorological stations under study: Bilje (Bilje), Črnomelj (Crnom), Ljubljana (Ljubl), Maribor (Marib), Murska Sobota (MurSo), Portorož-airport (Porto), and Rateče (Ratec); elevation (in meters) is pinned to each station.

Murska Sobota (MurSo), Portorož-airport (Porto), and Rateče (Ratec). Their location is shown in Figure 2. Portorož-airport is situated at sea level (elevation 2 m), Rateče is in the Alps (elevation 864 m), the other stations have elevation from 55 m to 299 m. The dataset is slightly incomplete: data for Portorož-airport started in 1975, for Bilje, Črnomelj, Maribor and Murska Sobota data for some years are inconsistently missing.

As already stated, we want to compare PCA results obtained on different data types which were aggregated from the same dataset. For each station and each variable, we aggregated the data in two different ways: the mean value and the $[min, max]$ interval which is based on the minimal and maximal values in the period under observation.

3.1 PCA on the Means

In Table 1, the classical covariance matrix calculated on the means is presented; the sum of variances (3891.8) is given below the matrix. Dominant variances are as follows: $Var(D.SnCov) = 1449.4$, $Var(D.Cold) = 1284.6$; dominant covariances are: $Cov(D.SnCov, D.Cold) = 1232.5$ (positive), $Cov(D.SnCov, D.Warm) = -679.3$ and $Cov(D.Warm, D.Cold) = -558.2$ (negative).

In Table 2, the PCA results are given. The first two principal components explain about 92% of total variance, the first three around 97%. The loads for the first three principal components are also presented; we shall interpret the first two principal components only. For the first principal component (PC_1) D.Cold and D.SnCov are dominant, for the second principal component (PC_2) D.Clear and D.Cloud show up. We can deduce that PC_1 is positively correlated with low air temperature and PC_2 with the surplus of cloudy

Table 1: Covariance matrix calculated on the means. The sum of the variances (in the table in bold) is given below the matrix.

	D.Cold	D.Warm	D.Storm	D.Prec	D.SnCov	D.Clear	D.Cloud
D.Cold	**1284.6**	-558.2	-182.0	308.1	1232.5	-262.3	204.8
D.Warm	-558.2	**356.7**	45.1	-96.1	-679.3	70.1	-28.3
D.Storm	-182.0	45.1	**47.3**	-29.3	-109.4	51.4	-29.5
D.Prec	308.1	-96.1	-29.3	**212.9**	337.8	-148.6	195.5
D.SnCov	1232.5	-679.3	-109.4	337.8	**1449.4**	-216.3	177.6
D.Clear	-262.3	70.1	51.4	-148.6	-216.3	**296.8**	-212.0
D.Cloud	204.8	-28.3	-29.5	195.5	177.6	-212.0	**244.1**

Sum of variances = **3891.8**

Table 2: PCA on the means, results for the first three principal components: cumulative percentage of variance explained, principal component loads (dominant loads are in bold).

	PC_1	PC_2	PC_3
Cum.% of var. exp.	79.2	92.2	96.7
D.Cold	**0.625**	0.014	**0.661**
D.Warm	-0.307	0.274	0.192
D.Storm	-0.071	-0.057	-0.391
D.Prec	0.172	0.385	-0.322
D.SnCov	**0.670**	-0.218	-0.473
D.Clear	-0.138	**-0.598**	-0.090
D.Cloud	0.113	**0.607**	-0.193

over clear days.

Figure 3 presents the seven stations in the space of PC_1 by PC_2. There is a positive trend with low air temperature along PC_1: Portorož-airport reveals few days with low air temperature and snow cover, Rateče the opposite. This is consistent with the fact that Portorož-airport is located near the Adriatic sea, Rateče is located in the Alps. There is a positive trend in the surplus of cloudy over clear days along PC_2; here, Portorož-airport has the lowest surplus (it has more clear than cloudy days), Ljubljana and Črnomelj have the highest (here, there are more cloudy than clear days).

3.2 Symbolic PCA on interval-valued variables

3.2.1 Symbolic covariance matrix and its decomposition

The symbolic covariance matrix Cov for the intervals is given in Table 3; also shown is the decomposition into $CovB$ and $CovW$. The term $CovB$ is identical to the classical covariance matrix on the interval midpoints. Values of $CovW$ reflect the internal variability and are all positive. Consequently, the terms in Cov are always larger than the corresponding terms in $CovB$; thus, there are fewer negative terms in Cov than in $CovB$.

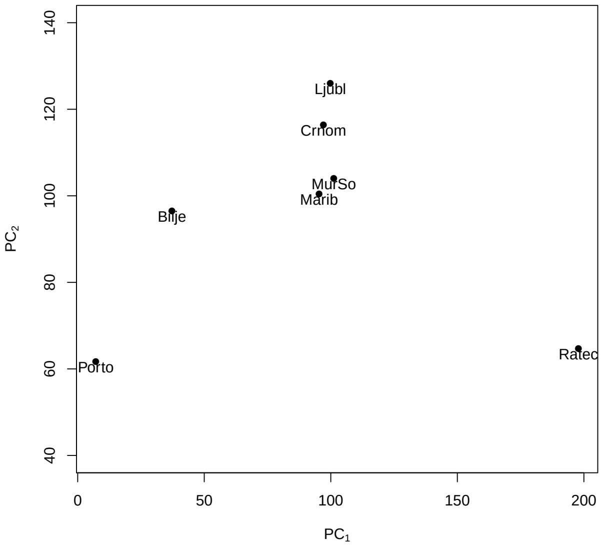

Figure 3: PCA on the means: presentation of seven stations in two-dimensional space of PC_1 by PC_2; 92.2% of total variance is explained. PC_1 reflects positive impact of low air temperature, PC_2 reflects positive impact of surplus of cloudy over clear days.

The sum of symbolic variances is 5754.2, the between component explains 3170 (55.1%), and the remaining 2563.2 (44.9%) is due to the within component. In this case, we can conclude that the gain in information, when we analyze the intervals instead of the interval midpoints, is large, it is nearly 45%. Let us find out the corresponding impact on the PCA results.

3.2.2 PCA on symbolic covariance matrix

Table 4 shows the PCA results based on the symbolic covariance matrix. The first two principal components explain 86.4% of variance, the first three 95.1%. For PC_1, the loads for D.Cold and D.SnCov are dominant, for PC_2 the dominant loads are D.Warm and D.Clear (positive), for PC_3 D.Clear (negative). Hence, the PC_1 is positively correlated with low air temperature, as in the PCA on the means; however, other results are different: PC_2 is positively correlated with D.Warm and D. Clear, PC_3 is negatively correlated with D.Clear.

Visualisation of these PCA results in two-dimensional space is based on the approach presented in Le-Rademacher and Billard (2012). For each station, a 7-dimensional polytope is obtained. Figure 4 (upper plot) presents the projection of these polytopes onto the PC_1 by PC_2 plane. Considerable overlapping is presented. The plot shows that the variability in PC_1 (D.Cold and D.SnCov) is dominant, for Rateče it is the greatest; however, variability in PC_2 (D.Warm and D.Clear) is comparable for all stations. Only two pairs of stations do not overlap: Portorož-airport and Rateče, Bilje and Rateče. The polytopes for two extreme stations, Portorož-airport and Rateče, are presented on Figure 4 (lower plot).

Table 3: Covariance matrix Cov for interval-valued variables, variances are in bold; it is decomposed into $CovB$ and $CovW$ (below). The respective sum of variances is presented below the corresponding matrix.

Cov	D.Cold	D.Warm	D.Storm	D.Prec	D.SnCov	D.Clear	D.Cloud
D.Cold	**1363.7**	-74.1	30.8	537.4	1313.6	105.0	630.2
D.Warm	-74.1	**720.3**	224.4	262.1	-53.4	399.3	407.2
D.Storm	30.8	224.4	**153.9**	146.1	127.6	265.9	191.4
D.Prec	537.4	262.1	146.1	**443.2**	615.6	163.6	557.6
D.SnCov	1313.6	-53.4	127.6	615.6	**1595.7**	187.5	750.6
D.Clear	105.0	399.3	265.9	163.6	187.5	**724.6**	202.2
D.Cloud	630.2	407.2	191.4	557.6	750.6	202.2	**752.9**

Sum of variances = **5754.2**

$CovB$	D.Cold	D.Warm	D.Storm	D.Prec	D.SnCov	D.Clear	D.Cloud
D.Cold	**1056.0**	-435.8	-144.9	258.0	941.1	-233.8	251.5
D.Warm	-435.8	**286.8**	19.2	-67.0	-482.5	8.8	-39.2
D.Storm	-144.9	19.2	**46.3**	-18.4	-68.8	60.1	-24.8
D.Prec	258.0	-67.0	-18.4	**183.1**	282.0	-148.7	209.8
D.SnCov	941.1	-482.5	-68.8	282.0	**997.7**	-196.6	270.7
D.Clear	-233.8	8.8	60.1	-148.7	-196.6	**322.3**	-209.5
D.Cloud	251.5	-39.2	-24.8	209.8	270.7	-209.5	**278.9**

Sum of between variances = **3171.0**

$CovW$	D.Cold	D.Warm	D.Storm	D.Prec	D.SnCov	D.Clear	D.Cloud
D.Cold	**307.7**	361.7	175.8	279.4	372.5	338.8	378.7
D.Warm	361.7	**433.5**	205.2	329.1	429.2	390.5	446.5
D.Storm	175.8	205.2	**107.6**	164.5	196.3	205.9	216.1
D.Prec	279.4	329.1	164.5	**260.2**	333.6	312.2	347.8
D.SnCov	372.5	429.2	196.3	333.6	**598.0**	384.1	479.9
D.Clear	338.8	390.5	205.9	312.2	384.1	**402.2**	411.7
D.Cloud	378.7	446.5	216.1	347.8	479.9	411.7	**474.1**

Sum of within variances = **2583.2**

Table 4: PCA on the intervals, results for the first three principal components: cumulative percentage of variance explained, principal component loads (dominant loads are in bold).

	PC_1	PC_2	PC_3
Cum.% of var. exp.	62.0	86.4	95.1
D.Cold	**0.569**	-0.277	-0.102
D.Warm	0.081	**0.663**	0.279
D.Storm	0.079	0.261	-0.119
D.Prec	0.309	0.172	0.271
D.SnCov	**0.636**	-0.220	-0.180
D.Clear	0.127	**0.512**	**-0.767**
D.Cloud	0.384	0.275	0.452

From these plots, it is observed that the internal variability for Rateče is greater than it is for Portorož-airport.

3.2.3 PCA on $CovB$ and $CovW$

We proceed with PCA on $CovB$, this is identical to the classical PCA on the interval midpoints, the results are in Table 5 (left) and are plotted in Figure 5 (upper plot); they are consistent with the PCA results on the means.

Since $CovW$ depicts the within interval information, PCA on $CovW$ allows an insight into the variability within the interval variables, see Table 5 (right) and Figure 5 (lower plot). In this case, the PC_1 explains 93.2%, the PC_2 explains additional 5.4%. The loads for PC_1 for all variables have similar magnitude, while for PC_2 the dominant load is D.SnCov; accordingly, PC_1 is positively related to all the variables, PC_2 is positively related to D.SnCov. The scores are calculated using the midpoints. The stations are located along the diagonal, from Portorož-airport at the lower end to Rateče at the upper end, revealing the increase of interval variability from the lower to the upper end. This result is consistent with the fact that Portorož-airport has tighter intervals, Rateče has larger intervals.

3.2.4 Programs used

Algorithms for deriving the PCA results on the symbolic covariance matrix along with the corresponding polytops are available at Le-Rademacher and Billard (2012, Supplementary material - online version). Their R script (R Core Team, 2013) was upgraded with PCA on $CovW$ and $CovB$ and adapted for our case-study.

3.3 Other PCA approaches for interval-valued variables

Other PCA approaches on interval data are described in the literature. As stated before, Le-Rademacher and Billard (2012) give a detailed insight into these methods. We shall limit ourselves to only some of them.

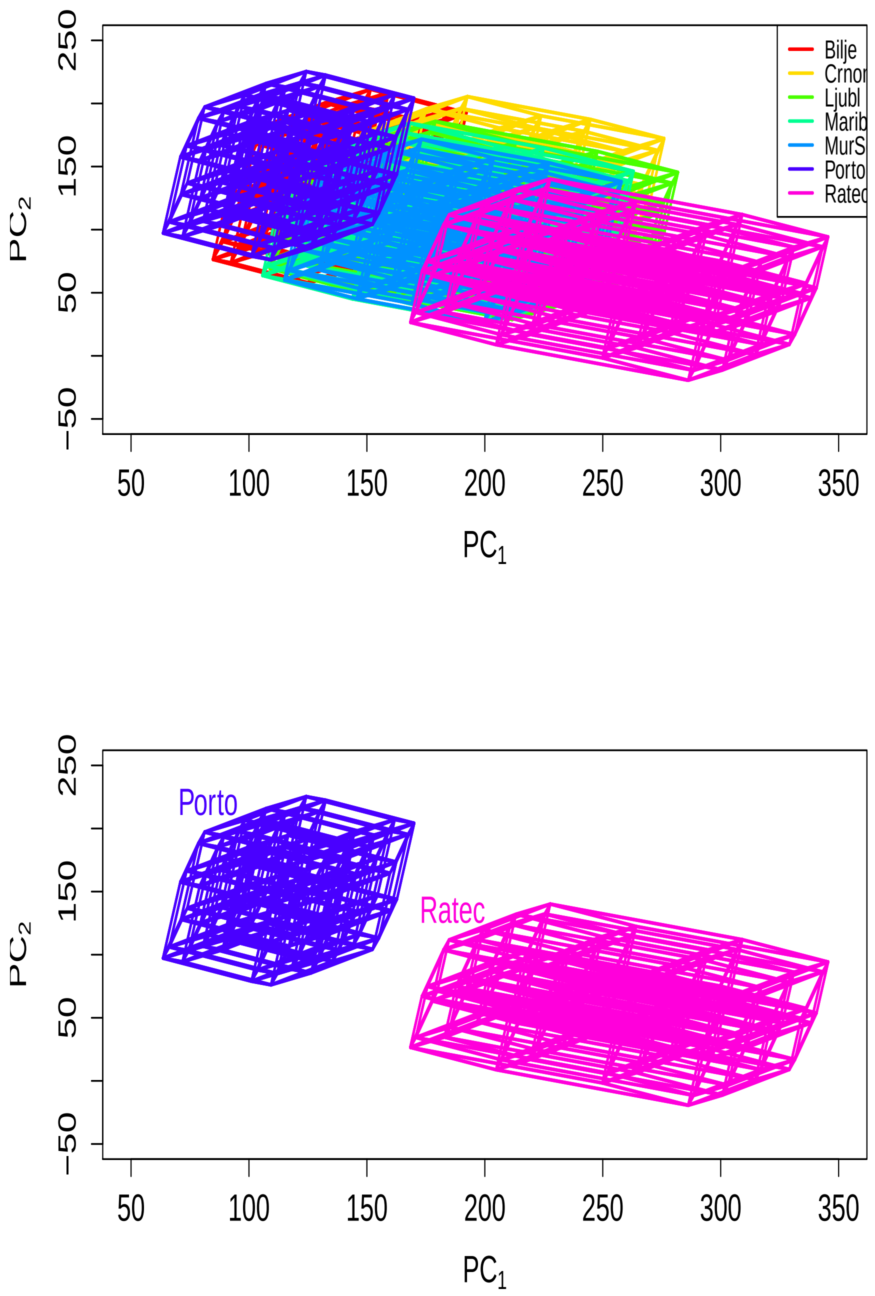

Figure 4: Projection of 7-dimensional polytopes onto 2-dimensional space of PC_1 by PC_2 , upper plot: for all 7 stations; lower plot: for Portorož-airport and Rateče. PC_1 explains 62.0% of variance, it reflects the positive impact of D.Cold and D.SnCov; PC_2 explains 24.4% of variance, it reflects the positive impact of D.Warm and D.Clear.

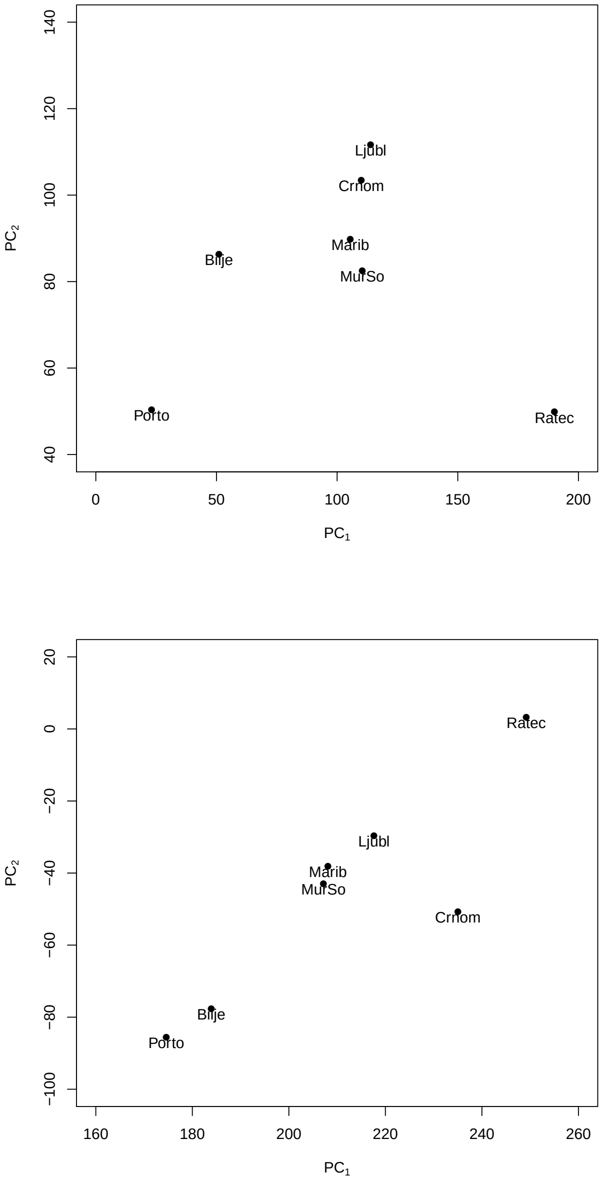

Figure 5: Upper plot: PCA on CovB (this is identical to classical PCA on midpoints); PC_1 reflects the positive impact of D.Cold and D. Sncov; PC_2 reflects the surplus of D.Cloud over D.Clear. Lower plot: PCA on CovW: PC_1 reflects the positive impact of all variables; PC_2 the positive impact of D.SnCov.

Table 5: PCA on the CovB (left), PCA on CovW (right); results for the first three principal components: cumulative percentage of variance explained, principal component loads (dominant loads are in bold).

	PC_1	PC_2	PC_3	PC_1	PC_2	PC_3
Cum.$\%$ of var. exp.	75.9	91.5	96.8	93.2	98.6	99.7
D.Cold	**0.642**	-0.118	**0.508**	0.355	-0.126	0.118
D.Warm	-0.286	0.376	0.175	0.416	-0.189	**0.603**
D.Thund	-0.068	-0.050	-0.423	0.203	-0.221	-0.194
D.Prec	0.193	0.349	-0.348	0.324	-0.172	0.049
D.SnCov	**0.630**	-0.161	-0.363	0.454	**0.847**	-0.207
D.Clear	-0.168	**-0.627**	-0.380	0.390	-0.391	**-0.700**
D.Cloud	0.198	**0.549**	-0.368	0.443	-0.018	0.221

3.3.1 Centers method

The centers method transforms the interval-valued matrix into a classical matrix of the interval midpoints. The results of this method are given as a part of the PCA approach on symbolic covariance matrix: see $CovB$ in Table 3, PCA results in Table 5 (left) and Figure 5 (upper plot). As already stated, in this approach the internal interval variance is completely ignored.

3.3.2 Vertices method

In this approach, the vertices of the hyper-rectangles (instead of the interval midpoints) are considered as the data-input. In our case, seven variables were taken into account; therefore, there are $2^7 = 128$ vertices. Thus, the dimension of the input matrix is $n = 128$, $p = 7$; classical PCA is performed on this matrix.

Here, we do not present the covariance matrix. The sum of variances equals 10932.8, which is approximately twice the value in the symbolic context (5754.2). Table 6 presents the PCA results for the first three principal components. In this case, PC_1 explains *only* 34.8% of the total variance; the first two principal components 51.3% and the first three 64.7%. For PC_1, D.Cold and D.SnCov are dominant; PC_2 is positively correlated with D.Cloud and negatively with D.Clear (as in the PCA on the means or midpoints); for PC_3 D.Warm and D.SnCov show up, surprisingly, both loads are positive.

We can summarize, that this approach is simple, it always works, but it fails to use all the variation in the data. The results reflect that the data matrix is artificially inflated; the vertices are treated as independent observations, this assumption is not sustainable. Our results are consistent with Douzal-Chouakria et al. (2011), who showed that the variance of the vertices in fact includes some but not all of the internal variation.

3.3.3 The midpoint-radii method

The *midpoint-radii* approach treats a single interval-valued variable as two variables: midpoints and midranges. A PCA can be performed on either of them. This is similar to the

Table 6: PCA on the vertices; results for the first three principal components: cumulative percentage of variance explained, principal component loads (dominant loads are in bold).

Cum.% of var. exp.	PC_1 34.8	PC_2 51.3	PC_3 64.7
D.Cold	**0.525**	0.037	-0.460
D.Warm	-0.276	0.284	**0.660**
D.Thund	-0.044	-0.024	0.041
D.Prec	0.151	0.181	-0.012
D.SnCov	**0.750**	-0.256	**0.579**
D.Clear	-0.149	**-0.514**	0.111
D.Cloud	0.196	**0.745**	0.050

PCA on $CovB$ and $CovW$; the only difference is, that $CovW$ is an uncentered covariance matrix on the ranges.

To analyze the midpoint and the range data simultaneously, Palumbo and Lauro (2003) propose to superimpose the PCs of the midrange onto the PCs on the midpoint and then rotate the midrange PC axes to maximize the connection between the midpoints and the midranges. It turns out the choice of rotation operator is subjective; the midpoints and the midranges are treated as independent (see Lauro et al., 2008). Le-Rademacher and Billard (2012) showed that the midpoint-radii approach is deficient and not working properly. Due to these facts, we believe that this approach should be replaced by the PCA on the symbolic covariance matrix; see the results given in Table 3 above, Table 4 and Figure 4.

4 Conclusions

A crucial advantage of the symbolic covariance matrix Cov is that it fully utilizes all the information in the data. It can be decomposed into a within part $CovW$ and a between part $CovB$. In the interpretation of the Cov term, we should recognise that: it is the sum of the classical covariance on the interval midpoints and a measure of variability (i.e., the size) of the intervals. Therefore, the sign of $CovB$ may be negative and the sign of Cov positive. Figure 1 illustrates such a case.

However, this decomposition allows for a deeper insight into the interval-valued dataset: from the traces of these matrices, the proportion of variance explained due to the within information and the proportion of variance explained due to the between information can be calculated. The information gain when performing PCA on the intervals instead of the interval midpoints (conditionally the means) is due to the within information.

We can summarize the PCA approach on Cov as follows: the interpretation of the PC should be the ”symbolic context”; visualization of the results using the projection of the polytopes is suitable for lower dimensions of p and n, for higher dimensions the plot can be unreadable. We suggest that separate PCA's on both the $CovB$ and the $CovW$ should be done additionally to allow for a deeper understanding of the between and within information. The analysis of PCA results on $CovB$ is straightforward, as in the classical context on the interval midpoints. However, the PCA results on $CovW$ are interpretable

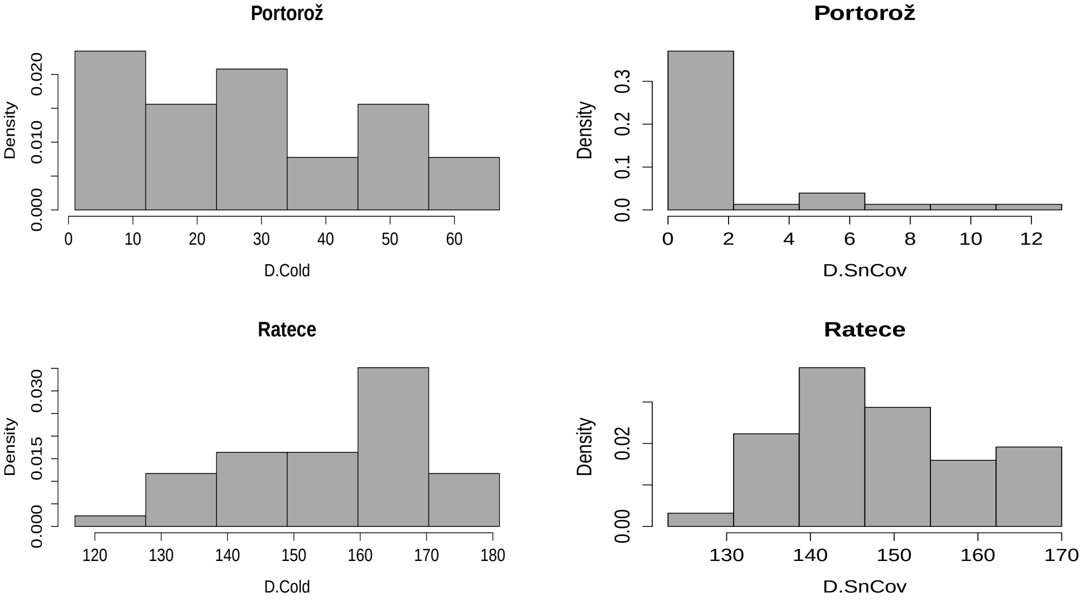

Figure 6: Histograms for D.Cold and D.SnCov for Portorož-airport and Rateče revealing different types of distribution.

in the context of the size of the rectangles.

In the case study presented, the information gain when performing PCA on the intervals instead of the interval midpoints (conditionally the means) is about 45%. For the PCA results on Cov, it may be difficult to grasp the meaning of the PC_2; however, the PCA results obtained on $CovB$ and $CovW$ are consistent with the subject-matter knowledge.

There is an important assumption hidden in this analysis: the distribution of the values along each $[min, max]$ interval should be uniform. This is often not the case, in particular when data for meteorological variables over a longer period are under study; for illustration, see some histograms of the raw data used herein in Figure 6. It is obvious that the uniformity assumption does not hold. Therefore, it may be interesting to analyze the histogram-valued variables and compare the results with the results obtained on the interval-valued variables.

References

[1] Bertrand, P. and Goupil, F. (2000): Descriptive statistics for symbolic data. In H.-H. Bock and E. Diday (Ed): *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, 103-124. Berlin: Springer-Verlag.

[2] Billard L. and Diday E. (2003): From the statistics of data to the statistics of knowledge: Symbolic data analysis. *Journal of American Statistical Association*, **98**, 470-487.

[3] Billard L. and Diday E. (2006): *Symbolic Data Analysis: Conceptual Statistics and Data Mining.* Wiley Series in Computational Statistics.

[4] Billard L. (2008): Sample Covariance Functions for Complex Quantitative Data. In Mizuta M. and Nakano J. (Ed): *Proceedings of the International Association of Statistical Computing Conference 2008*, 157-163. Yokohama.

[5] Billard L. (2011): Brief overview of symbolic data and analytical issues. *Statistical Analysis and Data Mining*, **4**, 149-156.

[6] Billard L. and Le-Rademacher J. (2012): Principal component analysis for interval data. *WIREs Comput Stat 2012*, 4:535-540. doi: 10.1002/wics.1231.

[7] Bock, H.-H. and Diday, E. (eds.) (2000): *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information form Complex Data*. Berlin: Springer-Verlag.

[8] Cazes, P., Chouakria, A., Diday, E. and Schektman, Y. (1997): Extension de l'Analyse en Composantes Principales à des Données de Type Intervalle. *Revue de Statistique Appliquée*, **45(3)**, 5-24.

[9] Chattfield C. and Collins A. J. (1980): *An Introduction to Multivariate Analysis*. Chapman and Hall.

[10] Douzal-Chouakria, A., Billard, L. and Diday, E. (2011): Principal Component Analysis for Interval-valued Observations. *Statistical Analysis and Data Mining*, **4**, 229-246.

[11] Gioia, F. and Lauro, C. (2006): Principal Component Analysis on Interval Data. *Computational Statistics*, **21**, 343-363.

[12] Giordani, P. and Kiers, H. A. L. (2006). A comparison of three methods for Principal Component Analysis of fuzzy interval data. *Computational Statistics and Data Analysis*, special issue *The Fuzzy Approach to Statistical Analysis*, **51(1)**, 379-397.

[13] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, **24**, 417-441, and 498-520.

[14] Hotelling, H. (1936). Relations between two sets of variates. Biometrika, **27**, 321-377.

[15] Irpino, A. (2006). Spaghetti PCA analysis: An extension of principal component analysis to time dependent interval data. Pattern Recognition Letters, **27**, 504-513.

[16] Johnson, R. A. and Wichern D.W. (2002): *Applied Multivariate Statistical Analysis* (5^{th} edition). Prentice Hall.

[17] Lauro, N. C. and Palumbo, F. (2000). Principal Component Analysis of Interval Data: A Symbolic Data Analysis Approach. *Computational Statistics*,**15**, 73-87.

[18] Lauro, N. C., Verde, R. and Irpino, A. (2008): Principal Component Analysis of Symbolic Data Described by Intervals. In E. Diday, and M. Noirhomme-Fraiture (Ed.): *Symbolic Data Analysis and the SODAS Software*, 279-311. Chichester: Wiley.

[19] Le-Rademacher J. (2008): Principal component analysis for interval-valued and histogram-valued data and likelihood functions and some maximum likelihood estimators for symbolic data. Doctoral Dissertation, University of Georgia.

[20] Le-Rademacher J. and Billard L. (2012): Symbolic-covariance principal component analysis and visualization for interval-valued data. *Journal of Computational and Graphical Statistics*, **21(2)**, 413-432. Epub 14 Jun 2012.

[21] Le-Rademacher J. and Billard L. (2013): Principal component histograms from interval-valued observations. *Computational Statistics 2013*, **28**, 2117-2138. doi 10.1007/s00180-013-0399-4.

[22] Le-Rademacher, J. and Billard, L. (2013). Principal component analysis for histogram-valued data. Technical report.

[23] Palumbo, F. and Lauro, N. C. (2003): A PCA for Interval-Valued Data Based on Midpoints and Radii. In H. Yanai, A. Okada, K. Shigemasu, Y. Kanu, and J. Meulman (Ed): *New Developments in Psychometrics*, 641-648. Tokyo: Psychometric Society and Springer-Verlag.

[24] Pearson, K. (1901): On lines and planes of closest fit to a system of points in space. *Philosophical Magazine*, **2**, 557-72.

[25] R Core Team (2013): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[26] Zuccolotto, P. (2007): Principal components of sample estimates: an approach through symbolic data analysis. Statistical Methods and Applications, **16 (2)**, 173-192.

A Appendix

The data used are yearly data from the period 1971-2010 for seven stations in Slovenia. The following variables are taken into account: number of cold days (D.Cold), number of warm days (D.Warm), number of days with storms (D.Storm), number of days with precipitations (D.Prec), number of days with snow cover (D.SnCov), number of clear days (D.Clear), and number of cloudy days (D.Cloud). For each station, min and max values are given for each variable under study.

	D.Cold		D.Warm		D.Storm		D.Prec		D.SnCov		D.Clear		D.Cloud	
Station	min	max	min	max	min	max	min	max	min	max	min	max	min	max
Bilje	38	96	67	125	9	63	97	160	0	12	10	114	67	133
Crnom	65	120	48	118	23	59	128	185	6	88	27	99	106	177
Ljubl	52	112	38	109	30	63	119	186	2	110	12	59	89	181
Marib	56	123	37	110	23	52	110	162	3	92	16	83	81	159
MurSo	77	131	33	109	18	47	107	154	0	85	29	77	79	155
Porto	1	67	34	125	37	71	88	143	0	13	68	128	56	124
Ratec	117	181	6	67	22	52	123	170	43	171	31	103	79	153

Metodološki zvezki, Vol. 11, No. 1, 2014, 21-29

Optimal Unbiased Estimates of $P\{X < Y\}$ for Some Families of Distributions

Marko Obradović, Milan Jovanović, Bojana Milošević[1]

Abstract

In reliability theory, one of the main problems is estimating parameter $R = P\{X < Y\}$. In this paper we shall present UMVUEs for R in different cases i.e. for different distributions of X and Y. Some of them are already existing and some are original.

1 Introduction

In reliability theory the main parameter is the reliability of a system, and its estimation is one of the main goals. The system fails if the applied stress X is greater than strength Y, so R is a measure of system performance. In most cases this parameter is given as $R = P\{X < Y\}$, although for some discrete cases the expression $R = P\{X \leq Y\}$ is also considered.

The problem was first introduced by Birnbaum (1956). Since then numerous papers have been published. Most of results are presented in (Kotz et al., 2003). The vast majority of papers presuppose independence of stress and strength variables, as well as that they come from the same family of, in most cases continuous, distributions. There exists a wide range of applications in engineering, military, medicine and psychology.

The unbiasedness of an estimator is a desired property especially when dealing with relatively small sample sizes, where we cannot count on asymptotic unbiasedness. Since in many cases most popular estimators are biased, it is often important to find the unique minimum variance unbiased estimator (UMVUE).

1.1 UMVUE of R

Let $\mathbf{X} = (X_1, \ldots, X_{n_1})$ and $\mathbf{Y} = (Y_1, \ldots, Y_{n_2})$ be the samples from the distributions of random variables X and Y. Then, using the following theorem we can construct UMVUEs.

Theorem 1 *If $V(\mathbf{X}, \mathbf{Y})$ is any unbiased estimator of parameter θ and T is a complete sufficient statistic for θ, then $E(V(\mathbf{X}, \mathbf{Y})|T)$ is the UMVUE of θ.*

[1] Faculty of Mathematics, University of Belgrade, Studenski trg 16, Belgrade, Serbia; email addresses: marcone@matf.bg.ac.rs, mjovanovic@matf.bg.ac.rs, bojana@matf.bg.ac.rs

This theorem is the combination of Rao-Blackwell and Lehmann-Scheffé theorems. Their proofs could be found in (Hogg et al., 2005).

However, in continuous case, the use of this theorem might not be technically convenient. Therefore, the following theorems were proposed to help deriving the UMVUEs (Kotz et al., 2003).

Theorem 2 *Let $\theta_0 \in \Theta$ be an arbitrary value of θ and let T be a complete sufficient statistic for θ. Denote by $g_{\theta_0}(t)$ and $g_{\theta_0}(t|X_1 = x_1, ..., X_k = x_k, Y_1 = y_1, ..., Y_k = y_k)$ the pdf of T and the conditional pdf of T for given $X_j = x_j, Y_j = y_j,\ j = 1, ..., k$, respectively. Then the $UMVUE$ of joint pdf $f_\theta(x_1, ..., x_k, y_1, ..., y_k)$ is of the form*

$$\widehat{f}(x_1, ..., x_k, y_1, ..., y_k) =$$
$$\prod_{j=1}^{k} f_{\theta_0}(x_j, y_j) \frac{g_{\theta_0}(t|X_1 = x_1, ..., X_k = x_k, Y_1 = y_1, ..., Y_k = y_k)}{g_{\theta_0}(t)}.$$

Theorem 3 *The UMVUE of R is*

$$\widehat{R} = \int\int I(x < y)\widehat{f}(x, y)dxdy,$$

where $\widehat{f}$ is given in theorem 2 for $k = 1$.

2 Existing results

In this section we present a brief summary of existing results obtained for UMVUEs of R for some distributions.

- **Exponential distribution**

  Let X and Y be independent exponentially distributed random variables with densities

  $$f_X(x; \alpha) = \alpha e^{-\alpha x},\ x \geq 0,$$

  $$f_Y(y; \beta) = \beta e^{-\beta y},\ y \geq 0,$$

  where α and β are unknown positive parameters. The complete sufficient statistics for α and β are $T_X = \sum\limits_{j=1}^{n_1} X_j$ and $T_Y = \sum\limits_{j=1}^{n_2} Y_j$.

  The UMVUE of R was derived by Tong (1974; 1977), and it is given by

  $$\widehat{R} = \begin{cases} \sum\limits_{i=0}^{n_1-2} (-1)^i \frac{\Gamma(n_1)\Gamma(n_2)}{\Gamma(n_1-i-1)\Gamma(n_2+i+1)} \left(\frac{T_Y}{T_X}\right)^{i+1}, & \text{if } T_Y \leq T_X \\ \sum\limits_{i=0}^{n_2-1} (-1)^i \frac{\Gamma(n_1)\Gamma(n_2)}{\Gamma(n_1+i)\Gamma(n_2-i)} \left(\frac{T_X}{T_Y}\right)^{i}, & \text{if } T_Y > T_X. \end{cases}$$

- **Normal distribution**

Let X and Y be normally distributed independent random variables with densities

$$f_X(x;\mu_1,\sigma_1) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}},\ x \in \mathbb{R},$$

$$f_Y(y;\mu_2,\sigma_2) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}},\ y \in \mathbb{R},$$

where μ_1, σ_1^2, μ_2 and σ_2^2 are unknown parameters. The complete sufficient statistics for (μ_1, σ_1^2, μ_2, σ_2^2) are $(\bar{X}, S_X^2, \bar{Y}, S_Y^2)$.

The UMVUE of R was derived by Downtown (1973) and it is given by

$$\widehat{R} = \left[B\left(\frac{1}{2}, \frac{n_1-2}{2}\right) B\left(\frac{1}{2}, \frac{n_2-2}{2}\right)\right]^{-1} \int\limits_{\Omega} (1-u^2)^{\frac{n_1-4}{2}} (1-v^2)^{\frac{n_2-4}{2}} du dv,$$

where $B(a,b)$ is the beta function and

$$\Omega = \left\{(u,v) \in [-1,1] \times [-1,1] | -u\frac{S_X(n_1-1)}{\sqrt{n_1}} + v\frac{S_Y(n_2-1)}{\sqrt{n_2}} + (\bar{Y}-\bar{X}) > 0\right\}.$$

- **Gamma distribution**

Let X and Y be independent gamma distributed random variables with densities

$$f_X(x;\alpha_1,\sigma_1) = \frac{x^{\alpha_1-1}}{\Gamma(\alpha_1)\sigma_1^{\alpha_1}} e^{-\frac{x}{\sigma_1}},\ x \geq 0,$$

$$f_Y(y;\alpha_2,\sigma_2) = \frac{y^{\alpha_2-1}}{\Gamma(\alpha_2)\sigma_2^{\alpha_2}} e^{-\frac{y}{\sigma_2}},\ y \geq 0,$$

where α_1 and α_2 are known integer values and σ_1 and σ_2 are unknown positive parameters. The complete sufficient statistics for σ_1 and σ_2 are $T_X = \sum\limits_{j=1}^{n_1} X_j$ and $T_Y = \sum\limits_{j=1}^{n_2} Y_j$.

The UMVUE of R was derived by Constantine et al. (1986), and it is given by

$$\widehat{R} = \begin{cases} 1 - \sum\limits_{k=0}^{(n_2-1)\alpha_2-1} \frac{B(\alpha_1+\alpha_2+k,(n_1-1)\alpha_1)}{B(\alpha_1,(n_1-1)\alpha_1)B(\alpha_2,(n_2-1)\alpha_2)} & \\ \quad \times \binom{(n_2-1)\alpha_2-1}{k} \frac{(-1)^k}{\alpha_2+k} \left(\frac{T_X}{T_Y}\right)^{\alpha_2+k}, & \text{if } T_Y \leq T_X \\ \sum\limits_{k=0}^{(n_1-1)\alpha_1-1} \frac{B(\alpha_2+\alpha_1+k,(n_2-1)\alpha_2)}{B(\alpha_2,(n_2-1)\alpha_2)B(\alpha_1,(n_1-1)\alpha_1)} & \\ \quad \times \binom{(n_1-1)\alpha_1-1}{k} \frac{(-1)^k}{\alpha_1+k} \left(\frac{T_Y}{T_X}\right)^{\alpha_1+k}, & \text{if } T_Y > T_X. \end{cases}$$

- **Gompertz distribution**

  Let X and Y be independent Gompertz distributed random variables with densities

$$f_X(x;c,\lambda_1)=\lambda_1 e^{cx}e^{\frac{-\lambda_1(e^{cx}-1)}{c}},x>0,$$
$$f_Y(y;c,\lambda_2)=\lambda_2 e^{cy}e^{\frac{-\lambda_2(e^{cy}-1)}{c}},y>0,$$

  where c is a known positive value and λ_1 and λ_2 are unknown positive parameters. The complete sufficient statisitcs for λ_1 and λ_2 are

$$W_X=\frac{1}{c}\sum_{j=1}^{n_1}(e^{cX_j}-1), W_Y=\frac{1}{c}\sum_{j=1}^{n_2}(e^{cY_j}-1).$$

  The UMVUE of R was derived by Saracoglu et al. (2009) and it is given by

$$\widehat{R}=\begin{cases} 1-\sum\limits_{k=0}^{n_2-1}(-1)^k\frac{\Gamma(n_1)\Gamma(n_2)}{\Gamma(n_1+k)\Gamma(n_2-k)}\left(\frac{W_X}{W_Y}\right)^k, & \text{if } W_X<W_Y\\ \sum\limits_{k=0}^{n_1-1}(-1)^k\frac{\Gamma(n_1)\Gamma(n_2)}{\Gamma(n_1-k)\Gamma(n_2+k)}\left(\frac{W_Y}{W_X}\right)^k, & \text{if } W_X\geq W_Y.\end{cases}$$

- **Generalized Pareto distribution**

  Let X and Y be independent random variables from generalized Pareto distribution with densities

$$f_X(x;\alpha_1,\lambda)=\alpha_1\lambda(1+\lambda x)^{-(\alpha_1+1)},\ x>0,$$
$$f_Y(y;\alpha_2,\lambda)=\alpha_2\lambda(1+\lambda y)^{-(\alpha_2+1)},\ y>0,$$

  where λ is known positive value and α_1 and α_2 are unknown positive parameters. The complete sufficient statistics for parameters α_1 and α_2 are $T_X=\sum\limits_{j=1}^{n_1}\ln(1+X_j)$ and $T_Y=\sum\limits_{j=1}^{n_2}\ln(1+Y_j)$.

  The UMVUE of R was derived by Rezaei et al. (2010), and it is given by

$$\widehat{R}=\begin{cases} 1-\sum\limits_{k=0}^{n_2-1}(-1)^k\frac{\Gamma(n_1)\Gamma(n_2)}{\Gamma(n_1+k)\Gamma(n_2-k)}\left(\frac{T_X}{T_Y}\right)^k, & \text{if } T_X<T_Y\\ \sum\limits_{k=0}^{n_1-1}(-1)^k\frac{\Gamma(n_1)\Gamma(n_2)}{\Gamma(n_1-k)\Gamma(n_2+k)}\left(\frac{T_Y}{T_X}\right)^k, & \text{if } T_X\geq T_Y.\end{cases}$$

- **Poisson distribution**

  Let X and Y be independent Poisson distributed random variables with mass functions

$$P\{X=x;\lambda_1\}=\frac{e^{-\lambda_1}\lambda_1^x}{x!},x=0,1,\dots,$$
$$P\{Y=y;\lambda_2\}=\frac{e^{-\lambda_2}\lambda_2^y}{y!},y=0,1,\dots,$$

  where λ_1 and λ_2 are unknown positive parameters. The complete sufficient statistics for λ_1 and λ_2 are $T_X=\sum\limits_{j=1}^{n_1}X_j$ and $T_Y=\sum\limits_{j=1}^{n_2}Y_j$.

The UMVUE of R was derived by Belyaev and Lumelskii (1988) and it is given by

$$\widehat{R} = \sum_{x=0}^{\min\{T_X, T_Y-1\}} \binom{T_X}{x} \frac{(n_1-1)^{T_X-x}}{n_1^{T_X}} \left(1 - \sum_{y=0}^{x} \binom{T_Y}{y} \frac{(n_2-1)^{T_Y-y}}{n_2^{T_Y}} \right).$$

- **Negative binomial distribution**

  Let X and Y be indepent random variables from negative binomial distributions with mass functions

  $$P\{X = x; m_1, p_1\} = \binom{m_1+x-1}{x} p_1^x (1-p_1)^{m_1}, \;\; x = 0, 1, \ldots,$$

  $$P\{Y = y; m_2, p_2\} = \binom{m_2+y-1}{y} p_2^y (1-p_2)^{m_2}, \;\; y = 0, 1, \ldots,$$

  where m_1 and m_2 are known integer values and p_1 and p_2 are unknown probabilities. The complete sufficient statistics for p_1 and p_2 are $T_X = \sum_{j=1}^{n_1} X_j$ and $T_Y = \sum_{j=1}^{n_2} Y_j$.

  The UMVUE of R was derived by Ivshin and Lumelskii (1995) and it is given by

$$\widehat{R} = \sum_{x=0}^{\min\{T_X, T_Y-1\}} \sum_{y=x+1}^{T_Y} \frac{\binom{m_1+x-1}{x}\binom{T_X-x+m_1(n_1-1)-1}{T_x-x}}{\binom{m_1n_1+T_X-1}{T_X}} \frac{\binom{m_2+y-1}{y}\binom{T_Y-y+m_2(n_2-1)-1}{T_y-y}}{\binom{m_2n_2+T_Y-1}{T_Y}}.$$

3 New results

In this section we shall derive the UMVUE of R for some new distributions. The first model is where stress and strength both have Weibull distribution with known but different shape parameter and unknown rate parameters. As a special case we present the model where stress has exponential and strength has Rayleigh distribution. An example with real data for Weibull model is presented. In the second model, both stress and strength have logarithmic distribution with unknown parameters.

3.1 Weibull model

Let X and Y be independent random variables from Weibull distribution with densities

$$f_X(x; \alpha_1, \sigma_1) = \alpha_1 \sigma_1^{\alpha_1} x^{\alpha_1-1} e^{-(\sigma_1 x)^{\alpha_1}}, \; x \geq 0,$$
$$f_Y(y; \alpha_2, \sigma_2) = \alpha_2 \sigma_2^{\alpha_2} y^{\alpha_2-1} e^{-(\sigma_2 y)^{\alpha_2}}, \; y \geq 0.$$

The Weibull distribution is one of the most used distribution in modeling life data. Many researchers have studied the reliability of Weibull model. Most of them did not consider unbiased estimators (e.g. Kundu and Gupta, 2006), and recently the case with common known shape parameter α has been studied in (Amiri et al., 2013).

We consider the case where shape parameters α_1 and α_2 are known positive values, while rate parameters σ_1 and σ_2 are unknown positive parameters.

The complete sufficient statistics for parameters σ_1 and σ_2 are $T_X = \sum\limits_{j=1}^{n_1} X_j^{\alpha_1}$ and $T_Y = \sum\limits_{j=1}^{n_2} Y_j^{\alpha_2}$. Since X^{α_1} and Y^{α_2} have exponential distributions with rate parameters $\sigma_1^{\alpha_1}$ and $\sigma_2^{\alpha_2}$, both statistics have gamma distribution, i.e. T_X has $\Gamma(n_1, \sigma_1^{-\alpha_1})$ and T_Y has $\Gamma(n_2, \sigma_2^{-\alpha_2})$. Similarly, for $k \leq \min(n_1, n_2)$, $T_X - \sum\limits_{j=1}^{k} X_j^{\alpha_1}$ has $\Gamma(n_1 - k, \sigma_1^{-\alpha_1})$ and $T_Y - \sum\limits_{j=1}^{k} Y_j^{\alpha_2}$ has $\Gamma(n_2 - k, \sigma_2^{-\alpha_2})$. Using this and transformation of random variables $(X_1, \ldots, X_k, \sum\limits_{j=k+1}^{n_1} X_j^{\alpha_1})$ to $(X_1, \ldots, X_k, T_X)$ we get, for $\sigma_1 = 1$,

$$g(t_X|X_1 = x_1, ..., X_k = x_k) = \frac{(t_X - \sum\limits_{j=1}^{k} x_j^{\alpha_1})^{n_1-k-1}}{\Gamma(n_1 - k)} e^{-(t_X - \sum\limits_{j=1}^{k} x_j^{\alpha_1})} I\{t_X \geq \sum_{j=1}^{k} x_j^{\alpha_1}\}.$$

Using theorem 2, we get that

$$\widehat{f}(x_1, ..., x_k) = \alpha_1^k \prod_{j=1}^{k} x_j^{\alpha_1 - 1} \frac{(t_X - \sum\limits_{j=1}^{k} x_j^{\alpha_1})^{n_1-k-1} \Gamma(n_1)}{t_X^{n_1-1} \Gamma(n_1 - k)} I\{t_X \geq \sum_{j=1}^{k} x_j^{\alpha_1}\}.$$

For $k = 1$ we obtain that

$$\widehat{f}(x) = \alpha_1(n_1 - 1)x^{\alpha_1 - 1} \frac{(t_X - x^{\alpha_1})^{n_1-2}}{(t_X)^{n_1-1}} I\{t_X \geq x^{\alpha_1}\}.$$

Analogously we get that

$$\widehat{f}(y) = \alpha_2(n_2 - 1)y^{\alpha_2 - 1} \frac{(t_Y - y^{\alpha_2})^{n_2-2}}{(t_Y)^{n_2-1}} I\{t_Y \geq y^{\alpha_2}\}.$$

Denote $M = \min\{t_X^{\frac{1}{\alpha_1}}, t_Y^{\frac{1}{\alpha_2}}\}$. Using the independence of samples and the theorem 3, we obtain

$$\begin{aligned}
\widehat{R} &= \int_0^\infty \int_0^\infty I\{x < y\} \widehat{f}(x) \widehat{f}(y) dx dy \\
&= \int_0^M \frac{\alpha_1(n_1 - 1)(n_2 - 1)x^{\alpha_1 - 1}(t_X - x^{\alpha_1})^{n_1-2}}{t_X^{n_1-1} t_Y^{n_2-1}} dx \int_x^{t_Y^{\frac{1}{\alpha_2}}} \alpha_2 y^{\alpha_2 - 1}(t_Y - y^{\alpha_2})^{n_2-2} dy \\
&= \int_0^M \frac{\alpha_1(n_1 - 1)x^{\alpha_1 - 1}}{t_X^{n_1-1} t_Y^{n_2-1}} (t_X - x^{\alpha_1})^{n_1-2} (t_Y - x^{\alpha_2})^{n_2-1} dx.
\end{aligned}$$

Now applying the binomial formula we obtain that the UMVUE of R is

$$\widehat{R} = \begin{cases} \sum\limits_{r=0}^{n_1-2}\sum\limits_{s=0}^{n_2-1} \frac{(-1)^{r+s}\alpha_1(n_1-1)}{\alpha_1(r+1)+\alpha_2 s}\binom{n_1-2}{r}\binom{n_2-1}{s}\frac{T_X^{\frac{\alpha_2 s}{\alpha_1}}}{T_Y^s}, & \text{if } T_X^{\frac{1}{\alpha_1}} \le T_Y^{\frac{1}{\alpha_2}} \\ \sum\limits_{r=0}^{n_1-2}\sum\limits_{s=0}^{n_2-1} \frac{(-1)^{r+s}\alpha_1(n_1-1)}{\alpha_1(r+1)+\alpha_2 s}\binom{n_1-2}{r}\binom{n_2-1}{s}\frac{T_Y^{\frac{\alpha_1(r+1)}{\alpha_2}}}{T_X^{r+1}}, & \text{if } T_X^{\frac{1}{\alpha_1}} > T_Y^{\frac{1}{\alpha_2}}. \end{cases} \tag{3.1}$$

3.1.1 Exponential-Rayleigh model

As a special case of Weibull model we have a model where X has exponential and Y has Rayleigh distribution with densities

$$f_X(x;\alpha) = \alpha e^{-\alpha x},\ x \ge 0,$$

$$f_Y(y;\beta) = 2\beta^2 y e^{-\beta^2 y^2},\ y \ge 0,$$

where α and β are unknown positive parameters. The complete sufficient statistics for α and β are $T_X = \sum\limits_{j=1}^{n_1} X_j$ and $T_Y = \sum\limits_{j=1}^{n_2} Y_j^2$.

The UMVUE of R is

$$\widehat{R} = \begin{cases} \sum\limits_{r=0}^{n_1-2}\sum\limits_{s=0}^{n_2-1} \frac{(-1)^{r+s}(n_1-1)}{(r+1)+2s}\binom{n_1-2}{r}\binom{n_2-1}{s}\left(\frac{T_X^2}{T_Y}\right)^s, & \text{if } T_X \le \sqrt{T_Y} \\ \sum\limits_{r=0}^{n_1-2}\sum\limits_{s=0}^{n_2-1} \frac{(-1)^{r+s}(n_1-1)}{(r+1)+2s}\binom{n_1-2}{r}\binom{n_2-1}{s}\left(\frac{\sqrt{T_Y}}{T_X}\right)^{r+1}, & \text{if } T_X > \sqrt{T_Y}. \end{cases}$$

3.1.2 Numerical example

Here we present an example with real data. We wanted to compare daily wind speeds in Rotterdam and Eindhoven. We obtained two samples of 30 randomly chosen daily wind speeds (in 0.1 m/s) from the period of April 1st 2010 to April 1st 2014 taken from the website of Royal Netherlands Meteorological Institute. The first sample is from Rotterdam and the second one is from Eindhoven:

Rotterdam (X): 48, 15, 27, 18, 40, 26, 84, 19, 35, 32, 55, 29, 45, 51, 47, 66, 38, 13, 39, 28, 50, 36, 15, 74, 53, 85, 18, 58, 18, 48.

Eindhoven (Y): 44, 25, 43, 35, 20, 59, 25, 38, 26, 15, 37, 16, 35, 17, 34, 27, 40, 37, 33, 17, 51, 50, 33, 52, 25, 21, 34, 39, 23, 60.

It is well known that wind speed follows Weibull distribution. To check this we used Kolmogorov-Smirnov test. Since this test requires that the parameters may not be estimated from the testing sample, we estimated them beforehand using some other larger samples from the same populations. We got that X follows Weibull distribution with shape parameter $\alpha = 2.8$ and rate parameter $\sigma = 1/47$ (Kolmogorov-Smirnov test statistics is 0.157 and the p-value is greater than 0.1), while Y follows Weibull distribution with shape parameter $\alpha = 2.6$ and rate parameter $\sigma = 1/41$ (Kolmogorov-Smirnov test statistics is 0.158 and the p-value is greater than 0.1).

Finally, using (3.1) we estimated the probability that the daily wind speed is lower in Rotterdam than in Eindhoven to be $\widehat{r} = 0.32$.

3.2 Logarithmic distribution

Let X and Y be independent random variables from logarithmic distribution with mass functions

$$P\{X = x; p\} = \frac{-1}{\ln(1-p)} \frac{p^x}{x}, \quad x = 1, 2, ...$$
$$P\{Y = y; q\} = \frac{-1}{\ln(1-q)} \frac{q^y}{y}, \quad y = 1, 2, ...,$$

where p and q are unknown probabilities.

The logarithmic distribution has application in biology and ecology. It is often used for modeling data linked to the number of species.

The complete sufficient statistics for p and q are $T_X = \sum_{j=1}^{n_1} X_j$ and $T_Y = \sum_{j=1}^{n_2} Y_j$. The sum of n independent random variables with logarithmic distributions with the same parameter p has Stirling distribution of the first kind $SDFK(n, p)$ (Johnson et al., 2005), so T_X has $SDFK(n_1, p)$ and T_Y has $SDFK(n_2, q)$ with the following mass functions

$$P\{T_X = x; n_1, p\} = \frac{n_1!|s(x, n_1)|p^x}{x!(-\ln(1-p))^{n_1}}, \quad x = n_1, n_1 + 1, ...,$$
$$P\{T_Y = y; n_2, q\} = \frac{n_2!|s(y, n_2)|q^y}{y!(-\ln(1-q))^{n_2}}, \quad y = n_2, n_2 + 1, ...,$$

where $s(x, n)$ is Stirling number of the first kind.

An unbiased estimator for R is $I\{X_1 < Y_1\}$. Since

$$E(I\{X_1 < Y_1\}|T_X = t_X, T_Y = t_Y) = \frac{P\{X_1 < Y_1, T_X = t_X, T_Y = t_Y\}}{P\{T_X = t_X, T_Y = t_Y\}}$$
$$= \frac{\sum_{x=1}^{M} \sum_{y=x+1}^{t_Y - n_2 + 1} P\{X_1 = x\}P\{Y_1 = y\}P\{\sum_{k=2}^{n_1} X_k = t_X - x\}P\{\sum_{l=2}^{n_2} Y_l = t_Y - y\}}{P\{T_X = t_X\}P\{T_Y = t_Y\}}$$
$$= \sum_{x=1}^{M} \sum_{y=x+1}^{t_Y - n_2 + 1} \frac{t_X! t_Y! |s(t_X - x, n_1 - 1)||s(t_Y - y, n_2 - 1)|}{n_1 n_2 (t_X - x)!(t_Y - y)! xy |s(t_X, n_1)||s(t_Y, n_2)|},$$

where $M = \min\{t_X - n_1 + 1, t_Y - n_2\}$, using theorem 1 we get that the UMVUE of R is

$$\widehat{R} = \sum_{x=1}^{\min\{T_X - n_1 + 1, T_Y - n_2\}} \sum_{y=x+1}^{T_Y - n_2 + 1} \frac{T_X! T_Y! |s(T_X - x, n_1 - 1)||s(T_Y - y, n_2 - 1)|}{n_1 n_2 (T_X - x)!(T_Y - y)! xy |s(T_X, n_1)||s(T_Y, n_2)|}.$$

4 Conclusion

In this paper we considered the unbiased estimation of the probability $P\{X < Y\}$ when X and Y are two independent random variables. Some known results of UMVUEs for R for some distributions were listed. Two new cases were presented, namely Weibull model with known but different shape parameters and unknown rate parameters and Logarithmic model with unknown parameters. An example using real data was provided.

References

[1] Amiri, N., Azimi, R., Yaghmaei, F. and Babanezhad, M. (2013): Estimation of Stress-Strength Parameter for Two-Parameter Weibull Distribution. *International Journal of Advanced Statistics and Probability*, **1**(1), 4-8.

[2] Belyaev, Y. and Lumelskii, Y. (1988): Multidimensional Poisson Walks. *Journal of Mathematical Sciences*, **40**, 162-165.

[3] Birnbaum, Z.W. (1956): On a Use of Mann-Whitney Statistics. *Proc. Third Berkeley Symp. in Math. Statist. Probab.*, **1**, 13-17. Berkeley, CA: University of California Press.

[4] Constantine, K., Carson, M. and Tse, S-K. (1986): Estimators of $P(Y < X)$ in the gamma case. *Communications in Statistics - Simulations and Computations*, **15**, 365-388.

[5] Downtown, F. (1973): On the Estimation of $Pr(Y < X)$ in the Normal Case. *Technometrics*, **15**, 551-558.

[6] Hogg, R.V., McKean, J.W. and Craig, A.T. (2005): *Introduction to Mathematical Statistics*, Sixth Edition. 348-349. New Jersey: Pearson Prentice Hall.

[7] Ivshin, V.V. and Lumelskii, Ya.P. (1995): *Statistical Estimation Problems in "Stress-Strength" Models*. Perm, Russia: Perm University Press.

[8] Johnson, N.L., Kemp, A.W. and Kotz, S. (2005): *Univariate Discrete Distributions*. New Jersey: John Wiley & Sons.

[9] Kotz, S., Lumelskii, Y. and Pensky, M. (2003): *The Stress-Strength Model and its Generalizations*. Singapore: World Scientific Press.

[10] Kundu, D. and Gupta, R.D. (2006): Estimation of $P[Y < X]$ for Weibull Distributions. *IEEE Transactions on Reliability*, **55**(2).

[11] Rezaei, S., Tahmasbi, R. and Mahmoodi, M. (2010): Estimation of $P[X < Y]$ for Generalized Pareto Distribution. *Journal of Statistical Planning and Inference*, **140**, 480-494.

[12] Saracoglu, B., Kaya, M.F. and Abd-Elfattah, A.M. (2009): Comparison of Estimators for Stress-Strength Reliability in the Gompertz Case. *Hacettepe Journal of Mathematics and Statistics*, **38**(3), 339-349.

[13] Tong, H. (1974): A Note on the Estimation of $P(Y < X)$ in the Exponential Case. *Technometrics*, **16**, 625. Errata: *Technometrics*, **17**, 395.

[14] Tong, H. (1977): On the Estimation of $P(Y < X)$ for Exponential Families. *IEEE Transactions on Reliability*, **26**, 54-56.

[15] *Website of Royal Netherlands Meteorological Institute*, downloaded from http://www.knmi.nl on April 12 th 2014.

Metodološki zvezki, Vol. 11, No. 1, 2014, 31-63

Testing Two Theories for Generating Signed Networks Using Real Data

Patrick Doreian[1] and Andrej Mrvar[2]

Abstract

Multiple social processes generate social network structures. We use relaxed structural balance, a generalization of classic structural balance, to facilitate a direct comparative test of two social psychological theories regarding network generation. One is structural balance theory. The other concerns differential popularity. These theories predict distinctive signed blockmodels. We use two well known empirical temporal signed data sets presenting an opportunity for comparing the two theories in terms of their predictions about blockmodel representations of these networks. The results provide strong support for differential popularity, differential disliking, and mutual disliking within a subset of actors. While there is evidence that structural balance was also operating, it seems the lesser process for the data used in these tests. We also examine the unequal distributions of receiving positive and negative ties. Both tend to become more unequal over time. Suggestions for future research are provided.

1 Introduction

Both social psychologists and social network analysts develop theories intended to help understand social processes in small social groups. To the extent that the former focus more on node-level (actor) characteristics while the latter are more attentive to the network structure as a whole, there is a tension between micro-level and macro-level phenomena (Robins and Kashima, 2008). Our focus here is on *understanding processes that generate network structures*. We provide comparative tests of two theories based on a simple assumption: social processes, if operative in small groups, leave traces of recognizable patterns of network ties. This comparative test is for *signed* networks. Our primary goal is disentangling the results from the operation of processes specified by two theories of social processes in groups. One is structural balance theory Heider (1946, 1958) The other concerns

[1] Faculty of Social Sciences, University of Ljubljana and Department of Sociology, University of Pittsburgh, 2602 WWPH, Pittsburgh, PA 15260; pitpat@pitt.edu
[2] Faculty of Social Sciences, University of Ljubljana, Kardeljeva pl. 5, 1000 Ljubljana, Slovenia; Andrej.Mrvar@fdv.uni-lj.si

differential popularity, a process described by Feld and Elsmore (1984) under which some group members receive more positive ties than others. The detailed predictions of the two theories differ.

As Taylor (1970) notes, Heider was credited with the initial statement of structural balance theory. While we focus attention on the Heider variant of consistency theories, Newcomb (1961), Festinger (1957), Osgood and Tannenbaum (1955) and others (see Abelson *et al.*, 1968) also formulated alternative consistency theories. We use Heider's approach because Cartwright and Harary's (1956) formal generalization of his theory laid formal foundations for analyzing signed social networks.

Feld and Elsmore (1984) drew a critical response from Hallinan (1984) regarding rival processes accounting for the unequal distributions in the receipt of signed ties in a group. Both papers considered rival theories about group processes by using distributions of particular triples of ties among trios of actors in the network of actors in the group.

Rather than use distributions of triple types, we examine the *overall structure* of a network using blocks located in signed blockmodels. Briefly, a blockmodel of a network is a simultaneous partition of both the actors and their social ties. The clusters of actors are called positions[3]. Using blockmodels delineating network structure provides an direct description of a network's overall structure.

The rest of this paper is organized as follows. Section 2 outlines substantive issues and Section 3 describes our data and methods. Section 4 presents our results and we conclude with a summary and discussion in Section 5.

2 Theories about processes that generate network structures

2.1 Structural balance theory

The intuitions of Heider's (1946) structural balance theory, formalized by Cartwright and Harary (1956), led to a sustained research effort of discerning the structure of signed networks (Doreian *et al.*, 2005: Chapter 10). Key in this development was a remarkable 'structure theorem' coupling *micro-processes* (of actors forming and/or dropping signed ties) and the resulting *macro-structure* of the group. Signed ties are either positive (e.g. liking, loving, supporting) or negative (e.g. disliking, hating, opposing). For three actors, denoted by *p*, *o* and *q*, in a signed network, the *poq* triple is made up of the ties ($p \rightarrow q$), ($q \rightarrow o$) and (p→o). The sign of every triple is the

[3] A formal statement can be found in Doreian *et al.* (2005). Ferligoj et al. (2011) contains a rigorous informal statement about positional analysis in terms of positions and roles.

product of its signed relations. A *poq*-triple is balanced if its sign is positive and imbalanced if the sign is negative[4]. There are four possible balanced triples and four imbalanced triples. A signed network is balanced if all of its *poq*-triples are balanced. Cartwright and Harary's main theorem states: the vertices of a balanced network can be partitioned into two positions where all of the positive ties are within positions and all of the negative ties are between members of different positions. This result links the micro-processes of tie formation and change within triads to a statement about the overall group structure for balanced networks. Davis (1967) noted human groups often have more than two mutually hostile subgroups. He generalized Cartwright and Harary's result by reconsidering one part of Heider's foundational statement: if all of the ties in a *poq*-triple were negative, the triple was imbalanced. Davis defined this all-negative triple as balanced. His result was: a 'clusterable' network[5] has two *or more* positions where all the positive ties were within clusters and all of the negative ties were between actors in different positions. This also links micro-processes to the macro-structure of a group. A signed network is *k*-balanced if it has the above partition structure. For $k=2$ it is Cartwright and Harary's structure theorem. For $k > 2$ it is the generalization.

Blockmodeling (see Breiger *et al.*, 1975; Doreian *et al.*, 2005) has techniques for partitioning network data into positions (containing actors) and blocks (of ties between positions). The *location* of an actor is the set of ties to and from all other actors in the group. These locations of actors are clustered to form the positions. For *n* actors, the *n* locations are partitioned into *k* positions with *k* is much smaller than *n*. A large network is reduced to a smaller image matrix with *k* positions and k^2 blocks representing the essential network structure. Doreian and Mrvar (1996) noticed the theorems of Cartwright and Harary (1956) and Davis (1967) can be viewed as leading to statements of specific blockmodels. A *positive block* is one having only positive ties and null ties while a *negative block* has only negative ties and null ties. From the structure theorems, in a *k*-balanced network, the signed blockmodel has positive blocks on the main diagonal (top left to bottom right) and negative blocks off the main diagonal. If, for example, $k=4$ and structural balance is the only process operating, then the blockmodel implied by structural balance is simple to describe. The block pattern for four positions is:

[4] This is expressed in folk aphorisms: "a friend of a friend is a friend", "a friend of an enemy is an enemy", "an enemy of a friend is an enemy" and "an enemy of an enemy is a friend". These have simple cognitive structures. As Mower White (1979) notes, simple cognitive structures are more likely than complex structures to exhibit balance. Also, "it is now recognized that if sentiment is restricted to the two values of positive and negative, balance is a simple implication of ordinary deductive logic (Montoya and Insko, 2008: 494)".

[5] To prove this theorem, Davis used the concept of a semiwalk, an alternating sequence of vertices and arcs where the direction of the arcs is irrelevant. For pairs of actors between whom there exist one or more semiwalks, the sign for each of these semiwalks is the product of the signs of the arcs in the semiwalk. These signs are positive or negative. He defined a network as 'clusterable' if it had no semiwalks with a single negative arc.

Positive	Negative	Negative	Negative
Negative	Positive	Negative	Negative
Negative	Negative	Positive	Negative
Negative	Negative	Negative	Positive

We refer to these as ideal blocks by location, call this blockmodel the *Structural Balance blockmodel*, and label it the ‘SB Model’.

Regardless of the number of positions, every blockmodel predicted by structural balance has this generic (ideal) SB Model form. The number of positions, *k*, has to be determined as a part of fitting blockmodels. Empirically, it is unreasonable to expect a perfect correspondence between an ideal structure and an empirical structure. If structural balance is appropriate we would anticipate the SB Model but with some inconsistencies compared to the ideal structure.

Doreian and Mrvar (1996) took the form of the idealized blockmodels implied by structural balance and proposed a partitioning approach for establishing empirical blockmodel structure(s) of signed networks closest to the ideal form implied by the structural theorems. When empirical blockmodels do not fit exactly there are some inconsistencies between the empirical blockmodel and the ideal counterpart. These will take the form of some negative ties in positive blocks and some positive ties in negative blocks. The former are termed negative inconsistencies, the latter are positive inconsistencies. For a binary network (where the ties are +1 or -1), the total number of positive inconsistencies is denoted by $\mathcal{P}$ and the total number of negative inconsistencies[6] by $\mathcal{N}$. A general measure of how poorly a blockmodel fits the data is given by $C_f = \alpha\mathcal{N} + (1 - \alpha)\mathcal{P}$ where[7] $0 < \alpha < 1$. With $\alpha = 0.5$, the two types of inconsistencies are weighted equally, a convention we use here. In essence, C_f is the line index of imbalance proposed by Harary et al. (1965: 348-350). C_f is a criterion function and the relocation clustering algorithm used by Doreian and Mrvar seeks optimal partition(s) minimizing this criterion function[8]. Structural balance implies an SB Model.

2.2 Differential popularity

In the main, social scientists collecting sociometric data focused on unsigned data with only positive ties. Undoubtedly, such data are easier to collect. Also, one rationale for making comparisons of the distribution of triples in unsigned

[6] If a network has weighted ties then P and N, respectively, are the *sums* of positive and negative inconsistencies.

[7] For α=1, positive inconsistencies are ignored and negative inconsistencies are ignored for α=0. Neither extreme weighting is useful when *both* positive and negative ties exist.

[8] It is a local optimization method so finding the optimal partition(s) is not guaranteed. Brusco *et al*. (2011) established this algorithm has, thus far, identified all of the optimal partitions for signed networks up to 40 actors.

networks, as used by Feld and Elsmore (1984) and by Hallinan (1984), is based on arguments of Davis and Leinhardt (1972) where signed graphs are 'converted' to unsigned counterparts. Rather than focus on signed ties (positive, null, and negative), attention was focused on mutual (M), null (N) and asymmetric (A) ties. Identifying clusters of positively connected actors, such as those among the positions of signed networks, was treated as evidence of a tendency towards clustering. Comparisons were then made of the distributions of the 16 possible triples involving M, A, and N ties. However, as using unsigned data handicaps any examination of balance theoretic ideas about signed networks, these efforts labored under a serious constraint: negative ties were excluded[9]. Feld and Elsmore (1984) focused primarily on transitivity. If $(p \rightarrow o)$ and $(o \rightarrow q)$ are present in an unsigned network then, under transitivity, the $(p \rightarrow q)$ tie will be present also. Empirically, there is a tendency towards transitivity in most unsigned networks with transitivity has regarded as a fundamental network process (Holland and Leinhardt, 1972; Wasserman and Faust, 1994: 243-248). Confirmation came with there being more transitive triples in a network than would be expected by chance. One key feature of Feld and Elsmore's argument is that some of the evidence for transitivity might be due to the operation of a process of differential popularity[10]. They provided some evidence in the form of distributions of *poq*-triples to support this claim. However, they were careful to *not* state differential popularity dominated transitivity. They suggested it could be a plausible generating process, one also creating some transitivity. In neutral terminology, transitivity and differential popularity are often confounded in empirical networks. When only one of them is considered, some of the support for it as *the* generating process will be spurious.

The idea of differential popularity extends straightforwardly to signed networks: some actors may be more popular and so receive more positive ties regardless of the presence of mutually hostile subgroups. If some members of a group are universally popular, then with *k=4,* the group structure, as a blockmodel, would be as follows if there were just two processes - structural balance and differential popularity – operating. An ideal blockmodel would look like:

Positive	Negative	Negative	Negative
Positive	Positive	Negative	Negative
Positive	Negative	Positive	Negative
Positive	Negative	Negative	Positive

[9] We do not dispute the value of the highly productive work on triadic censuses for unsigned networks and their extension to exponential random graph models. But when structural balance is involved, we contend that both positive and negative ties *must* be included.

[10] For example, given $p \rightarrow o$ and $o \rightarrow q$ as positive ties, if $p \rightarrow q$ exists then it can be viewed as being consistent with transitivity. It is consistent also with structural balance in a positive triple.

Note the column of positive blocks on the left of this ideal blockmodel. Except for the top block, *all* of the positive blocks in the first column are inconsistent with structural balance (and are bolded for this reason). We call this ideal blockmodel a *Structural Balance with Differential Popularity blockmodel* and label it the SB_DP Model. If some additional actors are popular but not universally popular, an ideal blockmodel would look like:

Positive	Negative	Negative	Negative
Positive	Positive	Negative	Negative
Positive	**Positive**	Positive	Negative
Positive	**Positive**	Negative	Positive

The additional bolded blocks (in the second column of blocks) are also inconsistent with structural balance but consistent with differential popularity. This blockmodel is a variant of the SB_DP Model. There may be less extreme configurations where only some blocks in the left hand column are positive. There could be other subgroups receiving positive ties from members of other positions. These can be accommodated. For now, we focus on the SB_DP Model in our comparative tests.

Discriminating between these two theories can be done in a direct fashion. If structural balance operates, then the SB Model is appropriate. Further, if differential popularity is not operative, the SB Model would fit the data and not the SB_DP Model. But if the SB_DP Model is identified empirically, greater credibility is given to differential popularity. The partitioning algorithm of Doreian and Mrvar (1996) is useless for this comparative test: a SB Model is the only fitted blockmodel. However, thinking in terms of relaxing structural balance (Doreian and Mrvar, 2009) led to the creation of an algorithm appropriate for distinguishing these two models.

2.3 Relaxed structural balance

In responding to Feld and Elsmore (1984), Hallinan (1984) argued at least five substantive processes could generate transitivity in unsigned networks: differential expansiveness; reciprocity; differential popularity; clustering and cognitive (structural) balance. Although we do not focus on transitivity and consider signed networks, we accept the point of analyses of network data requiring recognition, and consideration, of multiple processes. Incorporating them for signed networks, when considering balance theoretic ideas, requires a generalization of structural balance. Reciprocal positive ties can be accommodated easily to the extent that they occur among within actors in the same position. But, if there is positive reciprocity between pairs of actors in different positions, this creates problems for structural balance: positive inconsistencies contribute to C_f. If this involves multiple pairs in two positions there will be corresponding positive blocks above and below the main diagonal. If there is reciprocation of negative ties between actors in different positions this will be consistent with structural balance. However, we need to consider *subsets* of actors who, as individuals, are mutually hostile towards each other. Their presence also contradicts structural balance because this implies a negative diagonal block[11]. If we add mutual dislike at the actor level for a set of actors – a "nest of vipers" in the colorful terminology of Hummert *et al*. (1990) – to differential popularity and structural balance then we would expect a structure approximating the following blockmodel:

Positive	Negative	Negative	Negative
Positive	Positive	Negative	Negative
Positive	Negative	Positive	Negative
Positive	Negative	Negative	**Negative**

Locating the diagonal negative block on the bottom right of the blockmodel appears arbitrary. But if there is a differential popularity process then it is reasonable to anticipate differential disliking implies negative ties are concentrated actors other than popular actors[12]. This is represented by a column of off-diagonal negative blocks on the right of this blockmodel. Further, if those that are more disliked also tend to dislike each other this implies a diagonal negative block. To capture this, we locate (and bold) a diagonal negative block at the bottom right hand side while recognizing that there could be more than one such block and they could appear anywhere on the diagonal. The column of off-diagonal negative blocks on the right is consistent with *both* structural balance and differential dislike. The negative diagonal block is inconsistent with structural balance. We call this a *Structural Balance with Differential Popularity and Mutual Dislike blockmodel* and denote it as an SB_DP_MD Model.

To deal with these and other potential complications - including mediation - Doreian and Mrvar (2009) proposed 'relaxed structural balance' as a more general model for signed networks. Having only positive blocks and negative blocks was retained. However, they were allowed to appear *anywhere* in a blockmodel. Relaxed structural balance is a formal generalization of the structural balance. The criterion function, C_f, as described above and the relocation algorithm were retained for fitting relaxed structural balance models to network data. All that

[11] This pattern is present in Figure 2 and this prompted the notion of diagonal negative blocks.

[12] One mechanism is disliked attributes of some actors take time to be recognized more widely in a group.

changed under relaxed structural balance is the potential locations of the signed blocks. Relaxed structural balance permits the statement of another set of ideal blockmodels.

In partial summary, the first two primary substantive hypotheses are stated in a comparative form.

Hypothesis 1 If differential popularity operates for positive ties, there will be a column of positive blocks for the more popular actors and this tendency will increase through time[13]. If structural balance dominates differential popularity then there will be no positive off-diagonal blocks in a column corresponding to universally popular actors. Nor would there be positive off-diagonal blocks for other popular actors.

Hypothesis 2 If differential dislike is operative, there will be a column of negative blocks for the more disliked actors and this tendency will increase though time. In particular, there will be at least one diagonal negative block. If structural balance dominates then there will be no diagonal negative blocks.

Heider’s theory is essentially dynamic with actors striving to reduce inconsistencies. This is expressed as a tendency towards balance over time. Indeed, data for examining Heider’s theory *must* be temporal. However, all Heider’s imbalanced triples can be balanced in three ways. Alas, Heider was silent on how balance is achieved. It requires complex temporal processes in human groups (Hummon and Doreian, 2003). If differential popularity and differential dislike accumulate over time, this suggests:

Hypothesis 3 Increasing tendencies of differential popularity and differential dislike will create greater inequality on the receipt of *both* positive and negative ties over time.

The idea of moving towards certain structural forms stems from Heider's notion of tendencies towards balance being extended to relaxed structural balance. The concentration of both positive and negative ties (Hypothesis 3) could be the result of two social mechanisms. One is an individual level process where attributes making people popular (liked) or unpopular (disliked) are recognized more over time. The other is found in the idea of actors achieving consistency of views of people as driven by balance. Of course, this leaves open the issue of which of these processes are operative or the extent to which they are both operative. The data at our disposal do not permit an exploration of this issue. Even so, relaxed structural balance incorporates additional processes beyond structural balance.

The tests that we propose are facilitated by using the *same criterion function for all fitted models*. Relaxed structural balance models have structural balance as a special case. If structural balance dominates all other processes then the SB

[13] We allow less extreme versions with some actors more popular but not universally popular as shown in the one variant of the SB_DP Model. Positive valued actor attributes may also take time to be perceived widely.

Model will be identified implying structural balance is the generating process. But if both structural balance and differential popularity are operating without mutual dislike then a variant of the SB_DP Model will fit the signed data better. And if there is also mutual dislike in subgroups, the SB_DP_MD Model will fit. If any of the more general models within relaxed structural balance fit, there is evidence against structural balance being the sole, or even the main, generating process. Classic structural balance and relaxed structural balance partitions are rivals to be evaluated comparatively. They can be compared through their blockmodel signatures.

3 Data and methods

Brusco et al. (2011), based on Leik and Meeker (1975), argue it is more fruitful to have substance, data, and model (with the methods it implies) form a coherent whole. We achieve this here within the rubric of balance theoretic ideas. The SB Model and relaxed structural balance (RSB) models can be evaluated comparatively. Group trajectories towards balance, if they exist, need not imply strictly monotonic change in the level of imbalance. But there will be some overall movement in this direction over time. Given this empirical claim of Heider, it is necessary to examine signed structures over time using blockmodel structures. Given substance drove the hypotheses and the methods of relaxed structural balance are fully consistent with this, the coherence of Leik and Meeker’s substance-method-data triple is preserved.

Alas, there are few signed networks over enough time points to test Heider's theory. *We know of only two such data sets*. One is Newcomb’s (1961) data as recorded by Nordlie (1957). The other comes from Sampson’s (1968) study of trainee monks in a monastery. Neither data set is ideal. Newcomb collected network data from 17 students in a pseudo-fraternity. In partial exchange for room and board, these *previously unacquainted* students provided sociometric data for 15 time points over a semester. The strength of Newcomb’s study is the network formation process started from an initial state of no network ties. The recorded data were in the form of ranks with each actor ranking all of the other actors in terms of liking. Doreian *et al.* (1996) recoded these recorded ranked sociometric ties into a signed form. With this recoding, they established reciprocity, transitivity and structural balance had different time scales. The top four ranks were converted to +1 and the bottom three ranks were recoded to -1. The remaining ties were recoded as zero[14]. We use their (four positive ties and three negative ties) coding scheme here. Of course, as noted by Hallinan (1984) drawing on the arguments of Holland

[14] Their reasons for this coding and the formal methods for establishing it are found in their article. With regard to structural balance, other recoding options in terms of the number of positive and negative ties were tried without leading to substantively different results.

and Leinhardt (1973), there are problems with fixed choice designs. However, as we want our results to be comparable with prior analyses of the Newcomb data we used this coding.

Doreian *et al.* (1996) computed the imbalance over time for the recoded Newcomb data and showed a general decline over time. While this decline was not strictly monotonic, there was enough support for Heider's empirical hypothesis[15]. However, if the relaxed structural balance model is a better model, one that allows for multiple processes, then imbalance for relaxed structural balance will decline over time. More importantly, imbalance will be lower at each time point than for structural balance. To examine Hypothesis 3, we use Theil's (1967: 92) entropy index, as a measure of inequality, for receiving positive and negative ties at each time point.

The criterion function C_f can be viewed as merely descriptive and lacking tests of its utility for partitions established when using it. To address this, we use quadratic assignment regression, QAP, as formulated by Dekker et al. (2007) and implemented in Borgatti et al. (2002), to make statistical assessments of established signed blockmodels. The ideal blockmodels specify (by locations) the presence of positive and negative blocks. Given an established blockmodel (with inconsistencies), we can define the 'fitted' blockmodel that corresponds to the empirical blockmodel. In the following panel we show, on the left, a hypothetical pair of positive and negative blocks with some (bolded) inconsistencies. The c orresponding pair of 'predictions' implied by the blocks in an ideal blockmodel[16] are on the right.

A positive block (with inconsistencies) The corresponding fitted positive block

$$\begin{array}{cccccccc}
0 & \mathbf{-1} & 1 & 1 & 0 & 0 & \mathbf{-1} & 0 \\
1 & 1 & 0 & 1 & \mathbf{-1} & 0 & 1 & \mathbf{-1} \\
0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\
1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\
1 & 0 & \mathbf{-1} & 0 & 0 & 0 & 0 & \mathbf{-1}
\end{array}
\qquad
\begin{array}{cccccccc}
0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 \\
1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\
0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\
1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 1
\end{array}$$

[15] We emphasize the term 'enough support'. In a follow-up study using the Newcomb data, Doreian and Krackhardt (2001) showed that the incidence of *two of the imbalanced triples increased over time while the number of two of the balanced triples declined over time*.

[16] Borgatti and Everett (1999) propose using Pearsonian correlations in a similar fashion but with a crucial difference. Their ideal blocks are either complete or null. The latter are unproblematic but we differ here by 'predicting' only the implied value of a tie when there is an empirical tie in the data.

A negative block (with inconsistencies)

$$\begin{array}{cccccccc}
-1 & 0 & \mathbf{1} & –1 & 0 & 0 & 0 & -1 \\
\mathbf{1} & -1 & 0 & 0 & 0 & -1 & -1 & \mathbf{1} \\
-1 & 0 & -1 & 0 & 0 & 0 & -1 & -1 \\
0 & 0 & -1 & \mathbf{1} & \mathbf{1} & -1 & -1 & 0 \\
-1 & 0 & \mathbf{1} & 0 & 0 & -1 & 0 & 0
\end{array}$$

The corresponding fitted negative block

$$\begin{array}{cccccccc}
-1 & 0 & -1 & -1 & 0 & 0 & 0 & -1 \\
-1 & -1 & 0 & 0 & 0 & -1 & -1 & -1 \\
-1 & 0 & -1 & 0 & 0 & 0 & -1 & -1 \\
0 & 0 & -1 & -1 & -1 & -1 & -1 & 0 \\
-1 & 0 & -1 & 0 & 0 & -1 & 0 & 0
\end{array}$$

An empirical network with blocks and the fitted blockmodel can be compared by using QAP to assess the fit. QAP is used to ‘compare’ two whole matrix arrays to examine the extent to which they are the same or consistent with each other. In these analyses, the fitted blockmodel is used to predict the empirical data. If the correlations between the two are significant, the fitted blockmodel passes a test in terms of empirical adequacy. However, if the fit is poor, the blockmodel fails. It is possible also to compare the fitted blockmodel with a random partition as a secondary way of assessing the adequacy of its fit. We did this using the Adjusted Rand Index (ARI) and evaluative criteria put forth by Steinley (2004). He argues ARI values above 0.9 indicate an excellent correspondence in the composition of a pair of partitions; values above 0.8 suggest an acceptable correspondence and values below 0.8 are unacceptable.

Another potential problem with blockmodeling is finding multiple optimal partitions for a given value of *k*. If all have the same block structure, and attention is focused solely on the block structure, this is not a huge problem. But, if there are multiple ‘best’ partitions, having different block structures, this is a serious problem. A third potential problem is the presence of null blocks: they must be identified. For structural equivalence, only two block types are possible: complete blocks and null blocks. Differential penalties can be imposed on the two types of inconsistencies (ones in null blocks and null ties in complete blocks). Doreian et al. (2004), for partitioning two-mode data, imposed a heavy penalty on the former inconsistency to ensure null blocks appeared as fully null blocks[17].

For the Newcomb data, there are null blocks. Specifying a null block helps eliminate multiple equally well fitting partitions under relaxed balance. We used the algorithm of Doreian and Mrvar (2009) as implemented in pajek (Batagelj and Mrvar, 1998) for each time point in an inductive fashion with one null block specified. Having identified the ‘best’ partition structures for *k=4* inductively, we then, for each time point, pre-specified its delineated block structure in a deductive

[17] They used pre-specification but here only the presence of a null block was allowed.

fashion (with many repetitions) to make sure there were no additional partitions with the identified partition structure[18].

When comparing relaxed structural balance with structural balance we thought differential popularity would be important and, perhaps, dominate structural balance. The comparisons had to be fair. A crucial difference exists in the behavior of C_f as the number of clusters (k) increases fot structural balance and relaxed structural balance. For the former, the curve of the criterion function, C_f, when plotted against k, has a U-shape with a guaranteed minimum value (Doreian et al., 2005: Theorem 10.6). In contrast, for relaxed structural balance, C_f decreases monotonically with k (Doreian and Mrvar, 2009: Theorem 4). We chose k=4 primarily because the 'best' structural balance results were for this value of k. Increasing the value of k beyond 4 has two implications: i) values of C_f increase for structural balance while they decrease for relaxed structural balance. This creates a bias favoring the latter for higher values of k. For a fair comparative test we used the same value of k for relaxed structural balance and structural balance. If anything, this favored structural balance. At most time points, the optimal partition for structural balance occurs for $k=4$ in the Newcomb data. For the Sampson data, it is $k=3$ at all three time points. We then compared the fitted models with each other[19].

4 Empirical results

4.1 Using the Newcomb data

Figure 1 shows the criterion function values for $k=4$ over time for structural and relaxed balance. Both trajectories decrease overall. The values of the criterion function for relaxed balance are always lower than for structural balance, implying the RSB model fits the data better than the SB model. While this has little surprise value, it emphasizes limitations to structural balance. For each time point, we computed the ARI for pairs of partitions obtained from the two models. They ranged from 0.073 to 0.689. For each time point, the partitions obtained from the two approaches are not the same. Most often, they are not even close.

[18] In fitting blockmodels to signed networks where null blocks are specified, the criterion function $C_f = \alpha\mathcal{N} + (1 - \alpha)\mathcal{P}$ was modified by including a term for the null block that ensured that the null block would be as large as possible. (Small null blocks were penalized relative to larger null blocks so larger null blocks were identified.)

[19] For Sampson data we consider also $k=4$ for relaxed structural balance.

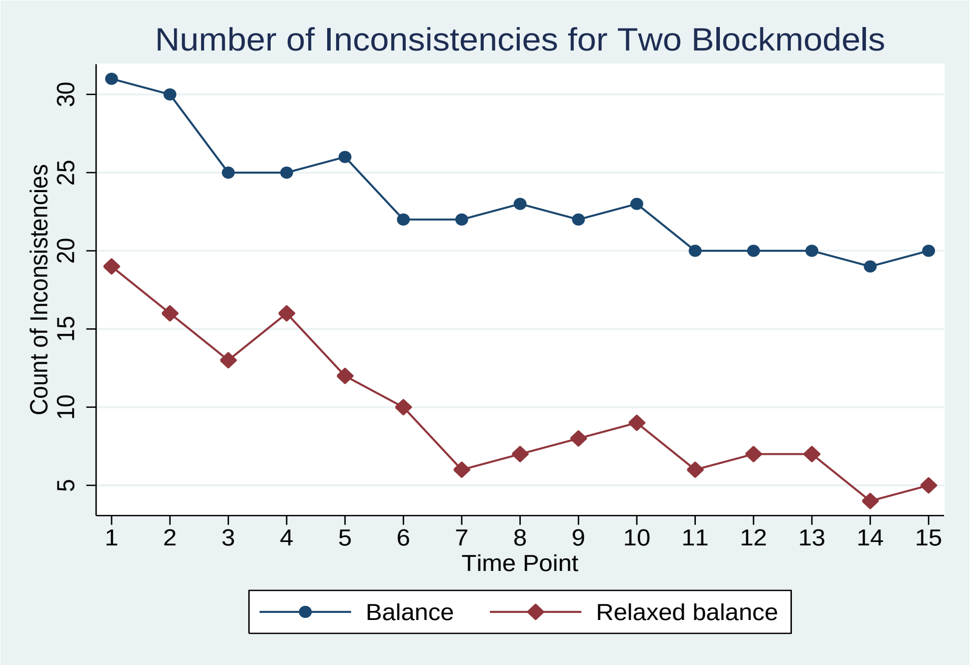

Figure 1: Inconsistency counts for the Structural Balance and the Relaxed Structural Balance models: Newcomb data.

There are additional issues in fitting blockmodels to network data meriting attention. The first concerns the predictive value of the fitted blockmodels. We computed the correlation, for the 15 time points labeled t_1 through t_{15}, between these QAP correlations and the value of the criterion function, C_f. The value of this correlation is -0.959 ($p < .0001$) indicating an very close correspondence between the two set of values. Table 1 provides the numerical values and the QAP correlations for both relaxed structural balance and structural balance. The QAP correlations in Table 1, using a permutation test, act as a close proxy for a permutation test for the criterion function. The p-values[20] for the QAP correlations are all less than 0.001. The values for structural balance have a similar temporal pattern but the correlation between the QAP correlations and the criterion function is slightly less. Even so, the lower QAP correlations for structural balance suggest poorer predictive performances consistent with the values of the criterion function for the two rival models.

[20] Most correlations are 'significant' which may be an inherent feature of QAP. However, our use of QAP is driven primarily by a need to compare the results from using relaxed structural balance and structural balance. It is unlikely that a bias towards significance affects the comparative results differently. Also there are non-significant QAP estimates in the results we report.

Table 1: QAP correlations and criterion function values: empirical and fitted blockmodels: Newcomb data.

T	RSB QAP Correlations*	RSB Criterion Function Values (C_f)	SB QAP Correlations*	SB Criterion Function Values (C_f)
t_1	0.679	9.5	0.499	15.5
t_2	0.740	8.0	0.502	15.0
t_3	0.779	6.5	0.588	12.5
t_4	0.752	8.0	0.598	12.5
t_5	0.810	6.0	0.579	13.0
t_6	0.754	5.0	0.511	11.0
t_7	0.911	3.0	0.633	11.0
t_8	0.881	3.5	0.619	11.5
t_9	0.865	4.0	0.633	11.0
t_{10}	0.860	4.5	0.617	11.5
t_{11}	0.899	3.0	0.674	10.0
t_{12}	0.898	3.0	0.669	10.0
t_{13}	0.881	3.5	0.671	10.0
t_{14}	0.932	2.0	0.687	9.5
t_{15}	0.915	2.5	0.669	10.0

RSB Relaxed Structural Balance; SB Structural Balance

* All p-values < 0.001. The correlation between QAP correlations and C_f is -0.959 for RSB and -0.858 for SB.

Table 2 presents the results of using QAP regressions comparing the predictive values of RSB and SB. Reading from the right, it appears both the fitted SB and the fitted RSB blockmodels have some predictive value. Further, the predictive value for each, roughly, increases through time. However, when the fitted SB blockmodel is included as a predictor with the fitted RSB blockmodel it seldom increases the predictive value of the QAP regression. Of course, when two predictors are correlated there is no unique partition of the variance explained between them. However, we note the following additional items in Table 2: i) the estimated intercept is near zero for each time point; ii) the *unstandardized* coefficients are such that the coefficients for RSB are always larger than the corresponding coefficients for SB[21]; iii) over time, the unstandardized coefficient for SB declines while the unstandardized coefficients for RSB increase; and iv) at each time point, the standardized coefficient for RSB is larger than the standardized coefficient for SB indicating it as the more potent predictor. In short, the fitted RSB blockmodel has superior predictive value than the fitted SB blockmodel.

[21] The two fitted blockmodels have the same density so there is not an issue of different scales inflating one coefficient relative to the other.

Table 2: QAP Regressions comparing Relaxed Structural Balance and Structural Balance: Newcomb data.

T.	Variable	Unstandardized Coefficient	Standardized Coefficient	p-value	R^2	R^2 (for RSB)	R^2 (for SB)
t_1	Intercept	0.051	0.000	-	0.47	0.46	0.25
	SB	0.134	0.133	0.0140			
	RSB	0.596	0.598	0.0005			
t_2	Intercept	0.015	0.000	-	0.58	0.55	0.25
	SB	0.201	0.202	0.0005			
	RSB	0.651	0.646	0.0005			
t_3	Intercept	0.013	0.000	-	0.64	0.61	0.35
	SB	0.208	0.209	0.0005			
	RSB	0.662	0.659	0.0005			
t_4	Intercept	0.042	0.000	-	0.62	0.57	0.36
	SB	0.285	0.284	0.0005			
	RSB	0.610	0.604	0.0005			
t_5	Intercept	0.041	0.000	-	0.66	0.66	0.34
	SB	0.089	0.089	0.0265			
	RSB	0.753	0.752	0.0005			
t_6	Intercept	0.010	0.000	-	0.57	0.57	0.26
	SB	0.085	0.085	0.0365			
	RSB	0.704	0.702	0.0005			
t_7	Intercept	-0.008	0.000	-	0.83	0.83	0.40
	SB	0.076	0.077	0.0100			
	RSB	0.868	0.861	0.0005			
t_8	Intercept	0.004	0.000	-	0.78	0.78	0.38
	SB	0.051	0.051	0.0880			
	RSB	0.848	0.847	0.0005			
t_9	Intercept	-0.020	0.000	-	0.77	0.75	0.40
	SB	0.172	0.173	0.0005			
	RSB	0.767	0.761	0.0005			
t_{10}	Intercept	0.028	0.000	-	0.75	0.74	0.38
	SB	0.108	0.108	0.0040			
	RSB	0.792	0.791	0.0005			
t_{11}	Intercept	0.021	0.000	-	0.81	0.81	0.45
	SB	0.022	0.022	0.2289			
	RSB	0.881	0.883	0.0005			
t_{12}	Intercept	-0.026	0.000	-	0.81	0.81	0.45
	SB	-0.069	-0.069	0.0475			
	RSB	0.957	0.952	0.0005			
t_{13}	Intercept	0.006	0.000	-	0.78	0.78	0.45
	SB	0.071	0.071	0.0440			
	RSB	0.831	0.830	0.0005			
t_{14}	Intercept	-0.004	0.000	-	0.87	0.87	0.47
	SB	0.084	0.085	0.0060			
	RSB	0.876	0.874	0.0005			
t_{15}	Intercept	-0.000	0.000	-	0.84	0.84	0.45
	SB	0.020	0.020	0.1964			
	RSB	0.902	0.901	0.0005			

The blockmodels for each time point are in Table 3 in three panels. The first row in each box gives the specific time point. The second row shows whether the partition reported was unique. A unique partition for 13 of the 15 time points was returned. For one time point (t_8) there are two partitions. In each case, the block structure is the same and the partitions differ only by a 'floater' moving between a pair of clusters[22]. For t_{12}, there were multiple partitions but one stands out[23]. The third row gives the value of the criterion function for $\alpha = 0.5$ (the inconsistency count is double the criterion function values reported in Figure 1). The final row in each cell gives the block structure where P, N and O denote, respectively, positive, negative and null blocks.

Table 3: Signed block structures over 15 time points: Newcomb data*.

t_1 Unique X(P)=9.5 PPNN PONN PNPN PNNP	t_2 Unique X(P)=8.0 PNPN PPNN PNNP NPPO	t_3 Unique X(P)=6.5 PNPN PPNN PNNP NPPO	t_4 Unique X(P)=8.0 POPN PPNP PNPN PPNN	t_5 Unique X(P)=6.0 PNPN PPNN PPNN PNON
t_6 Unique X(P)=5 PPON PNPN PPPN PPNN	t_7 Unique X(P)=3.0 POPN PPNN PPPN PPNN	t_8 Two X(P)=3.5 PPPN PNPN PPON PPNN	t_9 Unique X(P)=4.0 PPPN PPNN PNON PPPN	t_{10} Unique X(P)=4.5 PNPN OPPN PPNN PNNN
t_{11} Unique X(P)=3.0 PPPN PPON PNPN PPNN	t_{12} Unique** X(P)=3.0 PPNN PPNN PNOP PNPN	t_{13} Unique X(P)=3.5 PPNN POPN PPNN PNNN	t_{14} Unique X(P)=2.0 PPNN PNNN PONN PNPN	t_{15} Unique X(P)=2.5 PPPN PPON PNPN PNNN

*P denotes a positive block, N denotes a negative block and O denotes a null block.
** See footnote 15 for an explanation of this.

[22] The value of the ARI measure is 0.845 which is in the acceptable range specified by Steinley (2004).

[23] For t_{12}, it was necessary to specify two null blocks to have a unique solution. One of the identified null blocks contained a negative tie. We treated it (the third block in the first row) as a negative block. While there were multiple partitions using one specified null block, one is shown in Table 1. Specifying a second null block suggests a way of choosing a partition from the multiple equally well fitting partitions.

We examined the delineated signed blockmodel at each time point. We note that, especially towards the end of the process, the composition of the positions in terms of membership is quite stable. There are members of positions remaining firmly in place while a few do move between positions in transitions. We note also that the sizes of positions do not change abruptly in each transition. Illustrating the different partitions for structural balance and relaxed structural balance we show their unique partitions at t_{14} for $k=4$ in Figure 2. We chose this time point because it is near the end of the network evolution and the criterion functions are lowest at t_{14} for both models: each structure is closest to its ideal structure. The black squares represent positive ties with negative ties represented by grey diamonds. The SB partition is in the top panel. The RSB partition is in the bottom panel. The number of inconsistencies for structural balance is 19 while the corresponding number is 4 for relaxed balance. The reason for the sharp drop in the number of inconsistencies is clear. Structural balance struggles with the large number of off-diagonal positive ties. Also, the structural balance partition is unsatisfactory because it returns a partition with one large cluster, one pair, and two singletons. It misses the mutually hostile subgroup completely because negative blocks cannot appear on the main diagonal. The RSB partition returns an optimal partition with clusters of size 9, 3, 3 and 2. Many of the positive off-diagonal blocks are part of a coherent structure instead of contributing inconsistencies under structural balance. In short, the SB_DP_MD model fits these (t_{14}) data far better than the SB model.

It is apparent from Table 3 that *none* of the fitted RSB blockmodels conform to the SB Model. From Figure 1, the SB Model fares less well than a relaxed structural balance model, consistent with results shown in Table 2. Structural balance cannot be viewed as the sole generating process for these data. It may not be the dominant process. We next interpret the results in Table 3.

Differential popularity and Hypothesis 1 are considered first. The top left block is positive for all time points, a result consistent with *both* structural balance and differential popularity. The column of positive blocks in the left hand column is present for 12 of the 15 time points, including the last 5 leading to the final evolved structure. For t_2 and t_3, a negative off-diagonal block appears in this column. Even so, there are still two positive off-diagonal blocks. There is one null block with two positive blocks in the first column at t_{10}. This pattern provides overwhelming support for the presence of differential popularity (Hypothesis 1) and overwhelming support for Feld and Elsmore's (1984) arguments for it as a generative process. Hypothesis 1 is resolved in favor of the SB_DP model. A column of positive blocks appears early and is present for most time points. This feature is stable with decreasing inconsistencies.

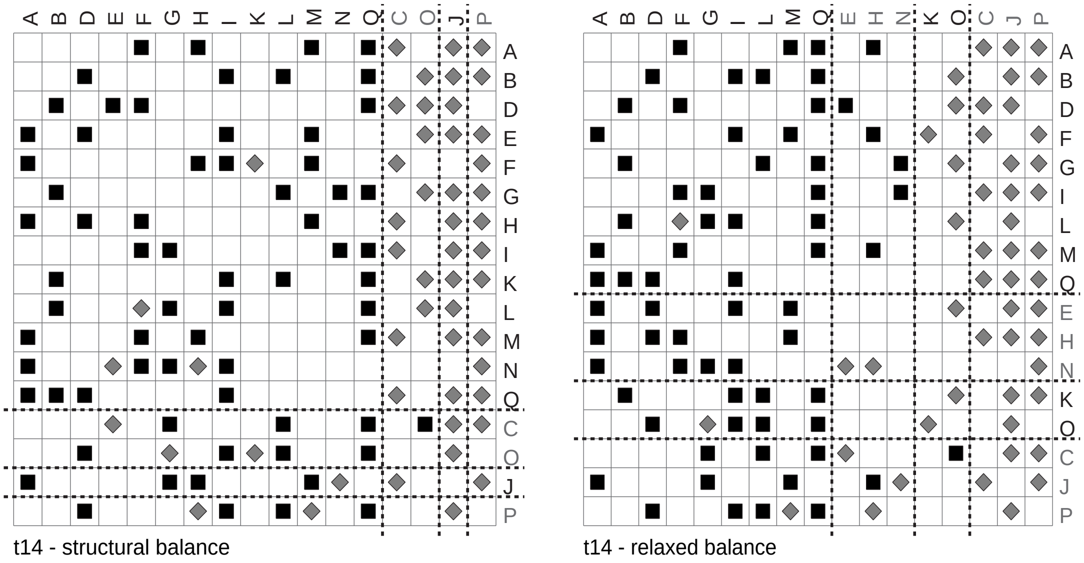
A B D E F G H I K L M N Q C O J P
A B D E F G H I K L M N Q C O J P
t14 - structural balance
A B D F G I L M Q E H N K O C J P
A B D F G I L M Q E H N K O C J P
t14 - relaxed balance

Figure 2: Structural Balance and Relaxed Balance partitions at t_{14} (Newcomb data).

Next, we consider Hypothesis 2. For differential dislike, including mutual dislike, the column of negative blocks on the right first appears at t_5. It was not there at the outset and emerged over time. It persisted through all subsequent time points. The bottom right (diagonal) negative block reveals a subgroup with mutual dislike. This also contradicts structural balance. However, negative off-diagonal blocks in this column are consistent with structural balance *and* differential dislike. Features of the SB_DP_MD Model are evident at multiple time points. Hypothesis 2 is resolved in favor of the SB_DP_MD model. There is evidence of differential popularity emerging earlier with a shorter time scale than differential dislike.

The signs of the blocks in the middle two columns for each of the fitted signed blockmodels have been treated as having secondary interest. Yet, for structural balance theory, additional positive blocks off the main diagonal and negative blocks on it provides further contradictory evidence. For eight time points there is one negative block on the main diagonal and for six there are two such negative blocks. There is strong evidence for differential popularity - in both a universal and less universal sense – as well as mutual dislike within a set of actors. These features are disentangled from balance processes because they leave observable traces inconsistent with that theory. Consistent with Hallinan's (1984) observation, structural features suggest the operation of multiple processes. Some cannot be completely distinguished by looking solely at blocks. However, there is some further evidence in favor of differential dislike.

The ideas of differential popularity and differential dislike imply that both positive and negative ties are concentrated on some actors but not on others. A natural way of considering this is by examining inequality in the receipt of these ties. Our third hypothesis claims that this inequality will increase over time. Figure 3 shows the values of the Theil entropy index over time[24]. Very similar results hold when the coefficient of variation (standard deviation/mean) or the Gini coefficient is used. The inequality for receiving negative ties increases over the first 7 time points, shows some oscillation for the next three time points, followed by a downwards drift, and then some more oscillation with increasing values. The over-time movement of inequality for the receipt of positive ties is quite different. It is flat over the first four time points, increases from t_4 through t_7, drops, and then oscillates while increasing. The inequality in the receipt of negative ties is always much higher than for the receipt of positive ties after t_1. The third hypothesis is strongly supported for received negative ties while, at best, it is supported for the receipt of positive ties from t_4 through t_7 and only weakly supported after t_7. The greater concentration of negative ties over time suggests that differential dislike generates more of the column of negative blocks than structural balance.

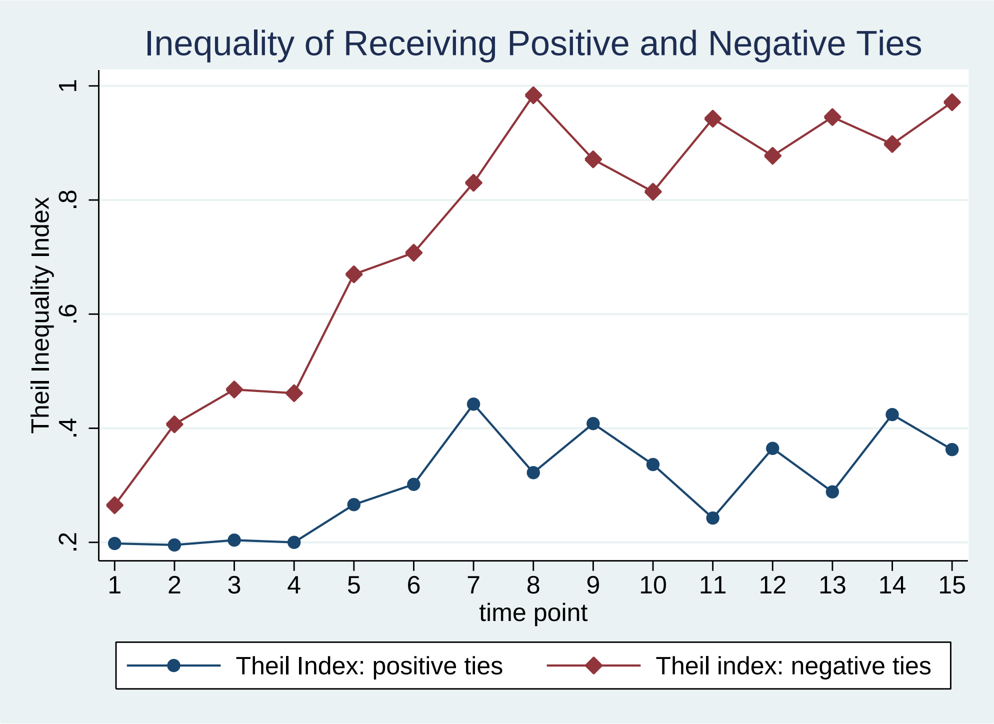

Figure 3: Inequalities in receiving positive and negative ties: Newcomb data.

[24] The results in Figure 3 are not due to having 4 positive ties and 3 negative ties from each actor. The trajectory of the Theil index, when using only 3 positive ties, is close to the trajectory of the index for 4 positive ties.

4.2 Using the Sampson Data

Sampson's (1968) data has three time points (labeled in the literature as T_2, $T_{3,}$ and $T_{4.}$ Sampson collected data for an earlier time point[25](T_1). He collected signed data on four relations: affect, esteem, influence, and sanction. Each took an apparent metric form with three ranked positive and three ranked negative ties. The sanction relation is problematic because some trainee monks refused to provide data (or claimed they sanctioned no-one). Doreian (2008) argued for using a multi-indicator approach for multiple relations. We do this here. We summed the binarized[26] affect, esteem and influence relations. The valued signed relation is the number of ties with a specific sign between pairs of actors. From prior analyses (Sampson, 1968; Breiger et al., 1975; Doreian and Mrvar, 1996), we know there are *k=3* clusters of monks. Figure 4 shows three trajectories for the criterion function. Two are for SB and RSB for *k=3*. We compare these first. The trajectory of the criterion function for relaxed structural balance for *k=4* has additional interest value regarding differential popularity.

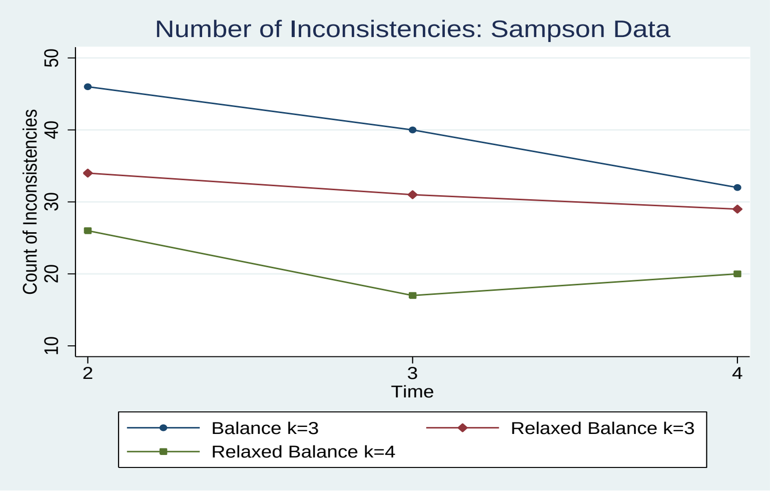

Figure 4: Inconsistency counts for the Structural Balance and the Relaxed Structural Balance models: Sampson data.

[25] The T_1 data were for a different set of monks. Some of them departed before T_2. Those who remained were joined by a group on new trainee monks at T_2.

[26] This was done because summing the ranks seems problematic with regard to measurement. The value of Cronbach's α for the three time points considered here are 0.795 (T_2), 0.777 (T_3) and 0. 849 (T_4), suggesting these three network relations are very consistent from a measurement point of view. Also, the comparisons of random partitions of the Sampson data into the same number of positions with the relaxed balance theoretic partitions, that value of the ARI ranges between -0.06 through -0.02 over the partitions reported in Table 3.

Both trajectories for *k=3* decline over time. The values of the criterion function for RSB are smaller than for SB. However, this evidence is modest: the declines for the RSB are small. For the last time point, the two values of the criterion function are close. The value of the criterion function for the RSB model for *k=4* declines from the first time point to the second but rises slightly at the third time point[27]. The values of the QAP correlations for *k*=3 are: 0.708 (T_2); 0.687 (T_3) and 0.737 (T_4). And for *k*=4 they are: 0.760 (T_2); 0.871 (T_3) and 0.816 (T_4). For all these QAP correlations p<0.001 confirming the descriptive values for the criterion function, C_f, are noteworthy.

Table 4: Signed block structures over 3 time points: Sampson data*.

Structural balance (k=3)

T_2	T_3	T_4
Unique X(P)=23 PNN NPN NNP	Unique X(P)=20 PNN NPN NNP	Unique X(P)= 16 PNN NPN NNP

Relaxed balance (k=3)

T_2	T_3	T_4
Unique X(P)= 17 PNN PPN NNP	Unique X(P) =15.5 PNN PPN NNP	Unique X(P) = 14.5 PNN PPN NNP

Relaxed balance (k=4)

T_2	T_3	T_4
Unique X(P)= 13 PNNN PPNN PNPP NNPP	Unique X(P) =8.5 PNPN PPNN PNPP NNPP	Unique X(P) = 10 PPNP PPNN PPPN PNNP

*P denotes a positive block, N denotes a negative block, O denotes a null block

[27] One problem with Sampson's data is the small number of time points. Also, the data collection, in contrast to Newcomb's data, did not start from a null network.

Table 4 presents the corresponding signed blockmodels for the three time points. For *k=3*, there are no large differences between the two blockmodels. The blockmodel for structural balance must be the SB model. For RSB, the same blockmodel existed at each time point with just one difference from the SB model: for all time points, one positive off-diagonal block is in the first column of blocks. In terms of Hypothesis 1, only a modest version of the SB_DP is present at each time point. Even so, it provided slightly better fits. Table 5 reports QAP regressions for the Sampson data. The top panel concerns the *k*=3 partitions. The RSB effect dominates SB only for T_2, consistent with the larger difference in the values of the criterion function at this time point in Table 3. In terms of Hypothesis 2, there is no for a SB_DP_MD model given the absence of a negative diagonal block. The off-diagonal negative blocks are consistent with both structural balance and differential dislike.

Table 5: QAP Regressions comparing Relaxed Structural Balance and Structural Balance: Sampson data.

A: Three positions (k=3)

T	Variable	Unstandardized Coefficient	Standardized Coefficient	p-value	R^2	R^2 (RSB)	R^2 (SB)
T_2	Intercept SB RSB	0.111 0.040 0.782	0.000 0.039 0.782	- 0.2324 0.0005	0.67	0.67	0.53
T_3	Intercept SB RSB	0.107 0.174 0.672	0.000 0.173 0.672	- 0.0075 0.0005	0.67	0.66	0.53
T_4	Intercept SB RSB	0.057 0.356 0.556	0.000 0.355 0.556	- 0.0005 0.0005	0.77	0.73	0.68

RSB Relaxed Structural Balance; SB Structural Balance

B: Four positions (k=4) RSB only

Time	Variable	Unstandardized Coefficient	Standardized Coefficient	p-value	R^2
T_2	Intercept RSB	0.031 0.858	0.000 0.859	- 0.0005	0.74
T_3	Intercept RSB	-0.001 0.889	0.000 0.889	- 0.0005	0.79
T_4	Intercept RSB	-0.045 0.903	0.000 0.902	- 0.0005	0.81

The lowest panel of Table 4 displays the blockmodel structure for relaxed balance with *k*=4. The evidence in these blockmodels is stronger for a SB_DP model fitting the Sampson data because of the presence of more off-diagonal positive blocks. At the last time point, T_4, there is a full column of positive blocks in the RSB blockmodel as well as other off-diagonal positive blocks. While structural balance works well for the Sampson data for k=3, for k=4 there is stronger evidence in favor of the SB_DP model. The corresponding results for prediction using only the RSB fitted blockmodel for the k=4 are provided in the lower panel of Table 5. This fitted blockmodel is a potent predictor of the signed relation for all three time points.

Figure 5 shows the structural balance partitions of the Sampson data for each time point. They are consistent with prior analyses with three clusters of actors: The Young Turks (John Bosco, Gregory, Mark, Winfrid, Hugh, Boniface and Albert); the Loyal Opposition (Peter, Bonaventure, Berthold, Ambrose, Victor, Romauld, Louis and Amand), and the Outcasts (Basil, Elias and Simplicius) were identified by Sampson (1968). There are some minor differences with Ambrose being in the Young Turk cluster at T3 and Amand joining the Outcasts[28] at T_4.

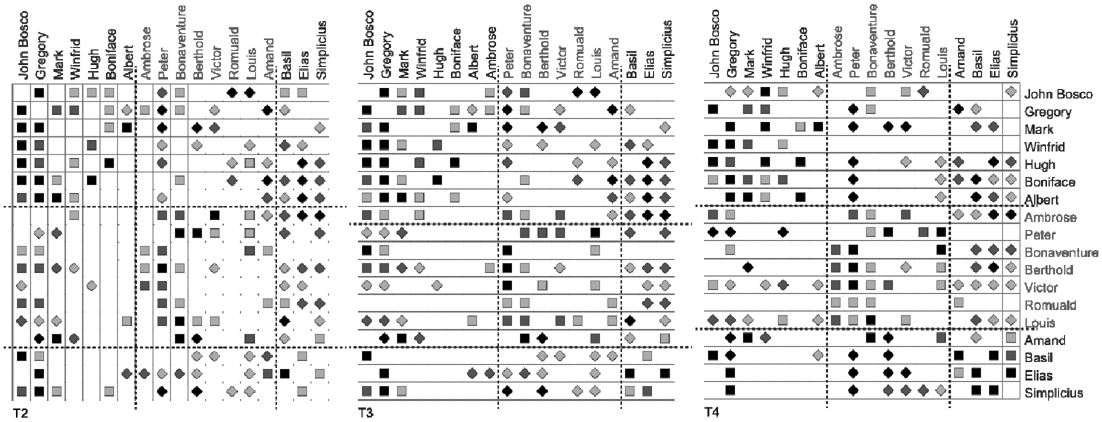

Figure 5: Structural Balance Partitions for the Sampson data at each time point.

Figure 6 shows the relaxed balance model as fitted for each time point with k = 4. For T_2, the Loyal Opposition has been split into two clusters. Four of their members (Bonaventure, Berthold, Ambrose and Romuald) send mainly positive ties to members of the Young Turks, a feature obscured in the structural balance partition. Consistent with structural balance, they send positive ties to others in the Loyal Opposition and negative ties to those in the Outcasts. The two partitions at T_3 differ only in the location of Albert, again with positive blocks off the main diagonal. At T_4, Bonaventure and Ambrose form a single cluster, receiving positive ties from members of the other three clusters. They also have reciprocated

[28] Doreian and Mrvar (1996) had Amand with the Outcasts at all three time points.

positive ties. This column of positive blocks supports the Sampson data conforming to the SB_DP model at T_4.

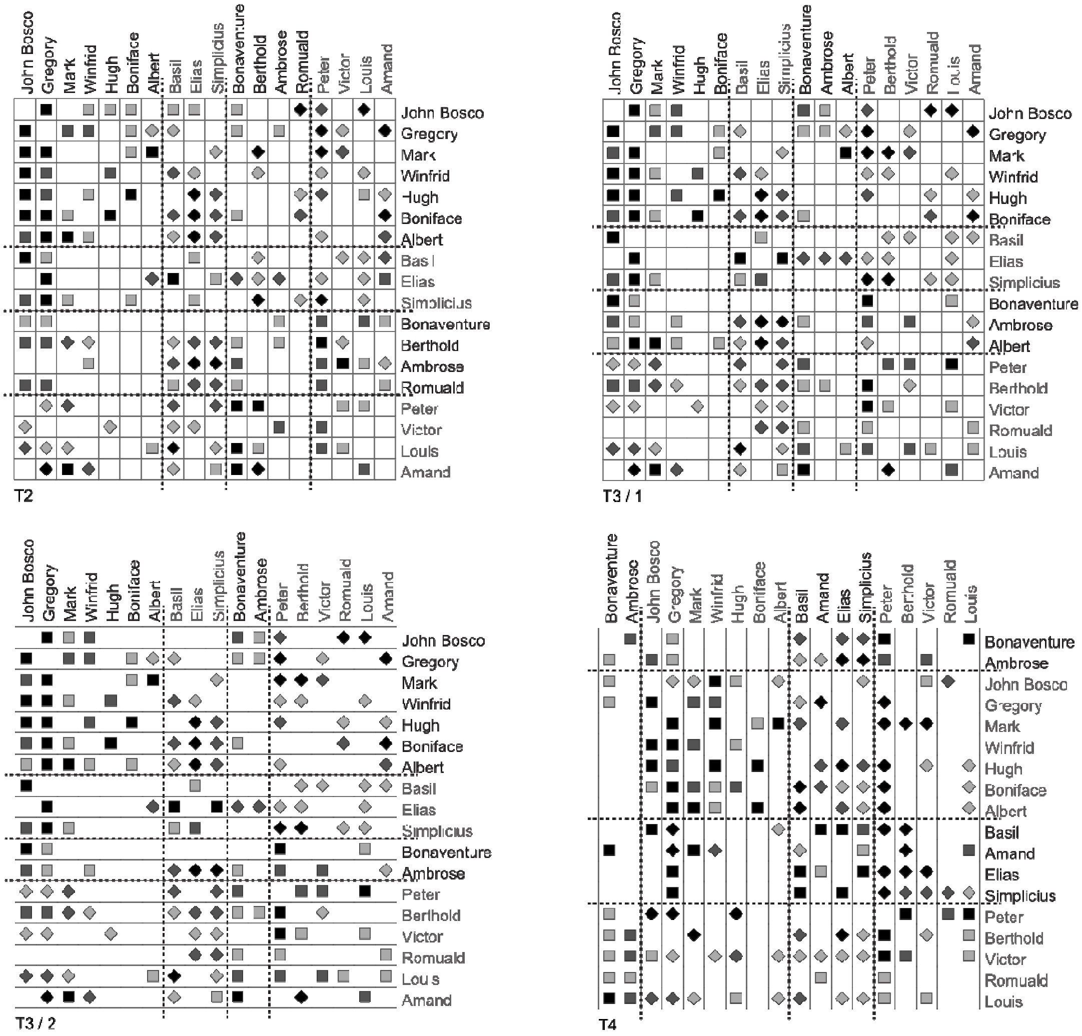

Figure 6: Relaxed Balance Blockmodels for the Sampson data at each time point (k = 4).

Figure 7 shows plots of inequality in the receipt of positive and negative ties. Consistent with the Newcomb data results, inequality of the receipt of negative ties increases across all time points. The pattern for inequality in the receipt of positive ties differs. From T_2 to T_3, it drops slightly before a sharp increase between T_3 and T_4. The highest value for each index is at T_4 providing support for Hypothesis 3 for the receipt of negative ties but only partial support for the receipt of positive ties.

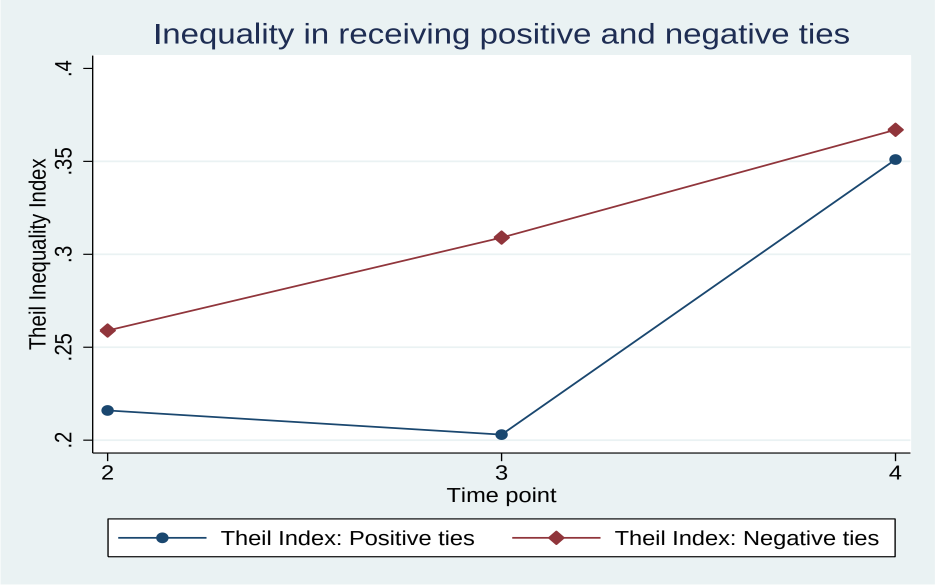

Figure 7: Inequalities in receiving positive and negative ties: Sampson data.

5 Summary and discussion

As multiple processes generate social relations among human actors, it is problematic to commit to examining *only* one process. The relevant processes include structural balance, differential popularity, differential dislike, and mutual hostility *within subgroups larger than dyads*. When processes operate they leave traces as structural features of networks. Our attempt to disentangle the results of these processes focused on the structure of the network as represented by locations of positive and negative block types in blockmodels. We used the generalized blockmodel of relaxed structural balance (Doreian and Mrvar, 2009) to fit blockmodels to signed networks. We found strong support for the operation of differential popularity in a column of off-diagonal positive blocks with the Newcomb data. Some actors in were universally popular, contrary to structural balance. Evidence was found also of subgroups of mutually hostile actors with persistent negative blocks on the main diagonal of the image matrix, also contradictory of structural balance.

The persistent presence of a column of off-diagonal negative blocks is consistent with both structural balance and differential dislike. By considering the increased concentration of negative blocks over time on a subset of actors, we infer that differential dislike contributes more than structural balance even though the results of these processes could not be disentangled completely. The results were less clear for the Sampson data where the structural balance model fared less badly than in the Newcomb data. There was some modest evidence for a weaker form of a model with differential popularity. Only at the last time point, in a model with four positions, was differential popularity more evident.

As a summary, in Newcomb's data, relaxed structural balance partitions provide strong support for the operation of differential capturing structural features at odds with structural balance. The increased concentration of negative ties on some actors suggests differential dislike is either a more potent process than structural balance or is an unrecognized component of it. The evidence for such outcomes was not as clear with Sampson's data. Yet there was support for the hypothesis regarding inequalities in receiving negative ties.

There are some caveats concerning our results because the data we used are not ideal. The recoding of Newcomb's data, used by others before us imposes the equivalent of a fixed choice design and is, at most, only an approximation of satisfactory temporal signed network data. Sampson also adopted a fixed choice design for the data he collected. Neither Newcomb's nor Sampson's data have systematic information regarding actor attributes. This imposes another limitation. Increasing concentration of receiving both positive and negative ties could rest on clearer perceptions of actor attributes and the accumulation of network processes. Without information on actor attributes and the recognition of this information by actors when forming and breaking signed ties these two processes cannot be disentangled. Some implications of these limitations are clear.

First, better over time network data for signed (and unsigned) networks in small groups are needed. Second, as networks and actors co-evolve, we need actor attribute data and (changing) actor perceptions of each other. Third, an adequate theory of network change requires reconsidering Heider's (1946) distinction between signed social relations and unit formation relations to incorporate both when studying actor and network co-evolution[29]. Using only structural (network) data is not enough. Even so, we have shown that network processes can be disentangled to some extent by delineating the structural traces that their operation leaves behind. This allowed for some comparative testing of theories about generating structures.

Such an approach can be made more fruitful by embedding signed blockmodeling in a richer substantive framework with more complete data. Here, we have written about tie formation without being attentive to the micro-processes involved for pairs of actors. Montoya and Insko (2008) analyze reciprocity in terms of affective, cognitive, and behavioral elements. Wojciszke *et al*. (2009) examine different mechanisms generating like-dislike and respect-disrespect relations. However these mechanisms operate, they will be constrained to some extent by the macro structure of the group within which they operate. It suggests also that a more general account will emerge from combining these different approaches.

[29] White (1979) notes empirical evaluations of balance theory differ according to whether *poq*-triples or *pox*-triples (with unit formation ties) are used.

Another item meriting attention comes from the differences between the two sites where Newcomb and Sampson collected their data. The students in the pseudo-fraternity of Newcomb had potential relations and contacts outside their residential hall. In contrast, the trainee monks were largely cut off from the outside world. Such differences could make a difference in the macro network structures formed (Doreian and Conti, 2012). In terms of substance, theories of how relational tie formation is dependent on the context within which relations are formed are needed for a better account of the processes of network formation and the resulting network structures.

Another very promising approach to social networks are exponential random graph models (ergms). It would seem useful to explicitly couple the micro-process generation of network structure represented in the use of dynamic exponential random graph models with the kind of block modeling approach used here. We think that coupling the ergm approach to block modeling is an step. The simplest way of doing this is to incorporate block structures as a covariate. Doreian and Conti (2012) provide an example where both estimated ergm parameters and a blockmodel covariate were significant. A much deeper approach is to develop an ergm and a blockmodel simultaneously.

We provide a different take on two classical data sets by using signed blockmodeling to comparatively assess two theories about the generation of structure. However, we are mindful that these data sets are unique and imply some problems with regard to generalization, especially to larger networks. Balance theoretic ideas were formed in the study of small networks but it is reasonable to anticipate their extension to larger signed networks where overall network density tends to be lower. This raises the issue of whether density could affect the use of relaxed structural balance and structural balance. We think this would not affect our methods, especially if fixed choice designs are avoided. However, this remains an empirical issue. In terms of formal analysis, Abell and Ludwig (2009) have launched a program of research based on simulation studies of balance processes in larger signed networks Their simulated networks are very dense and, while they are useful for studying the operation of balance processes, it is not clear that there is a direct extension to empirical signed networks.

If areas of differential density exist in large signed networks, then the empirical study of large ‘patchy’ signed networks could benefit from the kinds of community detection methods developed by Traag and Bruggeman (2009) for signed networks. We provide a methodological comparison of this algorithm with RSB in the Appendix A. For the Newcomb data, the results are mixed but point to the RSB approach as more useful. The criterion functions implied by the two algorithms are different and it may be useful in future work to try and combine them in some fashion. Having diagonal blocks with dense positive lines seems important provided that this does not destroy the block structures identified here.

References

[1] Abell, P. and Ludwig, M. (2009): Structural balance, a dynamic perspective, *Journal of Mathematical Sociology*, **33**, 1–27.

[2] Abelson, R.P., McGuire, W.J., Newcomb, T.E., Rosenberg, M.J., and Tannenbaum, P.H. (Eds.) (1968): *Theories of Cognitive Consistency: A Sourcebook*, Chicago: Rand-McNally.

[3] Batagelj, V. and Mrvar, A. (1998): Pajek program for large network analysis, *Connections*, **21***(2)*, 47-57.

[4] Borgatti, S.P. and Everett, M.G. (1999): Models of core/periphery structures, *Social Networks*, **21**, 375-395.

[5] Borgatti, S.P., Everett, M.G., and Freeman, L.C. (2002): *UCInet for Windows: Software for Social Network Analysis*. Harvard, MA: Analytic Technologies.

[6] Breiger, R.L., Boorman, S.A., and Arabie, P. (1975): An algorithm for clustering relational data with applications to social network analysis and comparison to multidimensional scaling, *Journal of Mathematical Psychology,* **12**, 328-383.

[7] Brusco, M.J., Doreian, P., Mrvar, A., and Steinley, D. (2011): Two algorithms for relaxed structural balance partitioning: Linking theory, methods and data to understand social network phenomena, *Sociological Methods and Research*, **40**, 57-87.

[8] Cartwright, D. and Harary, F. (1956): Structural balance: A generalization of Heider's theory, *Psychological Review*, **63**, 277-292.

[9] Davis, J.A. (1967): Clustering and structural balance in graphs, *Human Relations*, **20**, 181-187.

[10] Davis, J.A and Leinhardt, S. (1972): The structure of positive interpersonal relations in small groups, 218-251. In Berger, J., Zelditch Jr., M. and Anderson, B.(Eds.): *Sociological Theories in Progress Volume 2*, , Boston: Houghton Mifflin.

[11] Dekker, D., Krackhardt, D., and Snijders, T.A.B. (2007): Sensitivity of MRQAP tests to collinearity and autocorrelation conditions, *Psychometrika*, **72**, 563-581.

[12] Doreian, P. (2008): A multiple indicator approach to structural balance, *Social Networks*, **30**, 247-258.

[13] Doreian, P., Batagelj, V., and Ferligoj, A. (2005): *Generalized Blockmodeling*. New York: Cambridge University Press.

[14] Doreian, P., Batagelj, V., and Ferligoj, A. (2004): Generalized blockmodeling of two-mode network data, *Social Networks*, **26**, 29-53.

[15] Doreian, P. and Conti, N. (2012): Social infrastructure, geography and social network structure, *Social Networks*, **34**, 32-46.

[16] Doreian, P., Kapuscinski, R., Krackhardt, D., and Szczypula, J. (1996): A brief history of balance through time, *Journal of Mathematical Sociology*. **21**, 113-131 (reprinted in *Evolution of Social Networks*, Doreian, P. and Stokman, F. N. (Eds.) (pp 129-147), New York: Gordon and Breach.)

[17] Doreian, P. and Mrvar, A. (2009): Partitioning signed social networks, *Social Networks*, **31**, 1-11.

[18] Doreian, P. and Mrvar, A. (1996): A partitioning approach to structural balance, *Social Networks*. **18**, 149-168.

[19] Feld S.L. and Elsmore, R. (1984): Patterns of sociometric choices: Transitivity reconsidered, *Social Psychology Quarterly*, **45**, 77-85.

[20] Ferligoj, A., Doreian, P., and Batagelj, V. (2011): Positions and roles. In Scott, J. and Carrington, P.J. (Eds.): *Social Network Analysis*. 434-446 Los Angeles: Sage.

[21] Festinger, L. (1957): *A Theory of Cognitive Dissonance*, Evanston: Row, Peterson.

[22] Hallinan, M.T. (1984): Cognitive balance and differential popularity in social networks. *Social Psychology Quarterly*, **45**, 86-90.

[23] Harary, F., Norman, R. Z. and Cartwright, D. (1965): *Structural Models*. New York: John Wiley and Sons.

[24] Heider, F. (1946): Attitudes and cognitive organization, *Journal of Psychology*. **21**, 107-112.

[25] Heider, F. (1958): *The Psychology of Interpersonal Relations*. New York: Wiley

[26] Holland, P. and Leinhardt, S. (1973): The structural implications of measurement error in sociometry, *Journal of Mathematical Sociology*, **3**, 85-112.

[27] Holland, P. and Leinhardt, S. (1972): Some evidence on the transitivity of positive interpersonal sentiment, *American Journal of Sociology*. **72**, 1205-1209.

[28] Hummert, M.L., Crockett, W.H., and Kemper, S. (1990): Processing mechanisms underlying use of the balance schema, *Journal of Personality and Social Psychology*, **58**, 5-21.

[29] Hummon, N.P. and Doreian, P. (2003): Some dynamics of social balance processes: Bringing Heider back into balance theory, *Social Networks*. **25**, 17-49.

[30] Leicht, E.A. and Newman, M.E.J. (2008): Community structure in directed networks, *Physical Review Letters*, 100, 118703.

[31] Leik, R.K. and B.F. Meeker (1975): *Mathematical Sociology*. Englewood Cliffs, NJ: Prentice Hall.

[32] Montoya, R.M. and Insko, C.A. (2008): Towards a more complete understanding of the reciprocity of liking effect, *European Journal of Social Psychology*, **38**, 477-498.

[33] Mower White, C.J. (1979): Factors affecting balance, agreement and positivity biases in POQ and POX triads, *European Journal of Social Psychology***,** **7**, 129-148.

[34] Newcomb, T.M. (1961): *The Acquaintance Process*. New York: Holt, Rinehart, & Winston.

[35] Newman, M.E.J. (2006): Modularity and community structure in networks, *Proceedings of the National Academy of Sciences (USA),* **103** (23), 8577–8582.

[36] Nordlie, P. (1958): *A Longitudinal Study of Interpersonal Attraction in a Natural Setting*, Unpublished Ph. D. Dissertation, University of Michigan.

[37] Osgood, C.E. and Tannenbaum, P.H. (1955): The principle of congruity in the prediction of attitude change, *Psychological Review*, **62**, 42-55.

[38] Robins, G. and Kashima, Y. (2008): Social psychology and social networks: Individuals and social systems, *Asian Journal of Social Psychology*, ***11***, 1-12.

[39] Sampson, S.F. (1968): *A Novitiate in a Period of Change: An Experimental Case Study of Relationships*, Unpublished Ph.D. Dissertation, Department of Sociology, Cornell University, Ithaca, NY.

[40] Taylor, H.F. (1970): *Balance in Small Groups*. New York: Van Nostrand Reinhold.

[41] Theil, H. (1967): *Economics and Information Theory*. Chicago: Rand McNally.

[42] Traag, V.A. and Bruggeman (2009): Community detection in networks with positive and negative links, arXiv:0811.2329v3 [physics.soc=ph], 25 September, 2009.

[43] Wasserman, S. and Faust, K. (1994): *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

[44] Wojciszke, B., Abele, A.E., and Baryla, W. (2009): Two dimensions of interpersonal attitudes: Liking depends on communion, respect depends on agency, *European Journal of Social Psychology,* **39**, 973-990.

Appendix A

Another approach to partitioning networks exists within the community detection literature. Community detection and blockmodeling are two methods for partitioning social networks developed separately but with obvious parallels. In order to compare them, the algorithm of Traag and Bruggeman (2009), devised specifically for signed networks, is best placed for this. It has been implemented in pajek (Batagelj and Mrvar, 1998). The algorithm is based on an adaptation of modularity (Newman, 2006; Leicht and Newman, 2008) and maximizes positive and minimizes negative lines within diagonal blocks while it minimizes positive and maximizes negative lines in off-diagonal blocks. In using this approach, we obtained higher values of the modularity index for partitions having a high density of positive and low density of negative lines inside clusters and a high density of negative and low density of positive lines between clusters.

We note that partitioning signed networks using relaxed structural balance (RSB) is driven by substance concerning the dynamics of relations in small groups while community detection is driven more by the observation that communities have denser positive ties and sparser (or even no) negative ties within them compared to the ties to the rest of the network. It is useful for partitioning large networks. It is reasonable to compare them.

This comparison is purely methodological and takes the following form: i) produce the best partitions using the Traag and Bruggeman algorithm; ii) establish the corresponding RSB partitions (with the same values of *k*); iii) create the implied fitted matrix arrays for both; iv) establish how well they predict the actual data; and v) compare the two partitions in relation to each other. The results are shown in Table A.1: the first column lists time points; the second column has the number of positions (clusters) obtained by the community detection (CD) algorithm and used also for the corresponding RSB partitions; the third column has the variance explained by the community detection partitions; the fourth column has the variance explained by the RSB partitions; and the final column has a direct comparison of the pairs of fitted partitions. The comparison is made solely in terms of the number of clusters determined CD and defers to these values of *k*. The result is straightforward: at each time point the variance explained by the RSB approach is larger than the variance explained by community detection. However, for four time points the differences are trivially small and a reasonable conclusion is that the two partitions perform equally well in predicting the empirical relational arrays for these time points. Thereafter, in contrast, the differences are more substantial and sometimes the differences are large. We note that the correlations between the two fitted arrays are particularly high for t_4 and t_7. The variation of R^2 across the time points has more to do with the number of clusters: other things equal, using more positions leads to explaining more variance in the array of signed ties. Given that there are only 17 data points, even using 5 or 6 positions seems excessive. Using $k = 4$ for all time points, as done in the paper, seems preferable both in terms of substance and for uniform comparisons.

For the primary substantive concerns considered here, the results of using the signed community detection approach are mixed. For five time points (t_2, t_3, t_4, t_5, and t_{10}) there is no column of positive blocks. However, for the remaining times points, there is as least one column of positive blocks. This provides support for the SB_DP Model. Using this community detection algorithm permits a comparative test precluded by classical structural balance. For all time points, there are no diagonal negative blocks in the blockmodels obtained by the community detection approach: The presence of such blocks is missed and precluded the delineation (and examination) of the SB_DP_MD Model. We return to Leik and Meeker's point: coherence between substance, method, and data is important. The substantively driven RSB approach has this coherence while the community detection approach used here does not.

Table A.1: Comparing the predictive value of two partitions.

Time Point	k	R^2 (CD)	R^2 (RSB)	R^2 (CD_RSB)
t_1	3	0.27	0.33	0.22
t_2	3	0.32	0.35	0.28
t_3	3	0.35	0.41	0.50
t_4	3	0.42	0.46	0.86
t_5	3	0.30	0.53	0.34
t_6	5	0.38	0.57	0.53
t_7	6	0.81	0.93	0.87
t_8	5	0.51	0.90	0.48
t_9	4	0.28	0.75	0.30
t_{10}	4	0.29	0.74	0.30
t_{11}	5	0.66	0.90	0.64
t_{12}	5	0.66	0.90	0.69
t_{13}	4	0.64	0.78	0.64
t_{14}	5	0.75	0.90	0.72
t_{15}	4	0.42	0.84	0.40

CD - Community detection, RSB - Relaxed Structural Balance

Appendix B

All of the data analyses were done using three programs. The temporal plots in Figures 1, 3 and 4 were drawn using STATA. The fitting of blockmodels was done using Pajek (Batagelj and Mrvar, 1998) using pre-specified models. The commands for this are explained in the Pajek manual. The QAP regressions were done by using UCINET (Borgatti *et al.*, 2002). The Pajek files for doing this were imported into UCINET. Again, using QAP is documented in the manual for this suite of programs.

Metodološki zvezki, Vol. 11, No. 1, 2014, 65-78

A Comparison of Methods for the Estimation of Weibull Distribution Parameters

Felix Noyanim Nwobi[1] and Chukwudi Anderson Ugomma[2]

Abstract

In this paper we study the different methods for estimation of the parameters of the Weibull distribution. These methods are compared in terms of their fits using the mean square error (MSE) and the Kolmogorov-Smirnov (KS) criteria to select the best method. Goodness-of-fit tests show that the Weibull distribution is a good fit to the squared returns series of weekly stock prices of Cornerstone Insurance PLC. Results show that the mean rank (MR) is the best method among the methods in the graphical and analytical procedures. Numerical simulation studies carried out show that the maximum likelihood estimation method (MLE) significantly outperformed other methods.

1 Introduction

The Weibull Distribution has been widely studied since its introduction in 1951 by Professor Wallodi Weibull (Weibull, 1951). These studies range from parameter estimation; see for example, Mann et al. (1974), Johnson et al. (1994) and Al-Fawzan (2000) to diverse applications in reliability engineering especially in Tang (2004) and lifetime analysis in Lawless (1982, 2003). The popularity of the distribution is attributable to the fact that it provides a useful description for many different kinds of data, especially in emerging areas such as wind speed and finance (stock prices and actuarial data) in addition to its traditional engineering applications.

[1] Department of Statistics, Imo State University, Owerri 460222, Nigeria. Email: fnnwobi@imsu.edu.ng (corresponding author).

[2] Department of Statistics, Imo State University, Owerri 460222, Nigeria. Email: ugochukwu4all@yahoo.com

Engineers and statisticians relied mainly on probability plots, referred to as graphical procedure, to analyze life data prior to the advent of desktop computers and reliability analysis software became available. We discuss the three methods; the mean rank (MR), the median rank (MDR) and the symmetric cumulative distribution function (SCDF) in Section 2. Also in Section 2 we review three methods in the objective analytical procedure; the maximum likelihood estimation (MLE), the method of moments (MOM) and the least squares method (LSM). These methods are compared in Section 3, using the mean square error (MSE) and the maximum likelihood (LLH) criteria.

2 Methods for parameter estimation

Let $S_1, S_2, \ldots, S_N$ be a random sample of size N from a population. Define $r_t = \ln\left(s_t/s_{t-1}\right)$, $r_t \in \left(-\infty, \infty\right)$ as returns of the stock prices (say), $\left\{s_t : s_t > 0\right\}$. Let $x_t = r_t^2 \in R^+$ be hereinafter referred to as the squared returns.

2.1 The Weibull distribution

The general form of a three-parameter Weibull probability density function (pdf) is given by

$$f\left(x\right) = \frac{\beta}{\alpha}\left(\frac{x_t - \upsilon}{\alpha}\right)^{\beta-1} \exp\left\{-\left(\frac{x_t - \upsilon}{\alpha}\right)^{\beta}\right\}, \quad x, \upsilon \geq 0; \alpha, \beta > 0 \tag{2.1}$$

where; x_t is the data vector at time *t*; β is the shape parameter; α is the scale parameter that indicates the spread of the distribution of sampled data and υ is the location parameter. The Weibull probability density function satisfies the following properties:

a) If $0 < \beta < 1$, f is decreasing with $f(x) \to \infty$ as $x \to 0^+$.

b) If $\beta = 1$, f is decreasing with $f(x) \to 1$ as $x \to 0^+$.

c) If $\beta > 1$, f at first increases and then decreases, with a maximum value at the mode $x = \alpha\left(1 - 1/\beta\right)^{1/\beta}$.

d) For all $\beta > 0$, $f(x) \to 0$ as $x \to \infty$.

The cumulative distribution function (cdf) of the Weibull distribution is mathematically given as:

$$F\left(x_t\right) = 1 - \exp\left\{-\left(\frac{x_t - \upsilon}{\alpha}\right)\right\}. \tag{2.2}$$

In case of $\upsilon = 0$, the pdf in (2.1) reduces to (2.3)

$$f(x_t) = \begin{cases} \left(\frac{\beta}{\alpha}\right)\left(\frac{x_t}{\alpha}\right)^{\beta-1} \exp\left\{-\left(\frac{x_t}{\alpha}\right)^{\beta}\right\}, & x \geq 0; \alpha, \beta > 0 \\ 0, & \textit{otherwise} \end{cases} \tag{2.3}$$

with a corresponding cdf as

$$F(x_t) = \begin{cases} 1 - \exp\left\{-\left(\frac{x_t}{\alpha}\right)^{\beta}\right\}, & x \geq 0 \\ 0, \quad \textit{otherwise} \end{cases} \tag{2.4}$$

Cheng and Chen (1988) observed that the distribution interpolates between the exponential distribution $(\beta = 1)$ and Raleigh distribution $(\beta = 2)$. The mean and variance of the Weibull distribution are $E(X) = \alpha\Gamma(1+1/\beta)$ and $V(X) = \alpha^2\left[\Gamma(1+2/\beta) - \Gamma^2(1+1/\beta)\right]$ respectively, where $\Gamma(n)$ is a gamma function evaluated at *n*.

2.2 Estimation procedures

2.2.1 Graphical procedure

If both sides of the cdf in (2.4) are transformed by $\ln(1/(1-x))$, we get

$$\ln\left(\frac{1}{1-F(x_i)}\right) = \left(\frac{x_i}{\alpha}\right)^{\beta}$$

so that

$$\ln\left[\ln\left(\frac{1}{1-F(x_i)}\right)\right] = \beta \ln x_i - \beta \ln \alpha. \tag{2.5}$$

Here, x_i actually represents the order statistics $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$.

If we let $Y = \ln\left[\ln\left(1/\left(1-F(x_i)\right)\right)\right]$, $X = \ln x_i$ and $c = -\beta \ln \alpha$, then (2.5) represents a simple linear regression function corresponding to

$$Y = \beta X + c. \tag{2.6}$$

The unbiased estimate of α, the scale parameter, is calculated as

$$\hat{\alpha} = \exp\left[-\left(\frac{c}{\beta}\right)\right] \tag{2.7}$$

where *c* is the intercept of the linear regression (2.6).

Thus, we perform the estimation of α and β using the following methods of estimation in Table 1.

Table 1: Methods of estimation by graphical procedure

Method	$F(x_i)$
Mean Rank	$i/(n+1)$
Median Rank	$(i-0.3)/(n+0.4)$
Symmetric CDF	$(i-0.5)/n$

We plot Y_i, which is a function of $F(x_i)$, versus $X_i (= \ln(x_i))$, using the following procedure:

a) Rank the data $\{x_i\}$ in ascending order of magnitude;

b) Estimate $F(x_i)$ of the *i* th rank order; and

c) Plot Y_i versus X_i.

This plot produces a straight line from which we obtain $\hat{\beta}$ and $\hat{\alpha}$ (see (2.6) and (2.7)).

2.2.2 Analytical procedure

***Maximum Likelihood Estimation* (MLE)**

The method of maximum likelihood estimation is a commonly used procedure for estimating parameters, see, e.g., Cohen (1965) and Harter and Moore (1965). Let $x_1, x_2, ..., x_n$ be a random sample of size *n* drawn from a population with probability density function $f(x, \underline{\lambda})$ where $\underline{\lambda} = (\beta, \alpha)$ is an unknown vector of parameters, so that the likelihood function is defined by

$$L = f(\alpha, \beta) = \prod_{i=1}^{n} f(x_t, \underline{\lambda}) \tag{2.8}$$

The maximum likelihood of $\underline{\lambda} = (\beta, \alpha)$, maximizes L or equivalently, the logarithm of L when

$$\frac{\partial \ln L}{\partial \underline{\lambda}} = 0, \tag{2.9}$$

see, for example, Mood et al (1974). Consider the Weibull pdf given in equation (2.3), its likelihood function is given as:

$$L\left(x_1,x_2,...,x_n;\beta,\alpha\right)=\prod_{t=1}^{n}\left(\frac{\beta}{\alpha}\right)\left(\frac{x_t}{\alpha}\right)^{\beta-1}\exp\left[-\left(\frac{x_t}{\alpha}\right)^{\beta}\right]$$

$$=\left(\frac{\beta}{\alpha}\right)\left(\frac{x_t}{\alpha}\right)^{n\beta-n}\sum_{t=1}^{n}x_t^{(\beta-1)}\exp\left[-\sum_{t=1}^{n}\left(\frac{x_t}{\alpha}\right)^{\beta}\right]. \tag{2.10}$$

Taking the natural logarithm of both sides yields

$$\ln L=n\ln\left(\frac{\beta}{\alpha}\right)+(\beta-1)\sum_{t=1}^{n}x_t-\ln\left(\alpha^{\beta-1}\right)-\sum_{t=1}^{n}\left(\frac{x_t}{\alpha}\right)^{\beta} \tag{2.11}$$

and differentiating (2.11) partially w.r.t β and α in turn and equating to zero, we obtain the estimating equations as follows

$$\frac{\partial}{\partial\beta}\ln L=\frac{n}{\beta}+\sum_{t=1}^{n}\ln x_t-\frac{1}{\alpha}\sum_{t=1}^{n}x_t^{\beta}\ln x_t=0 \tag{2.12}$$

and

$$\frac{\partial}{\partial\alpha}\ln L=-\frac{n}{\alpha}+\frac{1}{\alpha^2}\sum_{t=1}^{n}x_t^{\beta}=0. \tag{2.13}$$

From (2.13) we obtain an estimator of α as

$$\hat{\alpha}_{mle}=\frac{1}{n}\sum_{t=1}^{n}x_t^{\hat{\beta}} \tag{2.14}$$

and on substitution of (2.14) in (2.12) we obtain

$$\frac{1}{\beta}+\frac{1}{n}\sum_{t=1}^{n}\ln x_t-\frac{\sum_{t=1}^{n}x_t^{\beta}\ln x}{\sum_{t=1}^{n}x_t^{\beta}}=0 \tag{2.15}$$

which may be solved to obtain the estimate of β using Newton-Raphson method or any other numerical procedure because (2.15) does not have a closed form solution. When $\hat{\beta}_{mle}$ is obtained, the value of $\hat{\alpha}$ follows from (2.14).

Method of Moments (MOM)

The second procedure we consider here is the MOM which is also commonly used in parameter estimation. Let $x_1,x_2,...,x_n$ represent a set of data for which we seek an unbiased estimator for the kth moment. Such an estimator is generally given by

$$\hat{m}_k = \frac{1}{n}\sum_{t=1}^{n} x_t^k \tag{2.16}$$

where $\hat{m}_k$ is the estimate of k^{th} moment. For the Weibull distribution given in (2.3), the k^{th} moment is given by

$$\mu_k = \left(\frac{1}{\alpha^k}\right)^{-\frac{k}{\beta}} \Gamma\left(1+\frac{k}{\beta}\right) \tag{2.17}$$

where Γ is as defined in subsection 2.1. From (2.17), we can find the 1st and 2nd moments about zero as follows

$$\hat{m}_1 = \hat{\mu}_1 = \left(\frac{1}{\alpha}\right)^{\frac{1}{\beta}} \Gamma\left(1+\frac{1}{\beta}\right) \tag{2.18}$$

and

$$\hat{m}_2 = \hat{\mu}^2 + \hat{\sigma}^2 = \left(\frac{1}{\alpha}\right)^{\frac{2}{\beta}} \left[\Gamma\left(1+\frac{2}{\beta}\right) - \Gamma\left(1+\frac{1}{\beta}\right)^2\right] \tag{2.19}$$

When we divide the square of $\hat{m}_1$ by $\hat{m}_2$, we get an expression which is a function of only β,

$$\frac{\hat{\mu}^2}{\hat{\sigma}^2+\hat{\mu}^2} = \frac{\Gamma\left(1+\frac{1}{\beta}\right)\Gamma\left(1+\frac{1}{\beta}\right)}{\Gamma\left(1+\frac{2}{\beta}\right)} \tag{2.20}$$

where $\hat{\mu} = E\left(X_t\right) = \frac{1}{n}\sum_{t=1}^{n} x_t$, $\hat{\sigma}^2 = E\left(X_t^2\right) - \left(E\left(X_t\right)\right)^2$ and letting $Z = 1/\beta$ (2.19) is easily transformed in order to estimate β so that the scale parameter α_{mom} can be estimated with the following relation

$$\hat{\alpha}_{mom} = \hat{\mu} \Big/ \Gamma\left(1+\frac{1}{\beta}\right). \tag{2.21}$$

The Least Squares Method (LSM)

The Least Squares method is commonly applied in engineering and mathematics problems that are often not thought of as an estimation problem. We assume that there is a linear relationship between two variables. Assume a dataset that constitute a pair $\left(x_t, y_t\right) = \left(x_1, y_1\right), \left(x_2, y_2\right), \ldots, \left(x_n, y_n\right)$ were obtained and plotted. The least squares principle minimizes the vertical distance between the data points and the straight line fitted to the data, the best fitting line to this data is the straight line: $y_t = \alpha + \beta x_t$ such that

$$Q(x;\alpha,\beta) = \sum_{t=1}^{n} (y_t - \alpha - \beta x_t)^2$$

To obtain the estimators of α and β we differentiate Q w.r.t α and β. Equating to zero subsequently yields the following system of equations:

$$\frac{\partial Q}{\partial \alpha} = -2\sum_{t=1}^{n} (y_t - \alpha - \beta x_t)^2 \qquad (2.22)$$

and

$$\frac{\partial Q}{\partial \beta} = -2\sum_{t=1}^{n} (y_t - \alpha - \beta x_t)^2 x_t = 0 \qquad (2.23)$$

Expanding and solving equations (2.21) and (2.22) simultaneously, we have

$$\hat{\beta} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2} \qquad (2.24)$$

and

$$c = \overline{y} - \hat{\beta}\overline{x} \; ; \quad \hat{\alpha} = \exp\left(-\frac{c}{\hat{\beta}}\right) \qquad (2.25)$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the unbiased estimators of α and β respectively.

3 Method assessment and selection

3.1 Comparison of estimation methods

The Mean Squared Error (MSE) criterion is given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n} \left[\hat{F}(x_i) - F(x_i)\right]^2 \qquad (3.1)$$

where $\hat{F}(x_i)$ is obtained by substituting the estimates of α and β (for each method) in (2.4) while $F(x_i) = i/n$ is the empirical distribution function. The method with the minimum mean squared error $(MSE_{\min})$ becomes the best method for the estimation of Weibull parameters among the candidate methods.

3.2 Goodness-of-fit tests

Goodness-of-fit test procedures are intended to detect the existence of a significant difference between the observed (empirical) frequency of occurrence of an item and the theoretical (hypothesized) pattern of occurrence of that item. Here, we assume that the Weibull distribution is a good fit to the given dataset; otherwise, this assumption is nullified if, for this test, the computed statistic is greater or equal to a defined critical value.

Kolmogorov–Smirnov test

The Kolmogorov-Smirnov test is used to decide if a sample comes from a population with specific distribution. It is based upon a comparison between the empirical distribution function (ECDF) and the theoretical one defined as $F(x)=\int_{\infty}^{x} f(y,\theta)dy$ where $f(x,\theta)$ is the pdf of the Weibull distribution. Given n ordered data points $X_1, X_2, ..., X_n$, the ECDF is defined as $F(X_i)=N(i)/n$ where $N(i)$ is the number of points less the X_i (X_i are ordered from smallest to highest value). The test statistic used is

$$D_n = \underset{1\le i\le n}{\mathrm{Sup}}\left|\hat{F}(x_i)-F(x_i)\right|. \tag{3.2}$$

The statistic D_n converges to zero almost surely as $n \to \infty$.

4 Implementation

4.1 Data

The data used for this study is the weekly stock prices (*N* = 100 weeks) collected from Cornerstone Insurance Company PLC, a public liability company listed in the Nigerian Stock Exchange (Appendix I). The squared returns, r^2, earlier defined in Section 2 are a measure of volatility in the stock prices and are multiplied by 100 without loss of generality. In Figure 1 we present a graphic relationship between the weekly stock prices and its squared returns. We perform the estimation of the parameters using the R software for the graphical and analytical procedures with $100r^2$ as the dataset and *r* is now of length *n*. R is a language and environment for statistical computing and graphics (from the R Foundation for Statistical Computing (2013)) ran on the Platform: i386-w64-mingw32/i386 (32-bit).

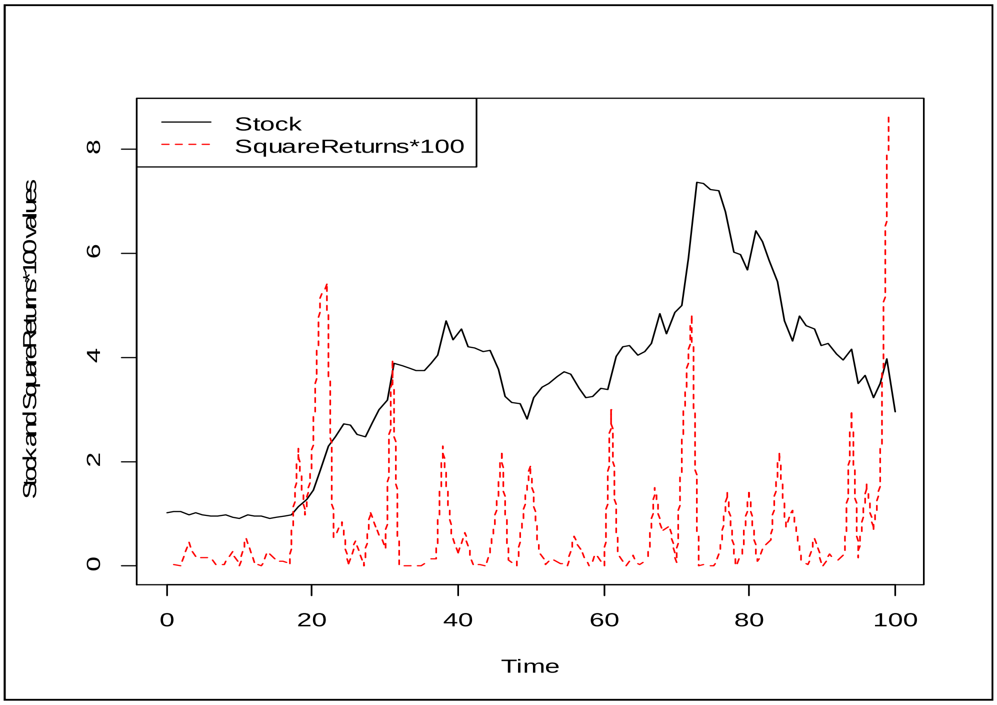

Figure 1: Plot showing relationship between Weekly Stock Prices and its Squared Returns*100

4.2 Simulation study

We carry out a numerical simulation study in order to investigate the behavior of the shape and scale parameters of the Weibull distribution. In the simulation experiment we set the Weibull distribution on the random variable X with shape parameter $\beta = 0.54$ with the aim of mimicking the squared returns $\left(100r^2\right)$. For the Weibull distribution on X, generate independently and identically distributed random sample $\left(x_1, x_2, ..., x_n\right)$ of size n (= 25, 50, 75, 100, 125, 150, 175, 200). Compute the mean of this sample and replicate this process N times to obtain a series. For each series of size n, estimate β and α using the methods described in Section 2, the MSE and the Kolmogorov-Smirnov (KS) statistic. This sequence is of the form $X^*_{1,...,N} = \text{mean}\left(x^*_1, ..., x^*_n\right)_1, \text{mean}\left(x^*_1, ..., x^*_n\right)_2, ..., \text{mean}\left(x^*_1, ..., x^*_n\right)_N$, $N = 10000$ times; and is accomplished in R for Windows 2013 by the replicate function: $\text{replicate}\left(N, \text{ mean}\left(\text{rweibull}\left(n, \text{ shape} = 0.54\right)\right)\right)$.

We remark here that the least squares method (LSM) is related to the graphical procedure in the estimation of Weibull parameters through (2.6), where $Y = \ln\left[\ln\left(1/\left(1 - F\left(x_i\right)\right)\right)\right]$ is dependent upon the particular graphical method (e.g., $F\left(x_i\right) = i/\left(n+1\right)$ for the mean rank) and $X = \ln x_i$; see also equations (2.7) and (2.25).

4.3 Results and Discussion

All computations and simulations in this investigation were done in R version 3.0.0. We relied on the functions fitdist() and fitdistr() respectively from R packages fitdistrplus and MASS (see, e.g., Delignette-Muller et al (2013) and Ripley (2013) respectively) for maximum likelihood estimation of the parameters and plots while codes were developed for the other methods. Results for the graphical procedure (MR, MDR and SCDF) were verified using the approach in Dorner (1999) on Microsoft Excel 2013. The R code used for this study is available from the first author on request.

Estimates of the parameters based upon both the graphical and theoretical procedures described in Section 2.2 are presented in Table 2. The shape parameter β lies within the interval (0, 1) which implies, as indicated in Section 2.1, that the function (irrespective of the method) decreases exponentially. We ranked the performance of the methods based on the least MSE criterion. In comparison, the Mean Rank (MR) method has the least MSE ($3.88\text{x}10^{-03}$) and at the same time has the least D_n (0.0563) making it the best among the five methods under study (graphical and analytical procedures) for this particular dataset. The Maximum Likelihood Estimation (MLE) method is, however, superior to Method of Moments in the analytical procedure. From these results the best estimate for the shape and scale parameters are respectively $\left(\hat{\beta},\hat{\alpha}\right)=\left(0.5325,0.4539\right)$ based on our dataset.

The visual assessments of fit are shown in the histogram (Figure 2(a)) overlaid with the Weibull densities generated from the different methods and in the empirical cumulative distribution function plot of Figure 2(b). The MOM is clearly different from other methods given their MSEs but this difference is not very clear in Figure 2. However, simulation results show (Table 3) that the MLE performed best 86% of the time when the n_i simulations are run 10,000 times. Similar result was obtained when the KS goodness-of-fit test was conducted to test the adequacy of the Weibull distribution in fitting the simulation data.

Table 2: Summary of results and comparison of methods for Weibull parameter estimation

Procedure	Method	$\hat{\alpha}$	$\hat{\beta}$	MSE	KS
Graphical	MR	0.4539	0.5325	$3.88\text{x}10^{-03}$	0.0563
	MDR	0.4494	0.5452	$4.21\text{x}10^{-03}$	0.0615
	SCDF	0.4461	0.5553	$4.49\text{x}10^{-03}$	0.0656
Analytical	MLE	0.4563	0.5421	$6.59\text{x}10^{-03}$	0.0617
	MOM	0.5244	0.6026	$1.18\text{x}10^{-01}$	0.1055

Table 3 Simulation results (based on 10,000 iterations)

		Method				
n	Measure	MR	MDR	SCDF	MLE	MOM
25	MSE	3.5726	3.5815	3.5837	**1.2557**	1.6770
	KS	0.0600	0.0600	0.0601	**0.0501**	0.9821
50	MSE	4.6281	4.6323	4.6282	**1.4930**	3.5122
	KS	0.0681	0.0682	0.0683	**0.0540**	0.9596
75	MSE	4.9234	4.9502	4.9407	**1.5438**	4.2108
	KS	0.0683	0.0684	0.0684	**0.0563**	0.9741
100	MSE	4.8839	4.9119	4.8985	**1.3216**	4.4869
	KS	0.0653	0.0654	0.0654	**0.0587**	0.0964
125	MSE	5.2496	5.2389	5.2598	**1.4261**	4.9398
	KS	0.0750	0.0750	0.0751	**0.0590**	0.9600
150	MSE	5.4266	5.4118	5.4341	**1.4671**	5.2043
	KS	0.0672	0.0671	0.0673	**0.0604**	0.9665
175	MSE	6.4067	6.3872	6.4096	**1.7235**	6.0586
	KS	0.0726	0.0726	0.0726	**0.0657**	0.9720
200	MSE	5.1548	5.1831	**1.3525**	1.4170	5.0833
	KS	0.0674	0.0675	0.0818	**0.0614**	0.9816

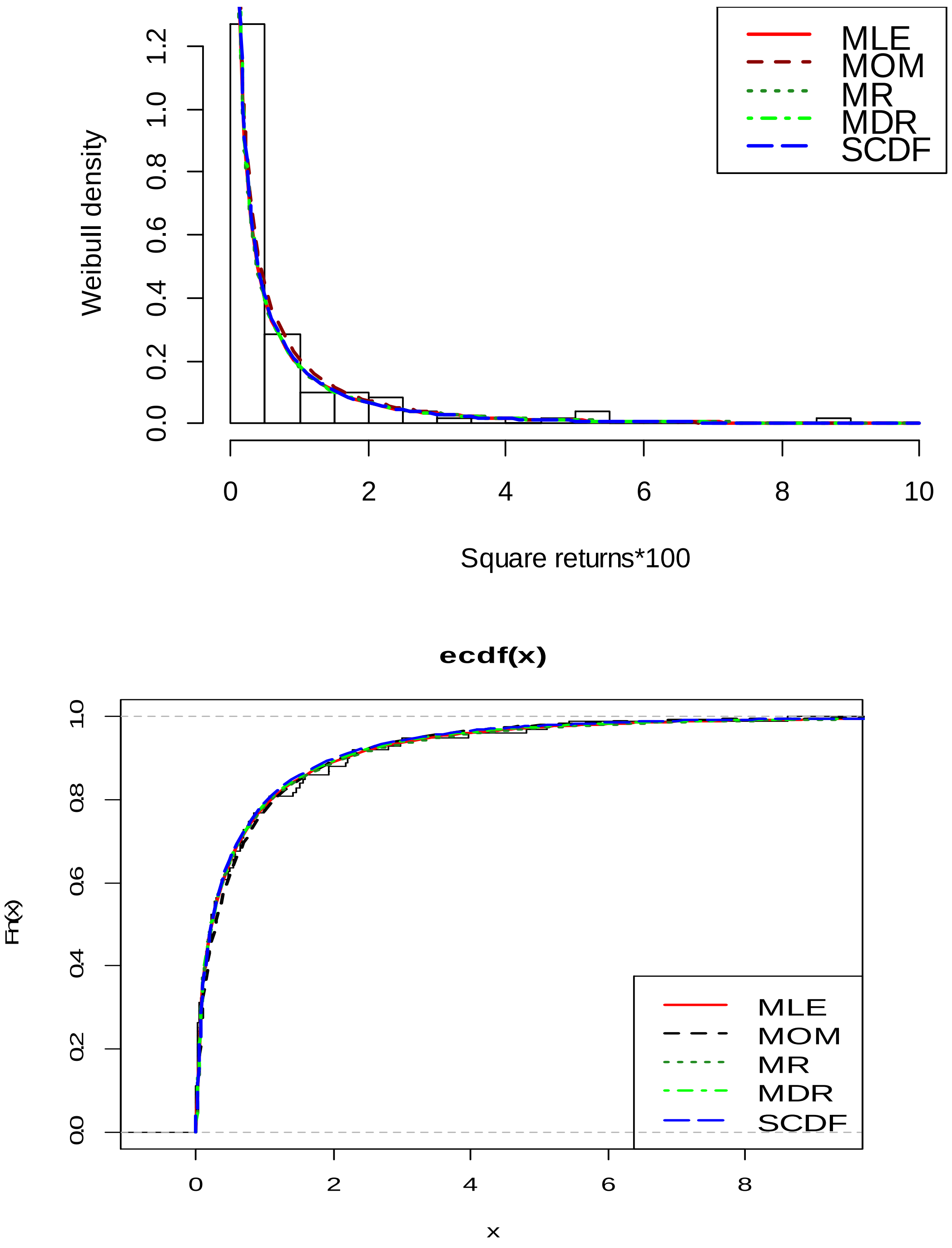

Figure 2: Fit of different methods (a) Density and Histogram (b) ECDF

5 Conclusion

The performances of five methods in the estimation of the parameters of the Weibull distribution were compared in this study. The MR was selected as the best method that gives the best estimates of the two-parameter model for square returns dataset, while the MLE is preferred over the MOM for the analytical procedure. These decisions were based on the minimum MSE criterion. When these methods were compared based upon simulation results, the maximum likelihood estimate method showed superiority over other methods. The least squares method (LSM), we remark, is also known as the rank regression method (RRM) because the estimation of the parameters of the Weibull distribution is dependent upon regressing some form of log and rank transformations of a given dataset according to the rank plotting position.

References

[1] Al-Fawzan, M. (2000): Methods for Estimating the Parameters of Weibull Distribution. King Abdulaziz City for Science and Technology, Saudi Arabia.

[2] Cheng, S. K. and Chen, C. H. (1988): Estimation of the Weibull parameters with grouped data. *Communications in Statistics: Simulation and Computation*, **11,** 197–216

[3] Cohen, A. C. (1965): Maximum Likelihood Estimation in the Weibull Distribution Based on Complete and on Censored Samples, *Technometrics*, **7** (3).

[4] Cornerstone Insurance Company Plc. www.conerstoneinsuranceplc.com Accessed 16th September, 2012.

[5] Delignette-Muller, M. L. Pouillot, R. Denis, J. Dutang, C. (2013): *R Package fitdistrplus*. http://www.cran.r-project.org/package=fitdistrplus

[6] Dorner, W. W. (1999): Using Microsoft Excel for Weibull Analysis. www.qualitydigest.com/jan99/html/weibull.html

[7] Harter, H. L. and Moore, A. H. (1956): Maximum Likelihood Estimation of the Parameters of Gamma and Weibull Populations from Complete and Censored samples. *Technometrics*, **7** (4)

[8] Johnson, N. L. Kotz, S. and Balakrishnan, N. (1994): *Maximum Likelihood Estimation for Weibull Distribution*. John Wiley & Sons, New York.

[9] Lawless, J. F. (1982): *Statistical Models for Lifetime Data*. 2nd Edition, John Wiley & Sons, New York.

[10] Lawless, J. F. (2003): *Statistical Models and Methods for Life time Data*. 3rd Edition, John Wiley and Sons, New York.

[11] Mann, N. R, Schafer, R. E. and Singpurwalla, N. D. (1974): *Methods of Statistical Analysis of Reliability and Life Data*. John Wiley & Sons, New York.

[12] Mood, A. M., Graybill, F. A. and Boes, C. D. (1974): *Introduction to the theory of Statistics*. 3rd Edition, McGraw Hill, Kogkusha.

[13] R Development Core Team (2013) http://www.r-project.org

[14] Ripley, B. (3013): R package *MASS*. http://www.cran.r- project.org/package =MASS

[15] Tang, Y. (2004): *Extended Weibull Distributions in Reliability Engineering.* A Thesis Submitted to Department of Industrial & System Engineering, National University of Singapore.

[16] Weibull, W. (1951): A Statistical Distribution of wide Applicability. *Journal of Applied Mechanics*, **18**, 239–296.

Appendix

Table A1: Weekly stock prices (read row-wise)

1.03	1.06	0.99	1.03	0.99	0.95	0.96	0.98	0.93	1.05
0.92	0.99	0.97	0.96	0.91	0.94	0.97	0.99	1.15	1.27
1.46	1.83	2.31	2.49	2.73	2.70	2.52	2.49	2.76	3.00
3.18	3.88	3.84	3.79	3.76	3.75	3.89	4.04	4.70	4.34
4.55	4.20	4.19	4.12	4.13	3.77	3.25	3.14	3.12	2.82
3.24	3.44	3.50	3.64	3.72	3.68	3.41	3.24	3.26	3.42
3.38	4.02	4.21	4.23	4.04	4.11	4.28	4.84	4.46	4.87
5.00	5.91	7.36	7.34	7.23	7.19	6.79	6.03	5.97	5.69
6.42	6.23	5.86	5.46	4.71	4.32	4.79	4.62	4.54	4.22
4.28	4.08	3.95	4.16	3.50	3.65	3.22	3.50	3.97	2.96

INSTRUCTIONS TO AUTHORS

Language: *Metodološki zvezki – Advances in Methodology and Statistics* is published in English.

Submission of papers: Authors are requested to submit their articles (complete in all respects) to the Editor by e-mail (MZ@stat-d.si). Contributions are accepted on the understanding that the authors have obtained the necessary authority for publication. Submission of a paper will be held to imply that it contains original unpublished work and is not being submitted for publication elsewhere. Articles must be prepared in LaTeX or Word. Appropriate styles and example files can be downloaded from the Journal's web page (http://www.stat-d.si/mz/).

Review procedure: Manuscripts are reviewed by two referees. The editor reserves the right to reject any unsuitable manuscript without requesting an external review.

Preparation of manuscripts

Tables and figures: Tables and figures must appear in the text (not at the end of the text). They are numbered in the following way: Table 1, Table 2,…, Figure 1, Figure 2,…

References within the text: The basic reference format is (Smith, 1999). To cite a specific page or pages use (Smith, 1999: 10-12). Use "et al." when citing a work by more than three authors (Smith et al., 1999). The letters a, b, c etc. should be used to distinguish different citations by the same author(s) in the same year (Smith, 1999a; Smith, 1999b).

Notes: Essential notes, or citations of unusual sources, should be indicated by superscript number in the text and corresponding text under line at the bottom of the same page.

Equations: Equations should be centered and labeled with two numbers separated by a dot enclosed by parentheses. The first number is the current section number and the second a sequential equation number within the section, e.g., (2.1)

Author notes and acknowledgements: Author notes identify authors by complete name, affiliation and his/her e-mail address. Acknowledgements may include information about financial support and other assistance in preparing the manuscript.

Reference list: All references cited in the text should be listed alphabetically and in full after the notes at the end of the article.

References to books, part of books or proceedings:

[1] Smith, J.B. (1999): *Title of the Book.* Place: Publisher.
[2] Smith, J.B. and White A.B. (2000): *Title of the Book.* Place: Publisher.
[3] Smith, J. (2001): Title of the chapter. In A.B. White (Ed): *Title of the Proceedings*, 14-39. Place: Publisher.

Reference to journals:

[4] Smith, J.B. (2002): Title of the article**.** *Name of Journal***,** **2**, 46-76.

Metodološki zvezki
Advances in Methodology and Statistics

Published by
Faculty of Social Sciences
University of Ljubljana, for
Statistical Society of Slovenia

Izdajatelj
Fakulteta za družbene vede
Univerze v Ljubljani za
Statistično društvo Slovenije

Editors — Urednika
Valentina Hlebec
Lara Lusa

Founding Editors — Prva urednika
Anuška Ferligoj
Andrej Mrvar

Cover Design — Oblikovanje naslovnice
Bojan Senjur
Gregor Petrič

Typesetting — Računalniški prelom
Lara Lusa

Printing — Tisk
Littera picta d.o.o.
Ljubljana, Slovenia

MZ
is indexed
and abstracted in
je indeksirana
in abstrahirana v

SCOPUS
EBSCO
ECONIS
STMA-Z
ProQuest

Home page URL — Spletna stran

http://www.stat-d.si/mz/

ISSN 1854 - 0023