23rd International Congress on Modelling and Simulation, Canberra, ACT, Australia, 1 to 6 December 2019
mssanz.org.au/modsim2019

Correspondence analysis approach to examine the Nobel Prize

T. Alhuzali [a], E. Stojanovski [a] and E. J. Beh [a]

[a] School of Mathematical and Physical Science, University of Newcastle
Email: T.Alhuzali@uon.edu.au

Abstract: The main goal of this study is to examine Nobel Prize data by exploring and studying the association between the Country of the nominated individual (or of the nominated team) and the Discipline in which the Nobel Prize was awarded. The sample studied comprises the eight the most developed countries that received at least one Nobel Prize in the period from 1901 to 2018; these being Canada, France, Germany, Italy, Japan, Russia, British Isles and the USA. The variables Country and Discipline are cross-classified to form a two-way contingency table. Simple correspondence analysis is performed to explore the nature of the association between these two variables.

Keywords: *Simple correspondence analysis, Chi squared test of independence, Nobel Prize*

1. INTRODUCTION

The Nobel Prize is considered internationally as one of the most prestigious awards whereby recipients receive prizes for groundbreaking contributions to the scientific disciplines of chemistry, physics, and physiology or medicine, as well as for the global impact of cultural activities related to literature, economics, and peace. The Nobel Prize in the discipline of Physics and Chemistry is granted from The Royal Swedish Academy of Sciences while The Nobel Assembly at the Karolinska Institute in Sweden awards this prize in Physiology or Medicine (Wallin, 2001). The Nobel Prize in Literature is awarded by The Swedish Academy. These three institutions also have special Nobel committees that grant a Nobel Prize in the discipline of Economics. The sixth prize is the Nobel Peace Prize, which is awarded by the Norwegian Nobel Committee, through a committee that is organised by the Norwegian Parliament (Levinovitz and Ringertz, 2001).

The history of the Nobel Prize dates back to Swedish businessman and engineer Alfred Nobel who invented dynamite in 1867. This invention was the main reason behind Nobel's fortune, which was estimated at 31.5 million Swedish Krone (The Nobel Peace Prize, 2019). In 1895, Alfred Nobel recommended directing his fortune to the creation of an institute to award prizes to those who contributed significant works that benefit each of chemistry, physics, physiology or medicine, literature, economic and peace disciplines and, as a result, each of these prizes were named in his honor (Morais, 2018). Alfred Nobel died in 1896, leaving behind substantial achievements which have contributed positively to the progress of mankind in many areas of science and culture and, through his will, directing most of his fortune to the creation of these prizes. The first prize was awarded in 1901 which coincided with the fifth anniversary of Alfred Noble's death. From then on, the Nobel Prize has been awarded annually to any individual or team in the scientific disciplines.

Nilesh and Pranav (2018) analysed data, using descriptive statistics, of approximately 204 Nobel laureates in the physical sciences domain. The analysis examined a diversity of variables including research fields in Physics, the age of the laureates at which the developments were carried out, the universities where their work was undertaken, migration status of the researchers as well as gender. Simple visual summaries have also been carried out to classify prize winners based on their year of birth and age at the time of their discovery (Nilesh and Pranav, 2018).

The purpose of the present paper is to analysethe Nobel Prize data using a data visualization technique that highlights the association between discipline of the awarded prize and the country where the individual, or team, carried out their work. Since the data we shall be analysing consists of categorical variables, the technique that we shall use is simple correspondence analysis. We shall confine our attention to the application and interpretation of the visual summary of association that comes from the analysis. However, there is plenty in the correspondence analysis literature that demonstrates how inferential statistics, clustering approaches and its parallels with other multivariate analytic techniques, play a role. One may, for example, refer to Greenacre (1984) and Beh and Lombardo (2014). Further details on the history of this technique, theoretical, practical and computational issues are described in Beh (2004) and Beh and Lombardo (2014). The aim is to make a statistical evaluation of the association for the data available from the time of the inception of prizes through to the present. There has been very little done in the correspondence analysis literature to account for time-dependent categories, so this is certainly one area for future investigation.

2. METHOD

2.1. Data

The sample examined consists of Nobel Prize data from the eight most developed countries where the work that was carried out earned their recipients a Nobel Prize. This data spans the period from 1901 to 2018 and was collected from two official websites; the Nobel Foundation (NobelPrize.org, 2019) and Encyclopedia Britannica (Encyclopedia Britannica, 2019). The eight *Country* included in the sample are Canada (CAN), France (FRA), Germany (DEU), Italy (ITA), Japan (JAP), Russia (RUS), British Isles (BI) and the United States of America (USA). The *Disciplines* are Chemistry (Chm), Economy (Eco), Literature (Lit), Medicine (Med), Peace (Pce) and Physics (Phy), giving a total sample size of N = 785 prizes.

2.2. Statistical analysis

An examination of the Nobel Prize data requires first exploring the association between the variables *Country* and *Discipline* of the Nobel Prize laureate(s). The association between these variables was analysed using a chi-squared test of independence to determine whether the association is statistically significant. However this method alone does not indicate the nature of this association. For any statistically significant association, a visual assessment of the data will be undertaken using a correspondence analysis to examine in greater detail

the nature of this association. All analyses will be performed using the online software package NCSS2019 (ncss.com, 2019).

For a simple correspondence analysis, consider a $I \times J$ a two-way contingency table, N, where the (i, j)'th cell entry is denoted by n_{ij} for $i = 1,2, \dots, I$ and $j = 1,2, \dots, J$. Let n be the grand total of N and $\boldsymbol{P} = (p_{ij})$ the simple correspondence matrix so that, $p_{ij} = n_{ij}/n$ and $\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1$. The *i-th* row marginal proportion and *j-th* column marginal proportion are defined as $p_{i.} = \sum_{j=1}^{J} p_{ij}$, $p_{.j} = \sum_{i=1}^{I} p_{ij}$ respectively. For a two-way contingency table, the strength of the association between the row and column categories should also be examined, and is done so by considering the model of complete independence $p_{ij} = p_{i.}p_{.j}$ for all i, j. This suggests that complete independence exists at each cell when the observed and expected cell proportions are identical. Complete independence will rarely be satisfied, and so a multiplicative measure of departure from the model of complete independence can be considered, such that $p_{ij} = \alpha_{ij}p_{i.}p_{.j}$. For the model of complete independence, $\alpha_{ij} = 1$ for all $i = 1,2, \dots \dots, I$ and $j = 1,2, \dots \dots, J$. The Pearson chi-squared statistic can be expressed by $X^2 = n\sum_{i=1}^{I}\sum_{j=1}^{J} p_{i.}\, p_{.j}\left(\alpha_{ij} - 1\right)^2$ (Beh, 2004). Therefore, a small chi-squared statistic, which is consistent with the hypothesis of independence, will be achieved when each $\alpha_{ij} = 1$ for all $i = 1,2, \dots \dots, I$ and $j = 1,2, \dots \dots,$.

As n increases by a factor of $C > 1$ so too does the Pearson chi-squared statistic, and this can often hinder tests of association in contingency tables. To avoid this problem, the Pearson chi-squared statistic is divided by n giving X^2/n, which is referred to as the total inertia of the contingency table in simple correspondence analysis (Beh, 2004). The matrix of centred Pearson ratios is $\boldsymbol{\Delta} = \boldsymbol{D_I}^{-1}\boldsymbol{P}\boldsymbol{D_J}^{-1} - \boldsymbol{U}$ where $\boldsymbol{U}$ is a $I \times J$ matrix of 1's (giving the unitary matrix). Note that the *(i, j)–th* cell of $\boldsymbol{\Delta}$ is $\frac{p_{ij}}{p_{i.}p_{.j}} - 1$ and are all zero when there is complete independence between the variables. To obtain a visual summary of the association between *Discipline* and *Country*, their *principal coordinate*s with respect to the (*principal*) axes of the summary are obtained, using a singular value decomposition (SVD) of $\boldsymbol{\Delta}$ so that $\boldsymbol{\Delta} = \boldsymbol{A}\boldsymbol{D}_\lambda\boldsymbol{B}^{\boldsymbol{T}}$ where:

* $\boldsymbol{A}$ is the *I x min (I – 1, J – 1)* column matrix of the left singular vectors of $\boldsymbol{\Delta}$. The matrix $\boldsymbol{A}$ satisfies the equation $\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{D}_I\boldsymbol{A} = \boldsymbol{I}$
* $\boldsymbol{B}$ is a *J x min (I – 1, J – 1)* column matrix of the right singular vectors of $\boldsymbol{\Delta}$**.** The matrix $\boldsymbol{B}$ satisfies the equation $\boldsymbol{B}^{\boldsymbol{T}}\boldsymbol{D}_J\boldsymbol{B} = \boldsymbol{I}$
* $\boldsymbol{D}_\lambda$ is the diagonal matrix of the *min (I – 1, J – 1)* singular vectors of $\boldsymbol{\Delta}$ that are arranged in descending order.

The row principal coordinates are defined by $\boldsymbol{F} = \boldsymbol{A}\boldsymbol{D}_\lambda$ where **F** is a column matrix of size *I x min (I – 1, J – 1)*. Similarly the column principal coordinates are defined by $\boldsymbol{G} = \boldsymbol{B}\boldsymbol{D}_\lambda$ which is a column matrix of size *J x min (I – 1, J – 1)*. For our purposes, we shall be confining our attention to the visual summary of the association between *Discipline* and *Country* by jointly plotting the first two columns $\boldsymbol{F}$ and $\boldsymbol{G}$ yielding a two-dimensional *correspondence plot*. For more on the technical details of simple correspondence analysis, and its numerous variations, see Greenacre (1984) and Beh and Lombardo (2014).

3. RESULTS

3.1. Preliminary overview of Nobel Prize Data

Table 1 shows the *8 x 6* contingency table obtained from the cross-classification of the sample of 785 Nobel Prizes awarded from 1901-2018 (inclusive) according to the variables *Country* and *Discipline*. These variables consist of eight (country) and six (discipline) categories respectively. The number of prizes is reported in the body of the table for each country and discipline pair and in parentheses, the percentage of the total sample that lies within each (*Country*, *Discipline*) pair.

3.2. Chi-squared test of independence

To determine whether there exists a statistically significant association between variables *Country* and *Discipline* in Table 1, a chi-squared test of independence was performed (X^2= 107.02, $p < 0.001$). This result shows that there is enough evidence to suggest that a statistically significant association exists between *Country* and *Discipline*.

To investigate further the nature of this association a simple correspondence analysis of the data presented in Table 1 was performed. This provides a visual representation of this association, which is based on the total inertia, X^2/n. This first requires a decomposition of the total inertia so that we know how many dimensions the visual summary shall consist of. Table 2 shows that a summary of the decomposition of the $8 \; x \; 6$ contingency table of Table 1 can be made by constructing a visual summary of the association (called a *correspondence plot*) consisting of a maximum of 5 dimensions; as derived from min (8 – 1, 6 – 1) = 5. The row labelled *Inertia* in Table 2 contains the X^2/n value accounted for by each category along each component. Of the total inertia (107/785 = 0.14), 62.3% is accounted for by the first component, 17.4% by the second component, and so on; see also the scree plot of Figure 1. Together, these two components explain 80% of the total inertia between *Country* and *Discipline*. Hence the correspondence plot based on these provides an excellent visual summary of association between the different countries and the different disciplines.

Table 1. Contingency table formed from the cross-classification of *Country* by *Discipline*

Country	**Discipline n (row %)**						
	Chemistry	**Economics**	**Literature**	**Medicine**	**Peace**	**Physics**	**Total**
Canada	6 (25%)	4 (17%)	2 (8%)	6 (25%)	1 (4%)	5 (21%)	24 (100%)
France	15 (21%)	3 (4%)	16 (22%)	13 (18%)	12 (16%)	14 (19%)	73 (100%)
Germany	30 (28%)	1 (1%)	10 (9%)	26 (24%)	7 (7%)	34 (32%)	108 (100%)
Italy	1 (5%)	1 (5%)	6 (30%)	6 (30%)	1 (5%)	5 (25%)	20 (100%)
Japan	7 (26%)	0 (0%)	3 (11%)	5 (19%)	1 (4%)	11 (41%)	27 (100%)
Russia	2 (8%)	2 (8%)	4 (16%)	2 (8%)	3 (12%)	12 (48%)	25 (100%)
British Isles	32 (24%)	11 (8%)	11 (8%)	35 (27%)	15 (11%)	28 (21%)	132 (100%)
USA	78 (21%)	61 (16%)	13 (4%)	103 (27%)	21 (6%)	100 (27%)	376 (100%)
Total:							
Count	171	83	65	196	61	209	785
% within countries	(21.8%)	(10.6%)	(8.3%)	(25%)	(7.8%)	(26.6%)	(100%)

3.3. Simple correspondence analysis results

Since the chi-squared test of independence showed that there exists a statistically significant association between *Country* and *Discipline*, the key feature obtained by performing a simple correspondence analysis on Table 1 is the two-dimensional correspondence plot of Figure 2 It shows how particular categories of the row (*Country*) and column (*Discipline*) variables are associated with each other. We can also identify specific categories that provide a strong or weak contribution to the association and those categories that are well explained, or not, by the correspondence plot.

Figure 2 shows that that the USA and Canada appear to provide a similar contribution to the association between *Country* and *Discipline* since they are in close proximity to one another. It also shows that these two countries dominate Nobel Prizes that are awarded in Medicine; this may seem evident due to the close scientific, and geological, ties between the two countries. Although it is not because of the relatively large joint population of the USA and Canada (among those countries studied); simple correspondence analysis deals with the analysis of proportional allocations instead of raw frequencies.

Figure 2 also shows that Nobel Literature laureates are strongly associated with Italy and France. The *Disciplines* of Chemistry and Physics appear to provide a similar contribution to the association since they lie close to one another in Figure 2. However, their close proximity to the origin, and that of Medicine, suggests that these three disciplines contribute relatively little to the association when compared with other Nobel Prize *Disciplines*. On the other hand, since Economics and Literature are a relatively long way from the origin, their distance suggests that these are the two most dominant *Disciplines* that define the association between the two variables. In terms of the countries analysed, a similar conclusion can also be made about Italy and Japan. Note that along the two dimensions of Figure 2 are two percentage values; 62.3% and 17.4%. These are the percentage contribution that the first two components from the decomposition of the total inertia provide. As implied in Section 3.2, this plot summarises about 80% of the association (numerically measured using the total inertia) between *Country* and *Discipline*.

Table 2. Summary of the features from the decomposition of the total inertia of Table 1

Rows:

	CAN	FRA	DEU	ITA	JPN	RUS	BI	USA
Mass	0.0306	0.0930	0.1376	0.0255	0.0344	0.0318	0.1682	0.4790
ChiDist	0.2630	0.6347	0.3421	0.8657	0.4859	0.6895	0.1900	0.2594
Inertia	0.0021	0.0375	0.0161	0.0191	0.0081	0.0151	0.0061	0.0322
Dim. 1	-0.5026	2.0229	0.5933	2.1271	0.7896	1.2851	0.2434	-0.8718
Dim. 2	0.6956	1.3198	-1.8524	1.7421	-2.5876	-0.2525	0.2259	0.2621

Columns:

	Chm	Eco	Lit	Med	Pce	Phy
Mass	0.2178	0.1057	0.0828	0.2497	0.0777	0.2662
ChiDist	0.2091	0.5892	0.7897	0.1761	0.4712	0.2248
Inertia	0.0095	0.0367	0.0516	0.0077	0.0173	0.0135
Dim. 1	-0.0975	-1.7010	2.5814	-0.4076	1.1271	0.0058
Dim. 2	-0.7901	1.9242	1.1543	0.1682	1.1555	-0.9717

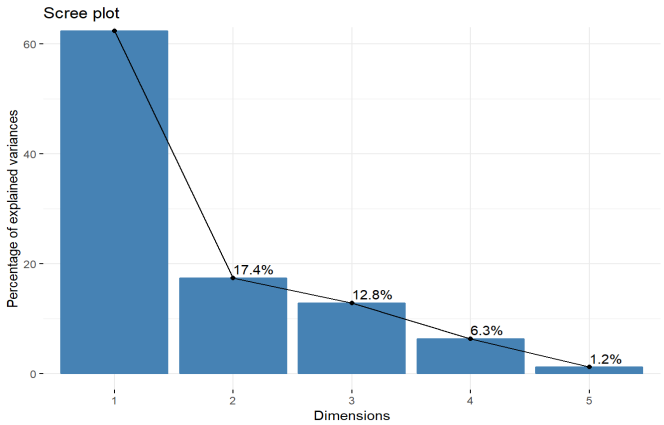

Figure 1. The scree plot of the weight of the five components from the correspondence analysis of Table 1.

To provide a more numerical description of the findings summarised in Figure 2, consider the following points based on the results given in Table 3. Table 3 summarises the quality (*qlt*) of visual representation made in the correspondence plot derived from the analysis of Table 1. Recall that the maximum number of dimensions needed for such an analysis is 5. So a two-dimensional plot will provide a good visual summary of some categories and a poor summary for others. For the two-dimensional correspondence plot, the column labelled *qlt* has, for the rows (countries) USA and FRA, *qlt* = 984 and 966, respectively. Therefore, the two-dimensional correspondence plot of Figure 2 depicts 98%, and 97%, of the contribution that the USA and France, respectively, make to the association between *Country* and *Discipline*. While these are the best represented categories in the two-dimensional plot, the country with the poorest quality of representation in Table 2 is that of British Isles (*qlt* = 173); meaning that the first two dimensions represent only 17% of the contribution British Isles makes to the association; the remaining 83% of the association can be accounted for by the third, fourth and fifth components (dimensions) which are not described here.

The column labelled *Cor* in Table 3 represents the correlation (or contribution) that each row and column combination of Table 1 makes to the first and second components of Figure 2. Consider, for example, the country USA; note that the sum of *cor* along the first two components is 960 + 24 = 984 which is equivalent to its overall *qlt* value. Thus, the position of the USA in a two-dimensional plot is heavily dominated by the first dimension while its coordinate along the second dimension is close to zero; this is certainly the case by observing the position of the USA in Figure 2. So, the first dimension (K = 1) of Figure 2 is dominated by USA, France and Italy (*cor* of 960, 863 and 513, respectively). Similarly, the second dimension (K = 2) is

dominated by Germany and Japan (*cor* of 696 and 673 respectively); this can be seen in Figure 2 which shows that both dimensions dominate the position of Italy in Figure 2.

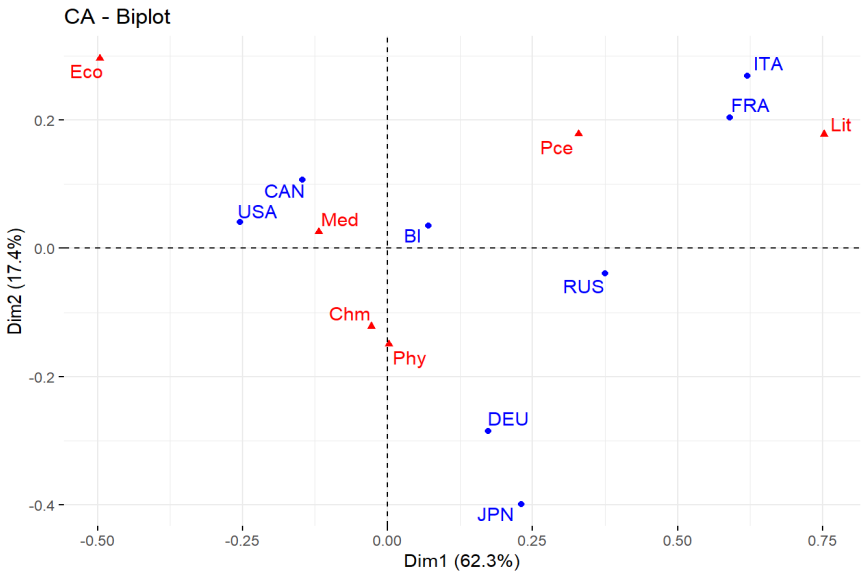

Figure 2. Two-dimensional correspondence plot of Table 1

Table 3. *Country* and *Discipline* Contributions to Figure 2

Rows:

name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
CAN	31	476	16	-147	310	8	107	166	15
FRA	93	966	275	590	863	381	203	103	162
DEU	138	951	118	173	256	48	-285	696	472
ITA	25	609	140	620	513	115	268	96	77
JPN	34	897	60	230	224	21	-399	673	230
RUS	32	298	111	375	295	53	-39	3	2
BI	168	173	45	71	139	10	35	34	9
USA	479	984	236	-254	960	364	40	24	33

Columns:

name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
Chm	218	357	70	-28	18	2	-122	339	136
Eco	106	961	269	-496	708	306	296	253	391
Lit	83	958	379	752	908	552	178	51	110
Med	250	477	57	-119	455	41	26	22	7
Pce	78	629	127	329	486	99	178	143	104
Phy	266	443	99	2	0	0	-150	443	251

A visual summary of how well each *Country* contributes to the quality of the two-dimensional correspondence plot of Figure 2 can be made by considering Figure 3. Those countries in yellow to red show an increase in the quality that the correspondence plot provides for these categories. For example, it shows that the USA, France and Japan and all colored toward the red end of the spectrum highlighting that Figure 2 provides a very good summary of these countries; we saw that this was the case by commenting on their *qlt* value. On the other hand, countries with colors ranging from light yellow to blue reflect an increasingly poor-quality representation of Figure 2, meaning that the addition of a third (or more) dimension is necessary to obtain a good visualisation of their contribution to the association. Figure 3 shows that the contribution Russia and, in particular, British Isles (with a *qlt* of 298 and 173 respectively) make to the association are not well captured by a two-dimensional correspondence plot. Thus, since Canada (in yellow) is poorly represented in Figure 2, this suggests that the USA and Canada do not, in fact, have a as strong an associative link to Medicine as we expected.

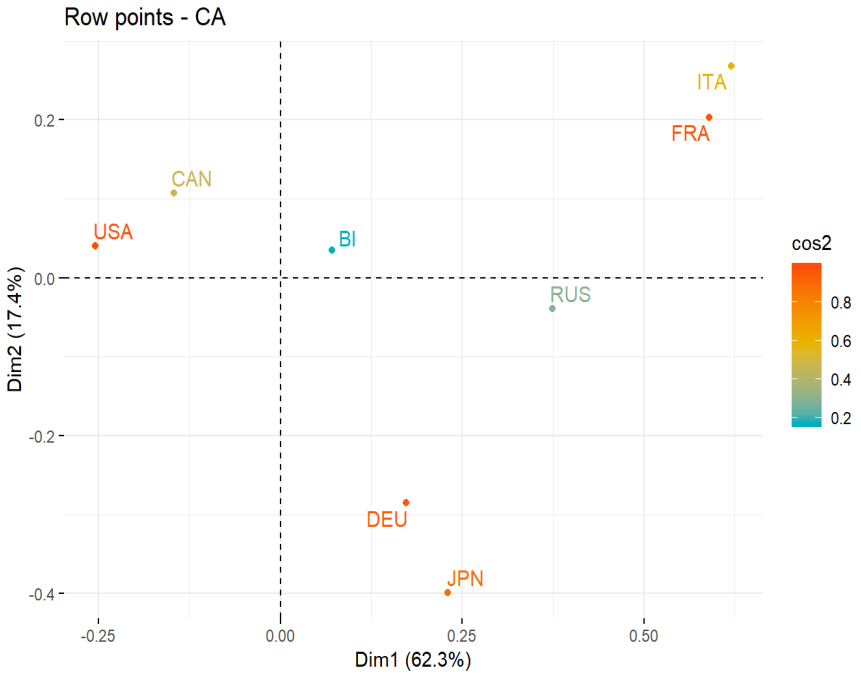

Figure 3. Quality of representation given by Figure 2 for the eight countries studied

4. CONCLUSION

These analyses have demonstrated the utilisation of correspondence analyses to help explore the association between discipline and recipient countries of the Nobel Prize using data from 1901 to 2018. While some countries appear to be more commonly associated with awards in some disciplines, correspondence analyses enabled the associations to be explored in much more detail than is able using standard tests of association. Correspondence analysis also provides an intuitive summary of the association between *Country* and *Discipline*. Future research on the data can be undertaken to further develop the links between clustering, modelling, inferential statistics and other topics commonly utilised for analysing categorical data.

VISUALISATION UTILISATION REFERENCES

Beh, E. J. (2004), Simple correspondence analysis: A bibliographic review, *International Statistical Review*, 72, 257 – 284.

Beh, E. J. and Lombardo, R. (2014). *Correspondence Analysis: Theory, Practice and New Strategies*, Wiley, Chichester.

Encyclopedia Britannica. (2019). Nobel Prize | Definition, History, Winners, & Facts. [online] Available at: https://www.britannica.com/topic/Nobel-Prize [Accessed 24 Apr. 2019].

Greenacre, M. (2001). Tying Up the Loose Ends in Simple Correspondence Analysis. *SSRN Electronic Journal, http://dx.doi.org/10.2139/ssrn.1001889.*

Morais, F. (2018). Vision and the Nobel Prize. *Arquivos Brasileiros de Oftalmologia*, 81(2), 161 – 165.

Ncss.com. (2019). Statistical Software | Sample Size Software | NCSS. [online] Available at: https://www.ncss.com/ [Accessed 24 April. 2019].

Nilesh, B. and Pranav, B. (2018). Statistical analysis of Nobel Prizes in physics: from its inception till date. *Journal of Physical Studies*, 22(3), (8 pages).

NobelPrize.org. (2019). All Nobel Prizes. [online] Available at: https://www.nobelprize.org/prizes/lists/all-nobel-prizes/ [Accessed 24 Apr. 2019].

The Nobel Peace Prize. (2019). Alfred Nobel's fortune. [online] Available at: https://www.nobelpeaceprize.org/History/Alfred-Nobel-s-fortune [Accessed 24 Apr. 2019].

Levinovitz, A. and Ringertz, N. (2001). *The Nobel Prize*. London: Imperial College Press.