Correspondence Analysis for Studying Global Malaria Mortality

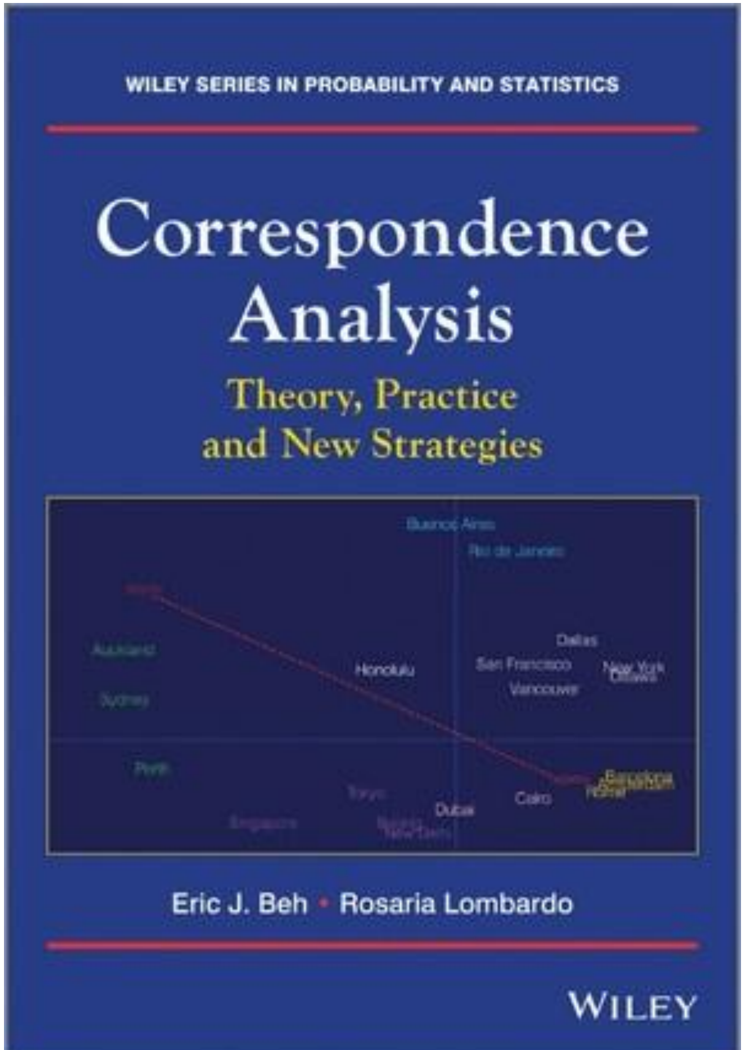
Eric J. Beh

School of Mathematics & Physical Sciences
University of Newcastle, Australia
eric.beh@newcastle.edu.au

London School of Hygiene & Tropical Medicine
University College London, 29th January, 2016

Brief History of Correspondence Analysis

Our 2014 Book



Correspondence Analysis: Theory, Practice and New Strategies

Eric J. Beh, Rosaria Lombardo

ISBN: 978-1-119-95324-1

576 pages
October 2014

Rosaria Lombardo

Department of Economics,
Second University of Naples, Italy
rosaria.lombardo@unina2.it

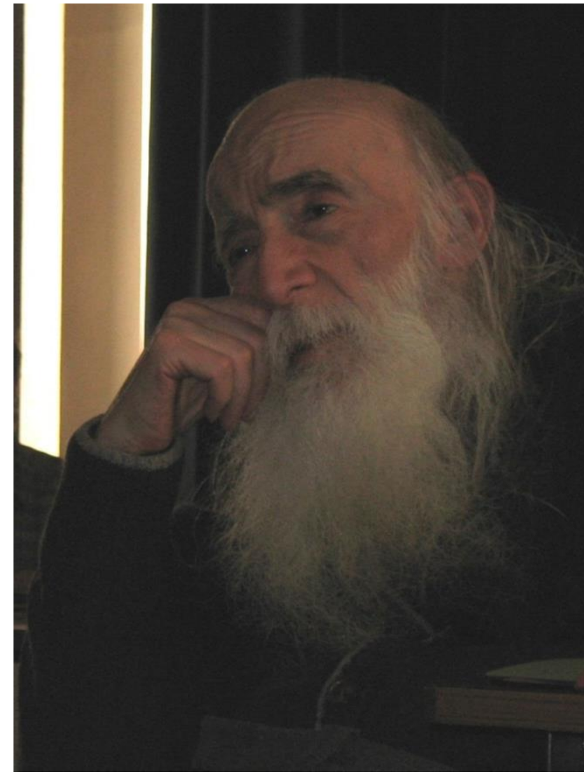
http://au.wiley.com/WileyCDA/WileyTitle/productCd-1119953243.html

Brief History of Correspondence Analysis

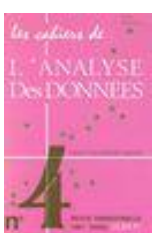
Jean-Paul Benzécri

The 1960’s saw the advances in categorical data analysis take on a geometric form with the development of correspondence analysis.

The “father” of modern day correspondence analysis is French linguist Jean-Paul Benzécri, and with his team of researchers, developed its foundations at the Mathematical Statistics Laboratory, Faculty of Science in Paris, France.



Jean-Paul Benzécri
Paris, 2011



As a result the method of *l’analyse des correspondances,* as coined by Benzécri, is very popular in France not just among statisticians, but among researchers from most disciplines in the country. The popularity of correspondence analysis in France resulted in a journal dedicated to correspondence analysis, *Cahiers de l’Analyse des Données*, founded by Benzecri (1976 – 1997). See the journal’s online site http://www.numdam.org/numdam-bin/browse?j=CAD&sl=0

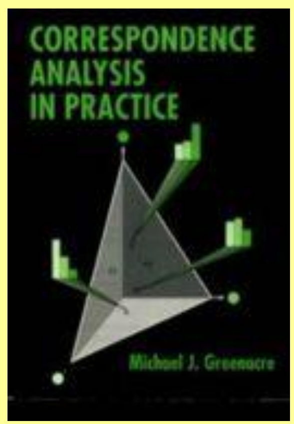
Brief History of Correspondence Analysis

Michael Greenacre

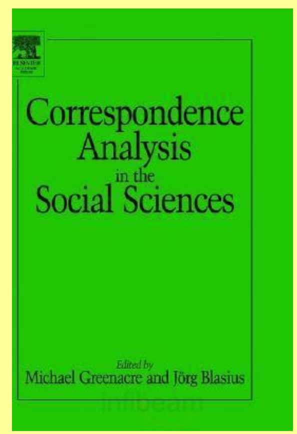
- Michael Greenacre
  Universitat Pompeu Fabra, Barcelona, Spain
- Former student of Benzecri

- Greenacre, M. J. (1978), Quelques methodes objectives de representation graphique d'un tableau de donnes, Unpublished PhD thesis, Universite Pierre et Marie Curie, Paris
  - Rough translation: *Some objective methods for the graphical representation of tabular data*

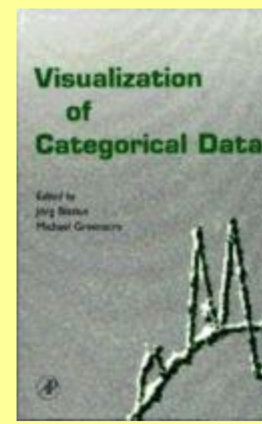
1984



1993



1994



1998

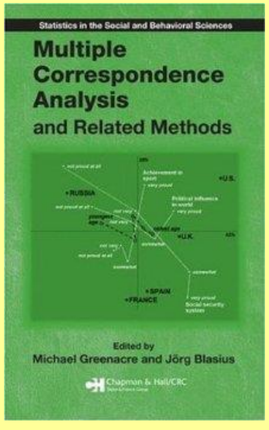
You are also invited to consider Beh and Lombardo (2012) who provide an extensive bibliography on the history and development of correspondence analysis up to 2012.

Brief History of Correspondence Analysis

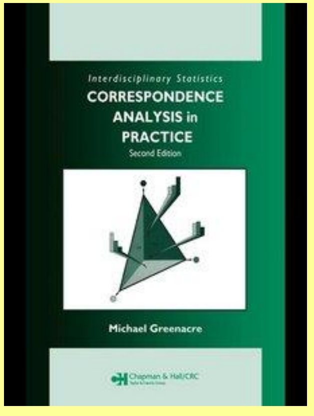
Michael Greenacre

- Michael Greenacre
  Universitat Pompeu Fabra, Barcelona, Spain
- Former student of Benzécri

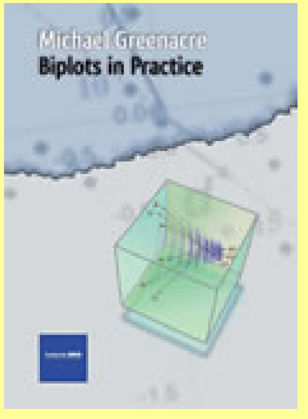
- Greenacre, M. J. (1978), Quelques methodes objectives de representation graphique d'un tableau de donnes, Unpublished PhD thesis, Universite Pierre et Marie Curie, Paris
  - Rough translation: *Some objective methods for the graphical representation of tabular data*



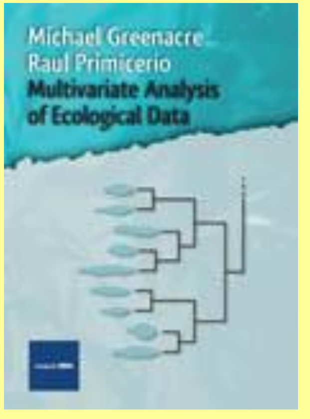
2006



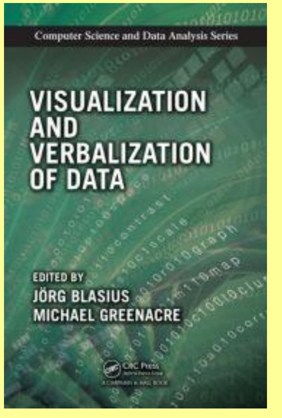
2007



2010



2013

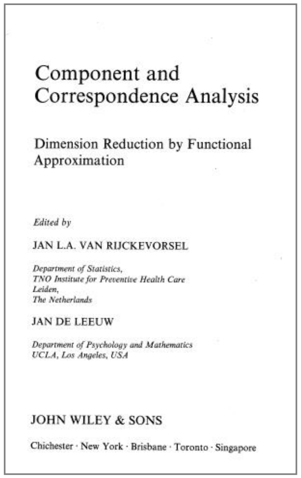


2014

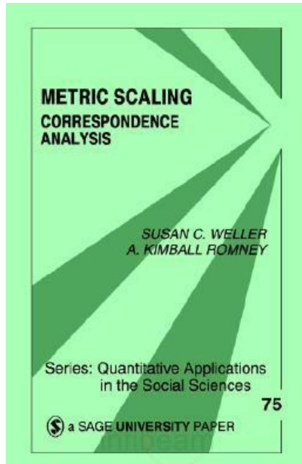
You are also invited to consider Beh and Lombardo (2012) who provide an extensive bibliography on the history and development of correspondence analysis up to 2012.

Brief History of Correspondence Analysis

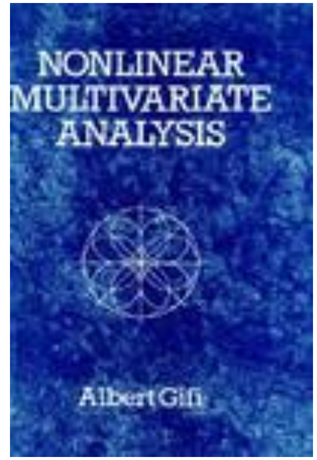
Some Key Books



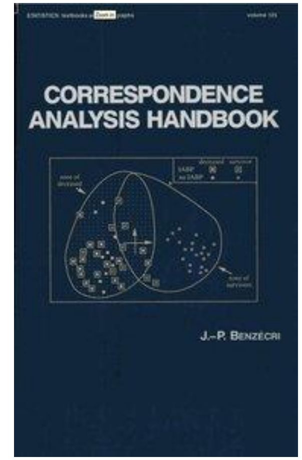
1988



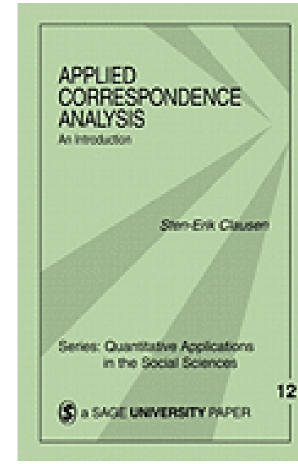
1990



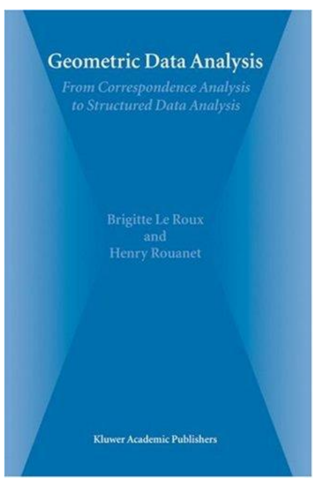
1991



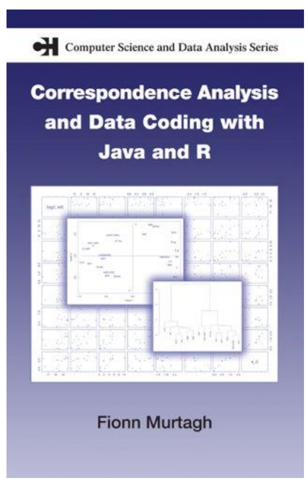
1992



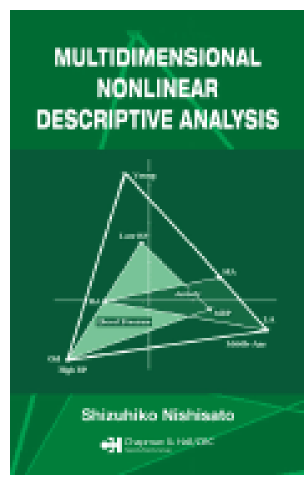
1998



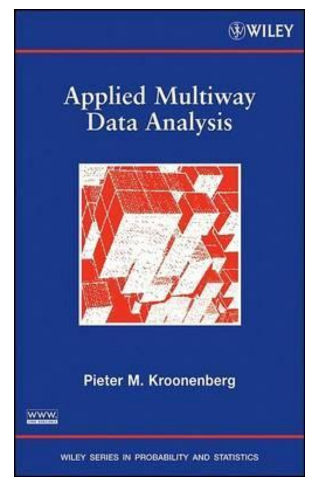
2004



2005



2007



2008

2010

What is Correspondence Analysis?

Suppose of n individuals/countries/things that are summarised according to two (A and B) variables which are cross-classified to form a contingency table

A/B	B_1	B_2	$\cdots$	B_j	$\cdots$	B_J	Total
A_1	n_{11}	n_{12}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{1\bullet}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_I	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{I\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet J}$	n

Basically, correspondence analysis is a way of **visualising the association** between categorical variables using as few dimensions as possible

What is Correspondence Analysis?

Suppose we have two (A and B) or more categorical variables and they are cross-classified to form a contingency table

A/B	B_1	B_2	$\cdots$	B_j	$\cdots$	B_J	Total
A_1	n_{11}	n_{12}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{1\bullet}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_I	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{I\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet J}$	n

From the analysis, we can visualise

- How similar, or different, categories from the same variable are to each other
- the association between the variables being studied

The Basic Idea of Correspondence Analysis

Suppose we consider the row and column categories of an IxJ contingency table.

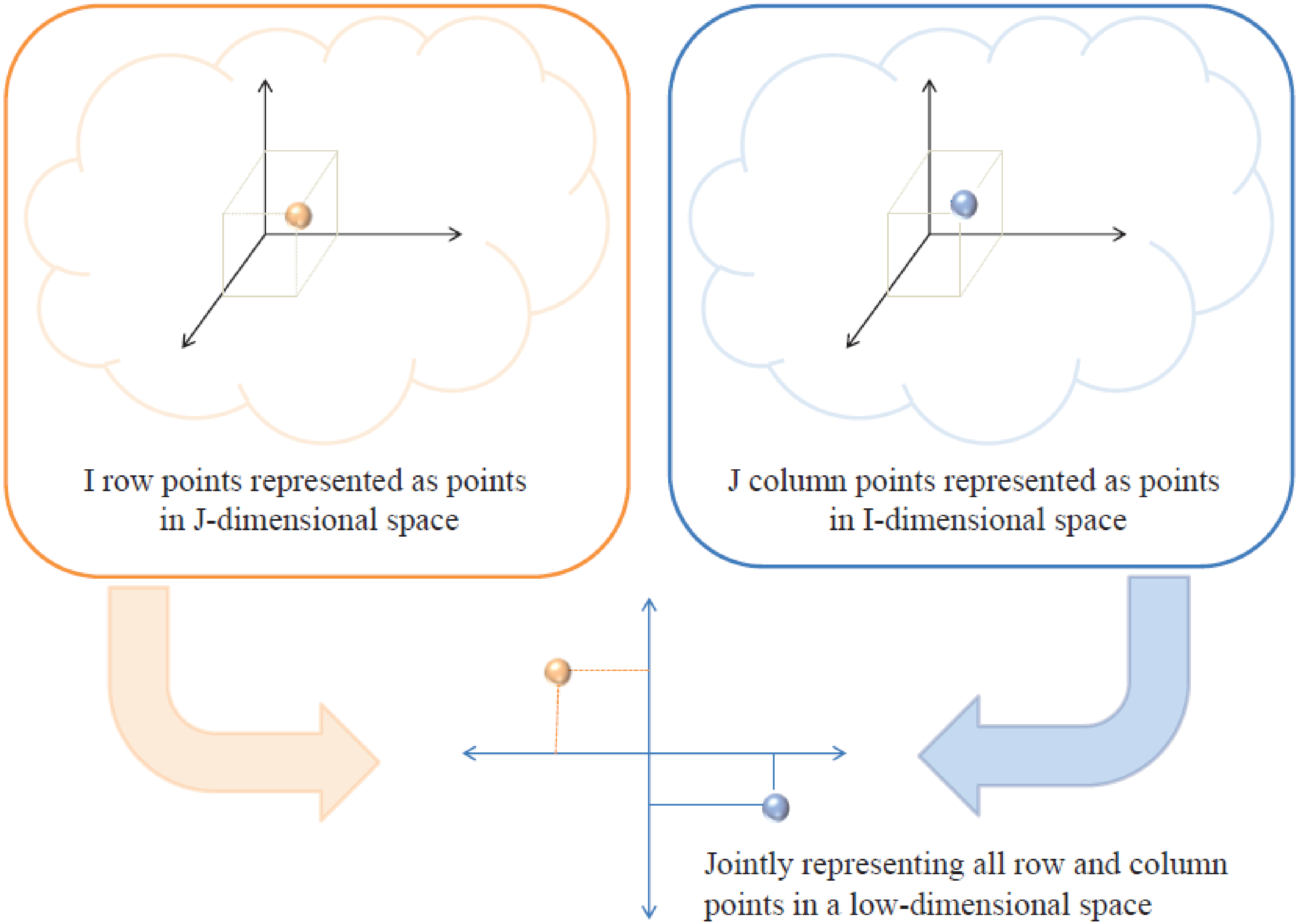
A/B	B_1	B_2	$\cdots$	B_j	$\cdots$	B_J	Total
A_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	$n_{1\bullet}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_I	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{I\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet J}$	n

- Each of the I rows can be thought of as a point in J – dimensional space

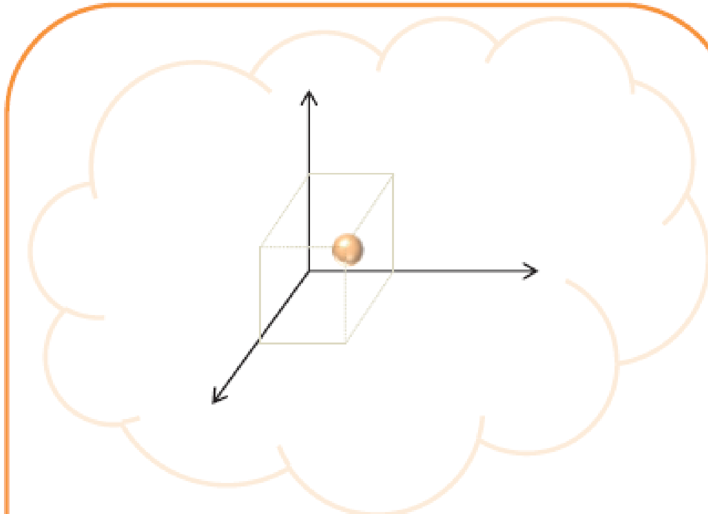
- Each of the J columns can be thought of as a point in I – dimensional space

The Basic Idea of Correspondence Analysis

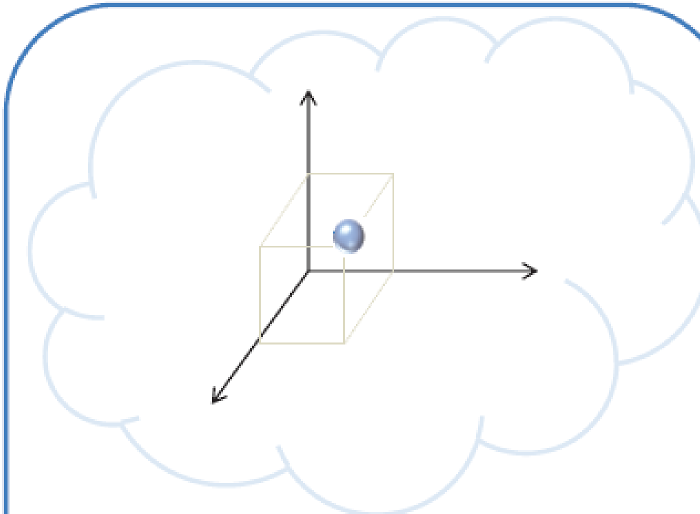
Clouds of Points



The Basic Idea of Correspondence Analysis



I row points represented as points
in J-dimensional space



J column points represented as points
in I-dimensional space

The data reduction is achieved through a variety means. Commonly . . .

- *singular value decomposition* is applied to a transformation of the data

Good news: We will skip all of the mathematical derivations

Global Study of Malaria

Some Facts

Some facts I have learned about malaria:

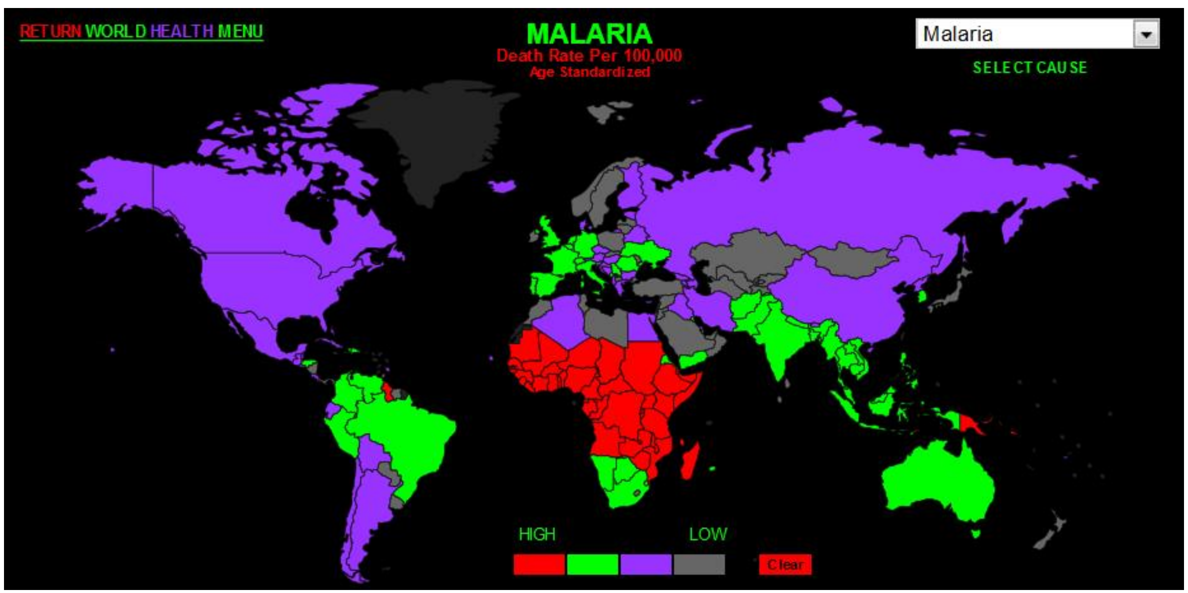
- Infectious disease caused by parasitic protozoans that belong to the genus Plasmodium
- Malaria is carried and transmitted through mosquito bites
- Symptoms include fever, fatigue, vomiting and headaches
- In severe cases it can cause seizures, coma or death
- Most deaths are caused by the Plasmodian strands *P. falciparum*
- Milder forms of malaria are caused by *P. vivax*, *P. ovale*, and *P. malariae*
- The strand *P. knowlesi rarely causes disease in humans*

Global Study of Malaria

Global Problem

Some facts I have learned about malaria:

- The disease is widespread in tropical and subtropical regions
- Widespread in sub-Saharan Africa, certain parts of Asia and Latin America



Rank	Country	Rate
1	CENTRAL AFRICA	75.61
2	CHAD	74.18
3	CONGO	70.41
4	SIERRA LEONE	69.56
5	GAMBIA	63.96
6	COMOROS	63.90
7	GUINEA-BISSAU	62.71
8	TOGO	62.09
9	GUINEA	61.91
10	GABON	61.72
11	GHANA	61.63
12	BURKINA FASO	61.39
13	NIGERIA	60.46
14	NIGER	58.61
15	DR CONGO	57.64
16	ANGOLA	56.81
17	BENIN	56.00
18	EQU. GUINEA	55.80
19	SENEGAL	55.51
20	MAURITANIA	53.04

http://www.worldlifeexpectancy.com/cause-of-death/malaria/by-country/

Global Study of Malaria

The Epidemic

To continue on with the rest of this story . . .

- WHO reported
  - 219 million cases of malaria in 2010
  - 212 million cases in 2012
  - 198 million cases in 2013
- WHO estimated that there were
  - About 660,000 deaths in 2010
  - were between 584000 to 855000 deaths 2013
- 90% of these deaths were in Sub-Saharan Africa
- 65% of deaths are children under the age of 15 years
- Trends indicate that up to 200,000 deaths of maternal malaria per year were seen in sub-Saharan Africa

WIKIPEDIA
The Free Encyclopedia

In 2014, Bill Gates announced that his foundation will be investing more than US$500million to reduce the prevalence of malaria, pneumonia, diarrheal diseases and other parasitic infections that are the leading cause of death and disability in developing countries.

BILL & MELINDA GATES foundation

Global Study of Malaria

This presentation will focus on the results published in

Murray CJL, Rosenfeld LC, Lim SS, Andrews KG, Foreman KJ, Haring D, Fullman N, Naghavi M, Lozano R and Lopez AD (2012), Global malaria mortality between 1980 and 2010: a systematic analysis, *The Lancet*, 379, 413 – 431.

- The authors study focus on the P.falciparum strand of malaria
- Was funded by the Bill & Melinda Gates Foundation

Study by researchers at the

- Institute for Health Metrics and Evaluation, University of Washington
- School of Population Health, University of Queensland

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

Global Study of Malaria

The authors summarise the *estimated* number of deaths due to malaria

- by age, gender and country
- in 105 countries (dominated by the sub-Saharan, Asian and Latin American regions)
- in 1980, 1990, 2000 and 2010.

Their estimates are based on

- Published and unpublished verbal autopsy studies
- Of these, population based studies that covered a period of at least 12 months
- They took into account other factors
  - Standardised age brackets where interval lengths were variable
  - Changes in the International Classification of Diseases and Injuries
  - The authors claim to model death's although they provide standardised summarises and present confidence intervals of number of deaths based on these standardisations

The Data

Rather than studying all 105 countries we shall focus on

- the 20 countries that have the highest death rate (per 100,000)
- Initially, infant malaria deaths in these countries (children less than 5 years of age)

Collectively, these 20 countries

- saw n = 1,607,161 infant deaths due to malaria (P.falcipalum)
- saw n = 579,309 individuals at least 5 years of age die of malaria
- All countries saw moderate to large increases in the number of infant deaths between 1980 and 2000
- most countries experienced moderate to large declines in the number of infant deaths between 2000 and 2010

Rank	Country	Rate	Ra
1	CENTRAL AFRICA	75.61	
2	CHAD	74.18	
3	CONGO	70.41	
4	SIERRA LEONE	69.56	
5	GAMBIA	63.96	
6	COMOROS	63.90	
7	GUINEA-BISSAU	62.71	
8	TOGO	62.09	
9	GUINEA	61.91	
10	GABON	61.72	
11	GHANA	61.63	
12	BURKINA FASO	61.39	
13	NIGERIA	60.46	
14	NIGER	58.61	
15	DR CONGO	57.64	
16	ANGOLA	56.81	
17	BENIN	56.00	
18	EQU. GUINEA	55.80	
19	SENEGAL	55.51	
20	MAURITANIA	53.04	

Est. Number of Deaths: Children < 5yrs

Country	1980	1990	2000	2010
Angola	4916	7241	13652	7777
Benin	4554	5939	8634	8251
Burkina Faso	9037	13305	28211	24656
Central African Republic	2257	3584	6676	5072
Chad	4194	5305	9776	9997
Comoros	152	200	359	235
Congo	1193	1746	3001	1869
Democratic Republic of Congo	31294	62676	108311	69505
Equatorial Guinea	174	516	564	310
Gabon	183	397	465	272
Gambia	665	971	1334	1594
Ghana	10335	14060	15560	10575
Guinea	8532	11099	17868	14208
Guinea-Bissau	1853	2233	2447	2678
Mauritania	307	393	810	758
Niger	6949	9735	16123	22984
Nigeria	130405	192945	304897	266429
Senegal	2917	4359	7939	4085
Sierra Leone	5978	7777	14101	8516
Togo	2982	3868	4987	4449

Table 1: Estimated number of infant deaths due to malaria (P.falciparum): children < 5 years

Est. Number of Deaths: Children < 5yrs

Country	1980	1990	2000	2010
Angola	[illegible]	[illegible]	[illegible]	[illegible]
Benin				
Burkina Faso				
Central African Republic				
Chad				
Comoros				
Congo				
Democratic Republic of Congo				
Equatorial Guinea				
Gabon				
Gambia				
Ghana				
Guinea				
Guinea-Bissau				
Mauritania				
Niger				
Nigeria				
Senegal				
Sierra Leone	5978	7777	14101	8516
Togo	2982	3868	4987	4449

Different countries have different levels of mortality over the 30 year period. This is because

- The population of each country is different
- The mortality *rate* is different for each country

We shall examine

- How the distribution of deaths compare for each country over the 30 year period
- What countries have a similar/different mortality distribution?

So we are interested in exploring malaria mortality by examining the association between these 20 countries and four time periods.

Table 1: Estimated number of infant deaths due to malaria (P.falciparum): children < 5 years

Est. Number of Deaths: Children < 5yrs

The Contingency Table

Country	1980	1990	2000	2010
Angola	4916	7241	13652	7777
Benin	4554	5939	8634	8251
Burkina Faso	9037	13305	28211	24656
Central African Republic	2257	3584	6676	5072
Chad	4194	5305	9776	9997
Comoros	152	200	359	235
Congo	1193	1746	3001	1869
Democratic Republic of Congo	31294	62676	108311	69505
Equatorial Guinea	174	516	564	310
Gabon	183	397	465	272
Gambia	665	971	1334	1594
Ghana	10335	14060	15560	10575
Guinea	8532	11099	17868	14208
Guinea-Bissau	1853	2233	2447	2678
Mauritania	307	393	810	758
Niger	6949	9735	16123	22984
Nigeria	130405	192945	304897	266429
Senegal	2917	4359	7939	4085
Sierra Leone	5978	7777	14101	8516
Togo	2982	3868	4987	4449

Each of the 20 rows can be thought of as a point in 4 dimensional space

Each of the 4 columns can be thought of as a point in 20 dimensional space

Table 1: Estimated number of infant deaths due to malaria (P.falciparum): children < 5 years

Est. Number of Deaths: Children < 5yrs

The Contingency Table

Country	1980	1990	2000	2010	% Diff ('00 -'10)
Angola	4916	7241	13652	7777	**-43.0%**
Benin	4554	5939	8634	8251	-4.4%
Burkina Faso	9037	13305	28211	24656	-12.6%
Central African Republic	2257	3584	6676	5072	**-24.0%**
Chad	4194	5305	9776	9997	2.3%
Comoros	152	200	359	235	**-34.5%**
Congo	1193	1746	3001	1869	**-37.7%**
Democratic Republic of Congo	31294	62676	108311	69505	**-35.8%**
Equatorial Guinea	174	516	564	310	**-45%**
Gabon	183	397	465	272	**-41.5%**
Gambia	665	971	1334	1594	**19.5%**
Ghana	10335	14060	15560	10575	**-32.0%**
Guinea	8532	11099	17868	14208	**-20.5%**
Guinea-Bissau	1853	2233	2447	2678	9.4%
Mauritania	307	393	810	758	-6.4%
Niger	6949	9735	16123	22984	**42.6%**
Nigeria	130405	192945	304897	266429	-12.6%
Senegal	2917	4359	7939	4085	**-48.5%**
Sierra Leone	5978	7777	14101	8516	**-39.6%**
Togo	2982	3868	4987	4449	-10.8%

Table 1: Estimated number of infant deaths due to malaria (P.falciparum): children < 5 years

Est. Number of Deaths: Children < 5yrs

The Contingency Table

Country	1980	1990	2000	2010	% Diff ('00 -'10)
Angola	4916	7241	13652	7777	**-43.0%**
Benin	4554	5939	8634	8251	-4.4%
Burkina Faso	9037	13305	28211	24656	-12.6%
Niger	6949	9735	16123	22984	**42.6%**
Nigeria	130405	192945	304897	266429	-12.6%
Senegal	2917	4359	7939	4085	**-48.5%**
Sierra Leone	5978	7777	14101	8516	**-39.6%**
Togo	2982	3868	4987	4449	-10.8%

Generally

- There was an increase in the number of deaths between 1980 and 2000
- For most countries, the number of deaths decreased between 2000 and 2010 (except for Gambia and Niger)

Let's look at the relative distribution of each country over this 30 year period

(Note: we could also do similar comparisons of the *relative distribution of each time period*, but – for brevity – we shall consider countries here)

Table 1: Estimated number of infant deaths due to malaria (P.falciparum): children < 5 years

Est. Number of Deaths: Children < 5yrs

Relative Proportions

Country	1980	1990	2000	2010
Angola	14.64%	21.56%	40.65%	23.16%
Benin	16.63%	21.69%	31.54%	30.14%
Burkina Faso	12.02%	17.69%	37.51%	32.78%
Central African Republic	12.83%	20.38%	37.96%	28.84%
Chad	14.33%	18.12%	33.40%	34.15%
Comoros	16.07%	21.14%	37.95%	24.84%
Congo	15.28%	22.36%	38.43%	23.93%
Democratic Republic of Congo	11.51%	23.06%	39.85%	25.57%
Equatorial Guinea	11.13%	32.99%	36.06%	19.82%
Gabon	13.90%	30.14%	35.31%	20.65%
Gambia	14.57%	21.28%	29.23%	34.93%
Ghana	20.45%	27.83%	30.79%	20.93%
Guinea	16.50%	21.47%	34.56%	27.48%
Guinea-Bissau	20.12%	24.24%	26.57%	29.07%
Mauritania	13.54%	17.33%	35.71%	33.42%
Niger	12.46%	17.45%	28.90%	41.20%
Nigeria	14.58%	21.57%	34.08%	29.78%
Senegal	15.11%	22.59%	41.13%	21.17%
Sierra Leone	16.44%	21.38%	38.77%	23.41%
Togo	18.31%	23.75%	30.62%	27.32%

Similar distribution of malaria mortality

Similar distribution of malaria mortality

Different distribution of malaria mortality

Distribution of infant malaria deaths across time (for each country): children < 5 years

Est. Number of Deaths: Children < 5yrs

Relative Proportions

Country	1980	1990	2000	2010
Angola	14.64%	21.56%	40.65%	23.16%
Benin	16.63%	21.69%	31.54%	30.14%
Burkina Faso	12.02%	17.69%	37.51%	32.78%
Central African Republic	12.83%	20.38%	37.96%	28.84%
Chad	14.33%	18.12%	33.40%	34.15%
Comoros	16.07%	21.14%	37.95%	24.84%
Congo	15.28%	22.36%	38.43%	23.93%
Democratic Republic of Congo	11.51%	23.06%	39.85%	25.57%
Equatorial Guinea	11.13%	32.99%	36.06%	19.82%
Gabon	13.90%	30.14%	35.31%	20.65%
Gambia	14.57%	21.28%	29.23%	34.93%
Ghana	20.45%	27.83%	30.79%	20.93%
Guinea	16.50%	21.47%	34.56%	27.48%
Guinea-Bissau	20.12%	24.24%	26.57%	29.07%
Mauritania	13.54%	17.33%	35.71%	33.42%
Niger	12.46%	17.45%	28.90%	41.20%
Nigeria	14.58%	21.57%	34.08%	29.78%
Senegal	15.11%	22.59%	41.13%	21.17%
Sierra Leone	16.44%	21.38%	38.77%	23.41%
Togo	18.31%	23.75%	30.62%	27.32%

1980 only

Relatively low malaria mortality

(possible weak association)

Relatively high malaria mortality

(possible strong association)

Distribution of infant malaria deaths across time (for each country): children < 5 years

Est. Number of Deaths: Children < 5yrs

Relative Proportions

Country	1980	1990	2000	2010
Angola	14.64%	21.56%	40.65%	23.16%
Benin	16.63%	21.69%	31.54%	30.14%
Burkina Faso	12.02%	17.69%	37.51%	32.78%
Central African Republic	12.83%	20.38%	37.96%	28.84%
Chad	14.33%	18.12%	33.40%	34.15%
Comoros	16.07%	21.14%	37.95%	24.84%
Congo	15.28%	22.36%	38.43%	23.93%
Democratic Republic of Congo	11.51%	23.06%	39.85%	25.57%
Equatorial Guinea	11.13%	32.99%	36.06%	19.82%
Gabon	13.90%	30.14%	35.31%	20.65%
Gambia	14.57%	21.28%	29.23%	34.93%
Ghana	20.45%	27.83%	30.79%	20.93%
Guinea	16.50%	21.47%	34.56%	27.48%
Guinea-Bissau	20.12%	24.24%	26.57%	29.07%
Mauritania	13.54%	17.33%	35.71%	33.42%
Niger	12.46%	17.45%	28.90%	41.20%
Nigeria	14.58%	21.57%	34.08%	29.78%
Senegal	15.11%	22.59%	41.13%	21.17%
Sierra Leone	16.44%	21.38%	38.77%	23.41%
Togo	18.31%	23.75%	30.62%	27.32%

1990 only

Relatively low malaria mortality

(possible weak association)

Relatively high malaria mortality

(possible strong association)

Distribution of infant malaria deaths across time (for each country): children < 5 years

Est. Number of Deaths: Children < 5yrs

Relative Proportions

Country	1980	1990	2000	2010
Angola	14.64%	21.56%	40.65%	23.16%
Benin	16.63%	21.69%	31.54%	30.14%
Burkina Faso	12.02%	17.69%	37.51%	32.78%
Central African Republic	12.83%	20.38%	37.96%	28.84%
Chad	14.33%	18.12%	33.40%	34.15%
Comoros	16.07%	21.14%	37.95%	24.84%
Congo	15.28%	22.36%	38.43%	23.93%
Democratic Republic of Congo	11.51%	23.06%	39.85%	25.57%
Equatorial Guinea	11.13%	32.99%	36.06%	19.82%
Gabon	13.90%	30.14%	35.31%	20.65%
Gambia	14.57%	21.28%	29.23%	34.93%
Ghana	20.45%	27.83%	30.79%	20.93%
Guinea	16.50%	21.47%	34.56%	27.48%
Guinea-Bissau	20.12%	24.24%	26.57%	29.07%
Mauritania	13.54%	17.33%	35.71%	33.42%
Niger	12.46%	17.45%	28.90%	41.20%
Nigeria	14.58%	21.57%	34.08%	29.78%
Senegal	15.11%	22.59%	41.13%	21.17%
Sierra Leone	16.44%	21.38%	38.77%	23.41%
Togo	18.31%	23.75%	30.62%	27.32%

Distribution of infant malaria deaths across time (for each country): children < 5 years

Burkina Faso has relatively low malaria mortality for 1980 and 1990

(possible weak association)

Ghana has a relatively high malaria mortality for 1980 and 1990

(possible strong association)

Est. Number of Deaths: Children < 5yrs

Relative Proportions

Country	1980	1990	2000	2010
Angola	14.64%	21.56%	40.65%	23.16%
Benin	16.63%	21.69%	31.54%	30.14%
Burkina Faso	12.02%	17.69%	37.51%	32.78%
Central African Republic	12.83%	20.38%	37.96%	28.84%
Chad	14.33%	18.12%	33.40%	34.15%
Comoros	16.07%	21.14%	37.95%	24.84%
Congo	15.28%	22.36%	38.43%	23.93%
Democratic Republic of Congo	11.51%	23.06%	39.85%	25.57%
Equatorial Guinea	11.13%	32.99%	36.06%	19.82%
Gabon	13.90%	30.14%	35.31%	20.65%
Gambia	14.57%	21.28%	29.23%	34.93%
Ghana	20.45%	27.83%	30.79%	20.93%
Guinea	16.50%	21.47%	34.56%	27.48%
Guinea-Bissau	20.12%	24.24%	26.57%	29.07%
Mauritania	13.54%	17.33%	35.71%	33.42%
Niger	12.46%	17.45%	28.90%	41.20%
Nigeria	14.58%	21.57%	34.08%	29.78%
Senegal	15.11%	22.59%	41.13%	21.17%
Sierra Leone	16.44%	21.38%	38.77%	23.41%
Togo	18.31%	23.75%	30.62%	27.32%

Distribution of infant malaria deaths across time (for each country): children < 5 years

Guinea's malaria mortality is "average" for 1990 and 2000

Nigeria's mortality is "average" for all four years

(makes relatively little contribution to the association)

Est. Number of Deaths: Children < 5yrs

Relative Proportions

Country	1980	1990	2000	2010
Angola	14.64%	21.56%	40.65%	23.16%
Benin	16.63%	21.69%	31.54%	30.14%
Burkina Faso	12.02%	17.69%	37.51%	32.78%
Central African Republic	12.83%	20.38%	37.96%	28.84%
Chad	14.33%	18.12%	33.40%	34.15%
Comoros	16.07%	21.14%	37.95%	24.84%
Congo	15.28%	22.36%	38.43%	23.93%
Democratic Republic of Congo	11.51%	23.06%	39.85%	25.57%
Equatorial Guinea	11.13%	32.99%	36.06%	19.82%
Gabon	13.90%	30.14%	35.31%	20.65%
Gambia	14.57%	21.28%	29.23%	34.93%
Ghana	20.45%	27.83%	30.79%	20.93%
Guinea	16.50%	21.47%	34.56%	27.48%
Guinea-Bissau	20.12%	24.24%	26.57%	29.07%
Mauritania	13.54%	17.33%	35.71%	33.42%
Niger	12.46%	17.45%	28.90%	41.20%
Nigeria	14.58%	21.57%	34.08%	29.78%
Senegal	15.11%	22.59%	41.13%	21.17%
Sierra Leone	16.44%	21.38%	38.77%	23.41%
Togo	18.31%	23.75%	30.62%	27.32%

2000 only

Relatively low malaria mortality

(possible weak association)

Relatively high malaria mortality

(possible strong association)

Distribution of infant malaria deaths across time (for each country): children < 5 years

Est. Number of Deaths: Children < 5yrs

Relative Proportions

Country	1980	1990	2000	2010
Angola	14.64%	21.56%	40.65%	23.16%
Benin	16.63%	21.69%	31.54%	30.14%
Burkina Faso	12.02%	17.69%	37.51%	32.78%
Central African Republic	12.83%	20.38%	37.96%	28.84%
Chad	14.33%	18.12%	33.40%	34.15%
Comoros	16.07%	21.14%	37.95%	24.84%
Congo	15.28%	22.36%	38.43%	23.93%
Democratic Republic of Congo	11.51%	23.06%	39.85%	25.57%
Equatorial Guinea	11.13%	32.99%	36.06%	19.82%
Gabon	13.90%	30.14%	35.31%	20.65%
Gambia	14.57%	21.28%	29.23%	34.93%
Ghana	20.45%	27.83%	30.79%	20.93%
Guinea	16.50%	21.47%	34.56%	27.48%
Guinea-Bissau	20.12%	24.24%	26.57%	29.07%
Mauritania	13.54%	17.33%	35.71%	33.42%
Niger	12.46%	17.45%	28.90%	41.20%
Nigeria	14.58%	21.57%	34.08%	29.78%
Senegal	15.11%	22.59%	41.13%	21.17%
Sierra Leone	16.44%	21.38%	38.77%	23.41%
Togo	18.31%	23.75%	30.62%	27.32%

2010 only

Relatively low malaria mortality

(possible weak association)

Relatively high malaria mortality

(possible strong association)

Distribution of infant malaria deaths across time (for each country): children < 5 years

Est. Number of Deaths: Children < 5yrs

Summary

So, let's summarise

- Angola and Senegal have very *similar* distributions across the four years
- Congo and Sierra Leone have very *similar* distributions across the four years
- Ghana and Niger have very *different* distributions across the four years

Focusing on the distribution for each country over the four year period . . .

- . . . there is a relatively *high* malaria infant mortality in
  - Ghana, Guinea-Bissau and Togo in 1980
  - Equatorial-Guinea, Gabon and Ghana in 1990
  - Senegal, Angola, Burkina-Faso and the Central African Republic in 2000
  - Chad, Gambia and Niger in 2010

- . . . there is a relatively *low* malaria infant mortality in
  - Burkina Faso, Democratic Republic of Congo and Equatorial Guinea in 1980
  - Burkina Faso, Mauritania and Niger in 1990
  - Gambia, Guinea-Bissau and Niger in 2000
  - Equatorial-Guinea, Ghana and Senegal in 2010

Nigeria has neither a relatively high, or relatively low, mortality compared with other countries (and so appears to contribute very little to the association)

Est. Number of Deaths: Children < 5yrs

To more formally study infant deaths due to malaria . . .

```
> chisq.test(malaria2.dat)

        Pearson's Chi-squared test

data:  malaria2.dat
X-squared = 18278.55, df = 57, p-value < 2.2e-16

>
```

- With a p-value < 0.0001, there is a very strong association between country and year
- This, really, isn't surprising . . .
- . . . the chi-squared statistic is linearly related to the sample size (if n doubles, so too does the chi-squared statistic)
- So the sample size of 1.6 million can "mask" the underlying association between categorical variables

Est. Number of Deaths: Children < 5yrs

Unfortunately, the chi-squared test of independence

- Does not provide any indication of **row** categories that provide a similar, or different, impact on the association structure
- Does not provide any indication of **column** categories that provide a similar, or different, impact on the association structure
- At the mercy of the sample size (as we just described)

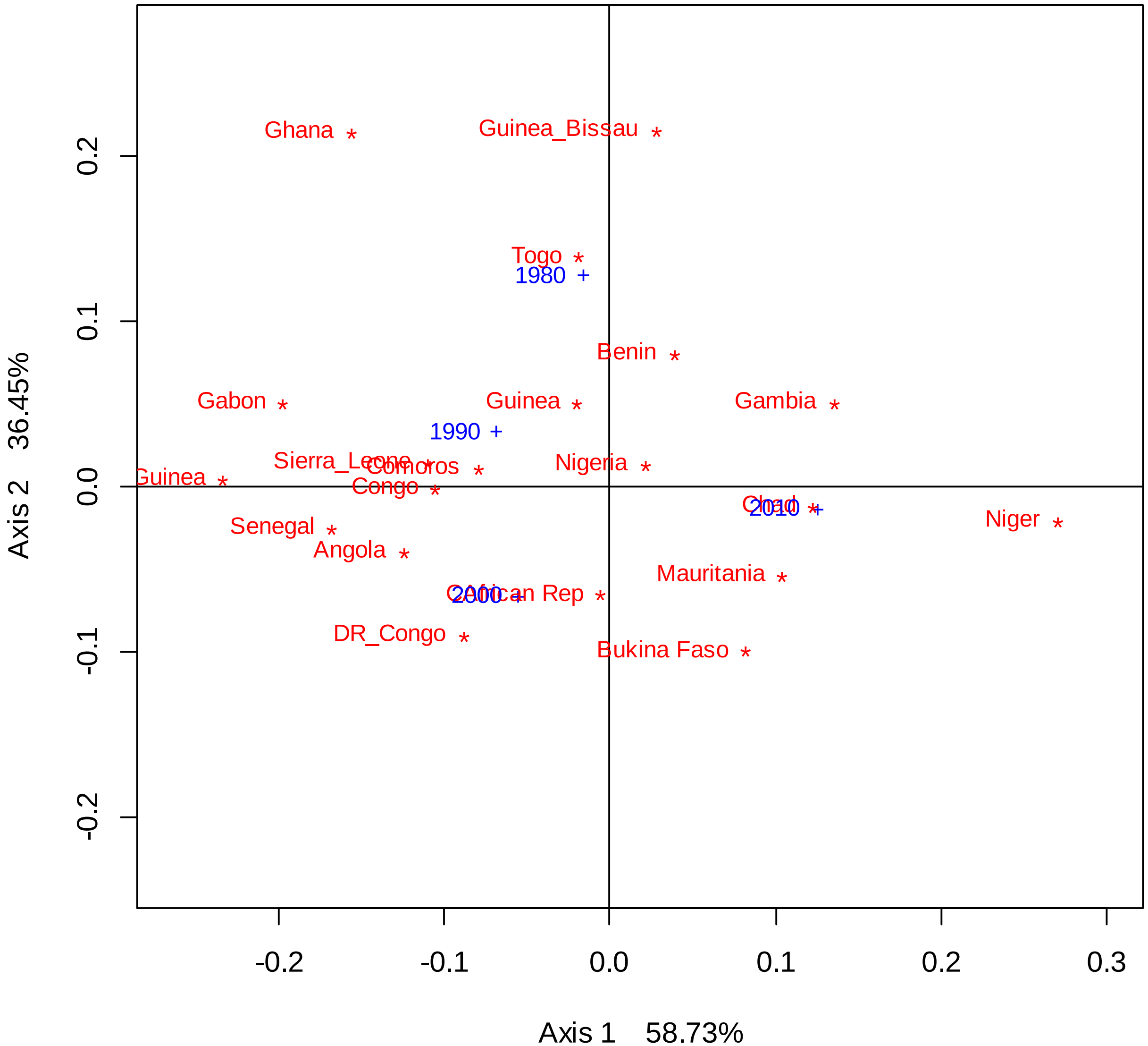
Correspondence analysis (CA) can be used to investigate further the association structure.

Rather than use the chi-squared statistic, X^2, CA uses

$$X^2/n$$

called the *total inertia*, to quantify the magnitude of the association. For our example Total inertia = 18278.5/1607161 = 0.0114

Est. Number of Deaths: Children < 5yrs



How many dimensions can we work with?

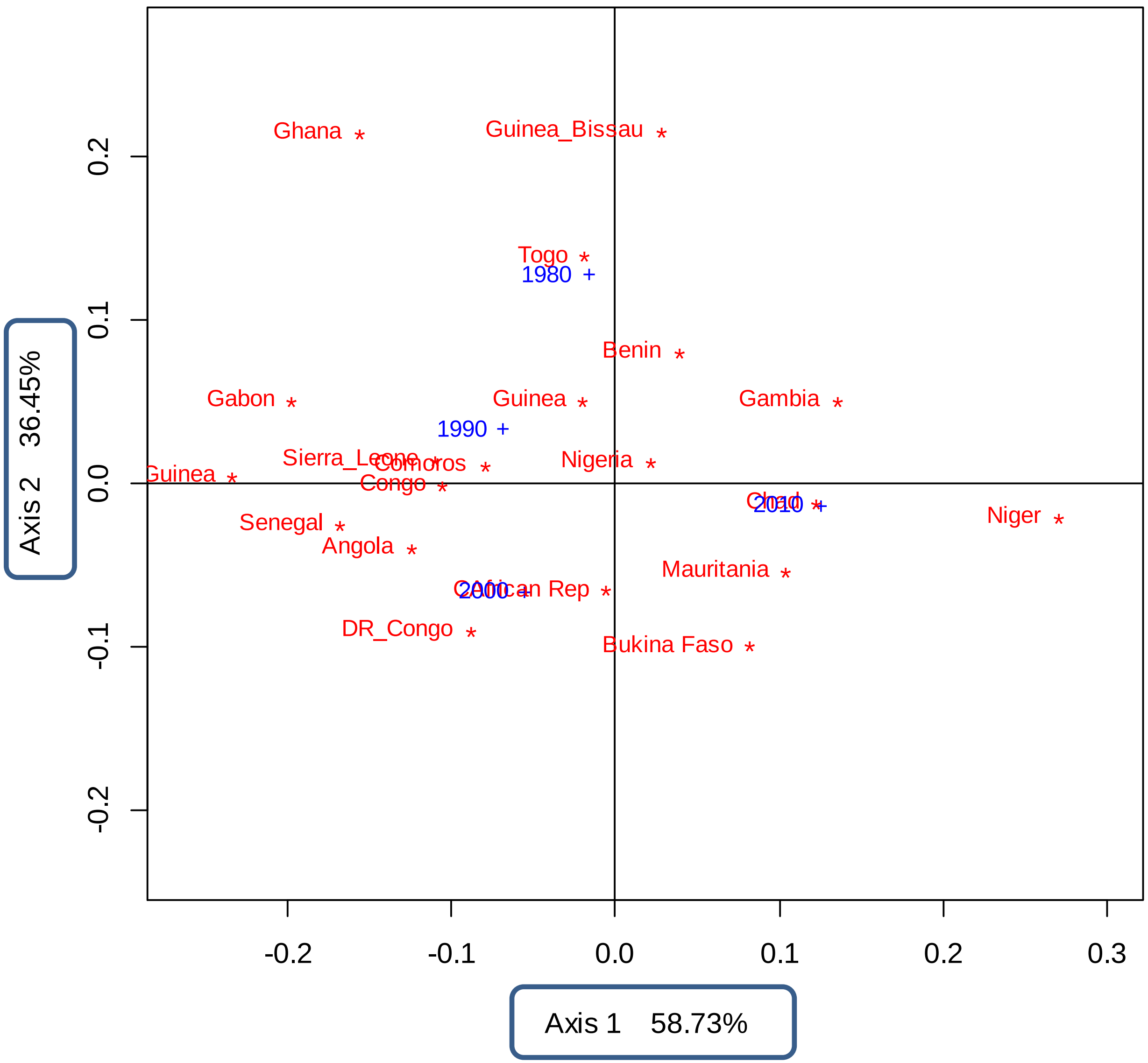
min(# rows, # cols) – 1
= min(20, 4) = 3

A three dimensional plot will graphical depict **ALL** of the association between *Country* and *Year*

Total Inertia = 0.0114

Analysis carried out using CAvariants package on the CRAN (Lombardo & Beh, 2015)

Est. Number of Deaths: Children < 5yrs



Quality of 2D Plot?

The first axis depicts 58.73% of the total inertia

The second axis depicts 36.45% of the association

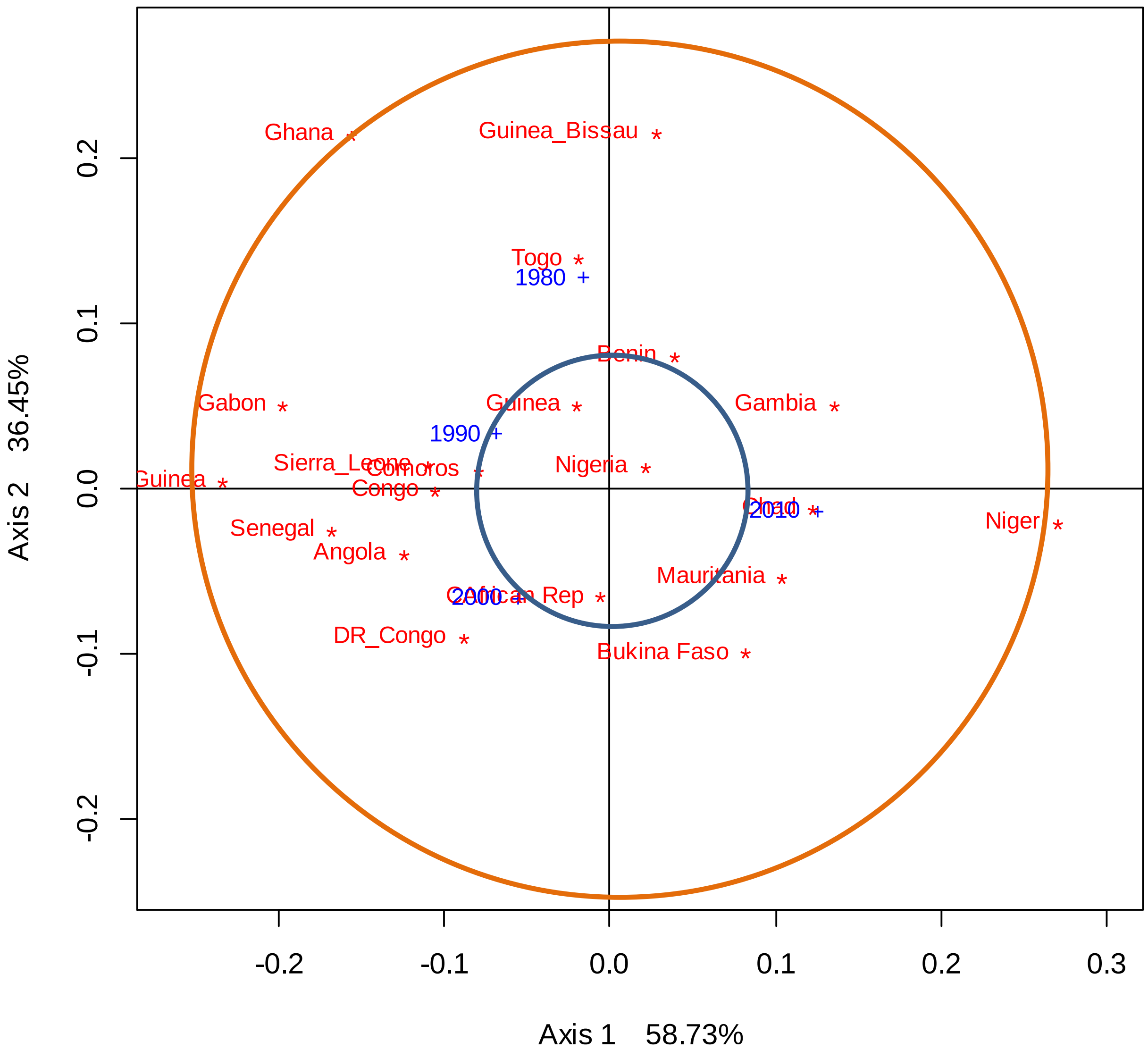
This two dimensional display graphical reflects

58.73% +
36.45% = 95.18%

of the association between country and year

Analysis carried out using CAvariants package on the CRAN (Lombardo & Beh, 2015)

Est. Number of Deaths: Children < 5yrs



Origin?

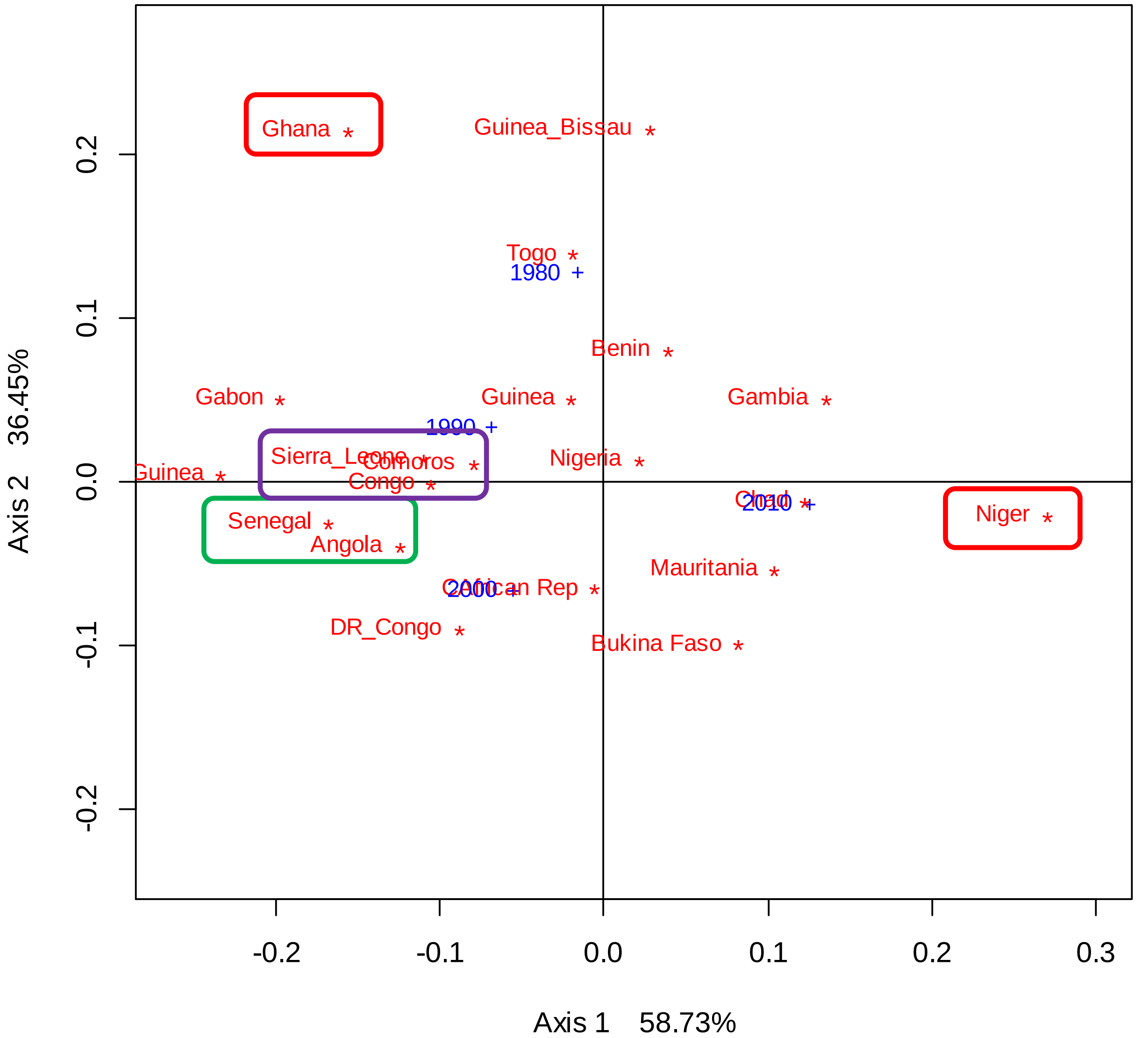
Generally, points close to the origin have no role, or a relatively small role, in defining the association

- Nigeria
- Guinea

Points far from the origin play are dominate categories in defining the association

- Niger, Ghana, Guinea, Gabon

Est. Number of Deaths: Children < 5yrs

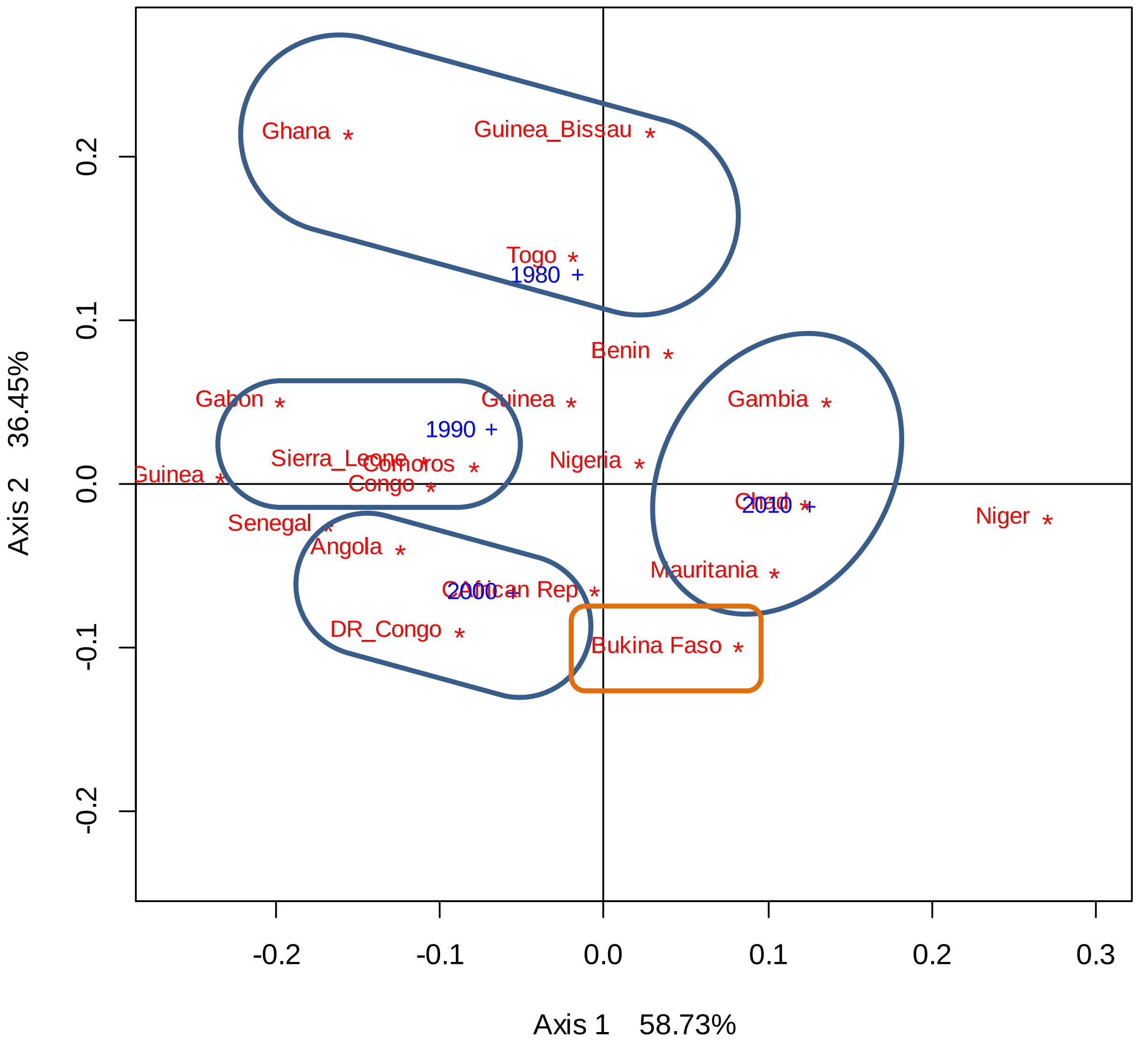


Distances?

- Points close to each other indicate a similar distribution across the years
  - Senegal/Angola
  - Congo/Sierra Leone
- Points far from each other indicate very different distributions across the years
  - Niger/Ghana

Our comparisons on the previous slides suggested this

Est. Number of Deaths: Children < 5yrs



Distances?

Relatively high mortality in

- Togo, Ghana, Guinea-Bissau in 1980
- Sierra Leone, Congo, Gabon in 1990
- C Afr Rep, DR Congo, Angola in 2000
- Gambia, Chad, Niger, Mauritania in 2010

Bukina Faso has relatively high mortality in 2010 and 2000

The Basic Idea of Correspondence Analysis

Dimensions

How Many Dimensions Should We Use to Visualise the Association?

There are various schools of thought on this . . .

- Blasius (1994) suggested choosing those dimensions whose percentage contribution exceeds the average

$$\text{average} = \frac{100}{\min(I, J) - 1} = \frac{100}{\min(20, 4) - 1} = 33.33$$

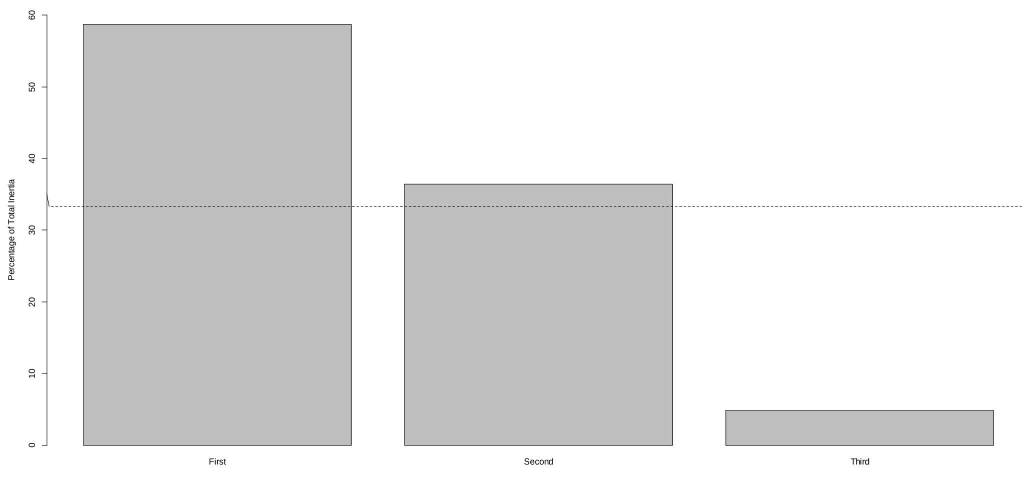
- For our malaria data this suggests only two dimensions are needed.
- Sometimes overestimates the number of dimensions really needed

- There are inferential procedures based on Monte-Carlo p-values for each axis. Although this can be computationally intensive (and defeats the purpose of a simple visualisation of the association)

How Many Dimensions Should We Use to Visualise the Association?

There are various schools of thought on this . . .

- The scree-plot . . . Simply put, construct a scree-plot (basically a barchart) of the percentage contribution each axis makes to the association (measured using the total inertia).
- Where there is a natural “cliff” in the bars, that defines how many dimensions you should consider



This scree-plot suggests only two dimensions are needed

The Basic Idea of Correspondence Analysis

How Many Dimensions Should We Use to Visualise the Association?

- Jolliffe (1986), who considered the same issue but from a principal component analysis issues says

  "the rules of which have more sound statistical foundations seem, at present, to offer little advantage over the simpler"

- Benzecri (1992, pf 398) believes the decision should be made based on the researchers personal judgement rather than by any mathematical procedure. - However, the analyst should at least be aware that potentially important information is lost if higher dimensions are not considered.

- In many practical problems, the issue of "how many dimensions", and the quality of a display rarely is considered (this is a problem).

Est. Number of Deaths: Individuals ≥ 5yrs

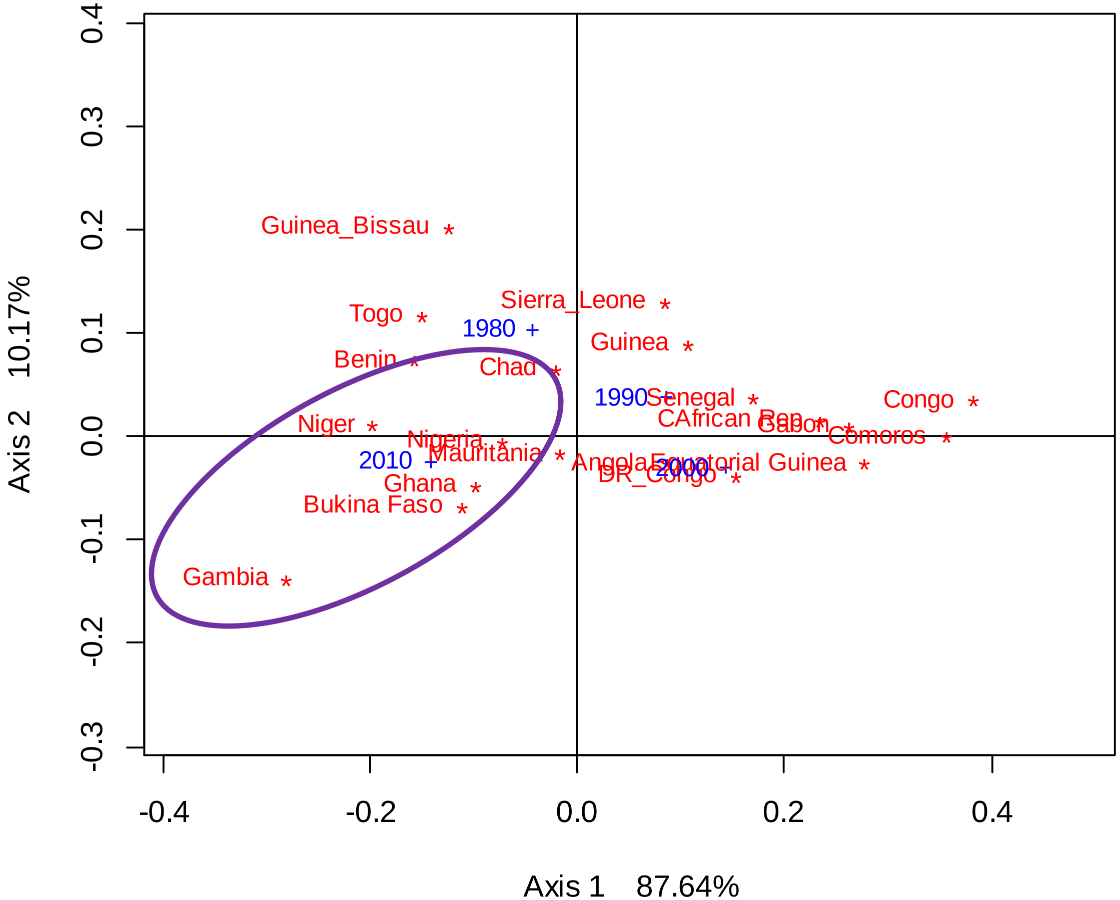
Country	1980	1990	2000	2010	% Diff ('00 -'10)
Angola	1937	2706	6947	6737	-3.0%
Benin	1724	2122	3414	6164	**80.6%**
Burkina Faso	3109	4886	10331	16074	**55.6%**
Central African Republic	1226	2370	5076	3641	-28.3%
Chad	1442	1820	3548	4516	27.3%
Comoros	87	154	402	205	**-49.0%**
Congo	912	2215	4481	2243	**-49.9%**
Democratic Republic of Congo	10146	18992	44004	38045	-13.5%
Equatorial Guinea	125	301	650	431	-33.7%
Gabon	292	862	1494	1080	-27.7%
Gambia	90	140	296	662	**123.6%**
Ghana	2400	4046	7805	12049	**54.4%**
Guinea	2005	2831	5568	5299	-4.8%
Guinea-Bissau	358	408	536	941	**75.6%**
Mauritania	189	266	593	737	24.3%
Niger	1727	2181	3830	7428	**93.9%**
Nigeria	27865	39966	79395	114213	43.9%
Senegal	1971	3670	7186	6066	-15.6%
Sierra Leone	1566	2088	3810	3827	0.4%
Togo	1135	1453	2035	3767	**85.1%**

Table 2: Estimated number of deaths due to malaria (P.falciparum): individuals ≥ 5 years

Dimensions

Est. Number of Deaths: Individuals ≥ 5yrs

Dimensions

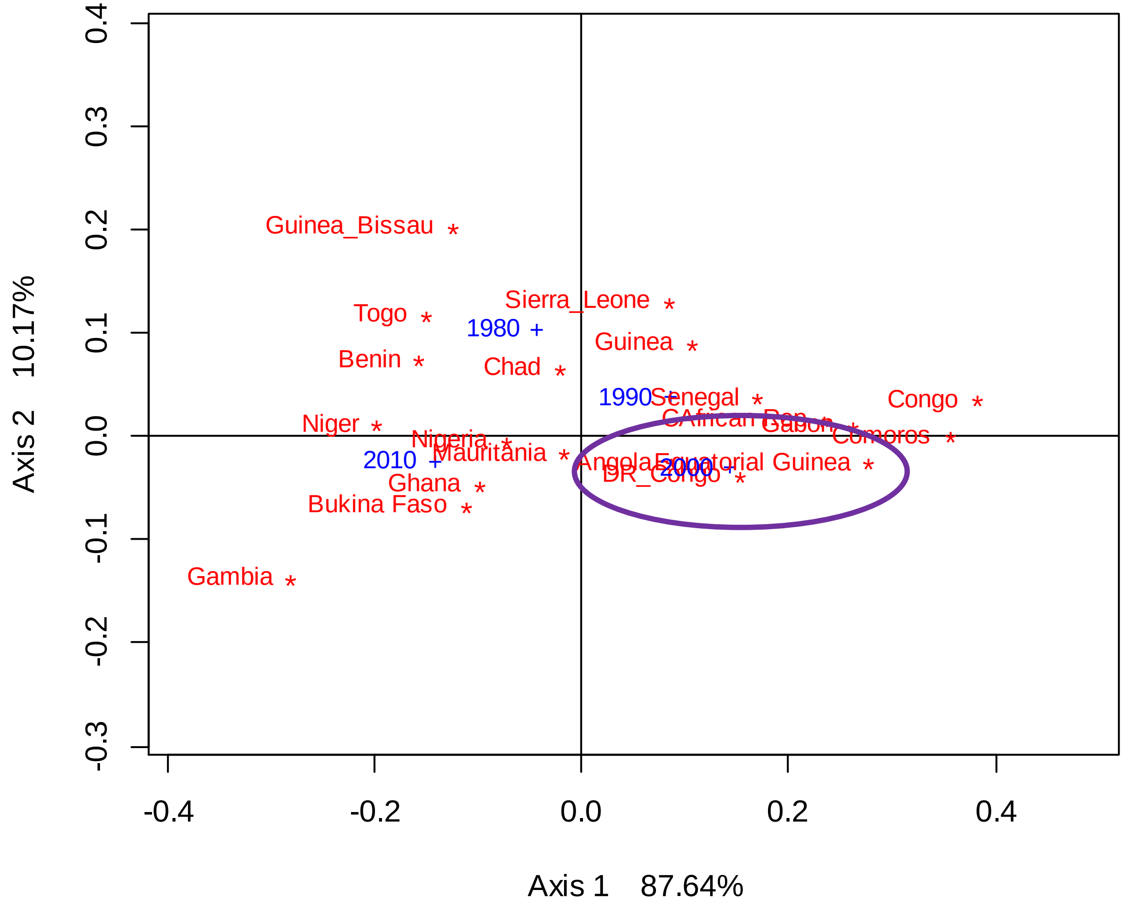


In 2010, relatively high mortality was seen (when compared to other periods) in

- Gambia
- Burkina Faso
- Ghana
- Niger
- Nigeria

Est. Number of Deaths: Individuals ≥ 5yrs

Dimensions



In 2000, relatively high mortality was seen (when compared to other periods) in

- DR Congo
- Equatorial Guinea

Multiple Correspondence Analysis

The data of malaria consisted of **two** variables.

Row Variable = *Country*
Column Variable = *Year*

However, the study also involves a third (dichotomous) variable – *Age*

For more than three categorical variables, we need to consider

- an alternative way of summarising the data

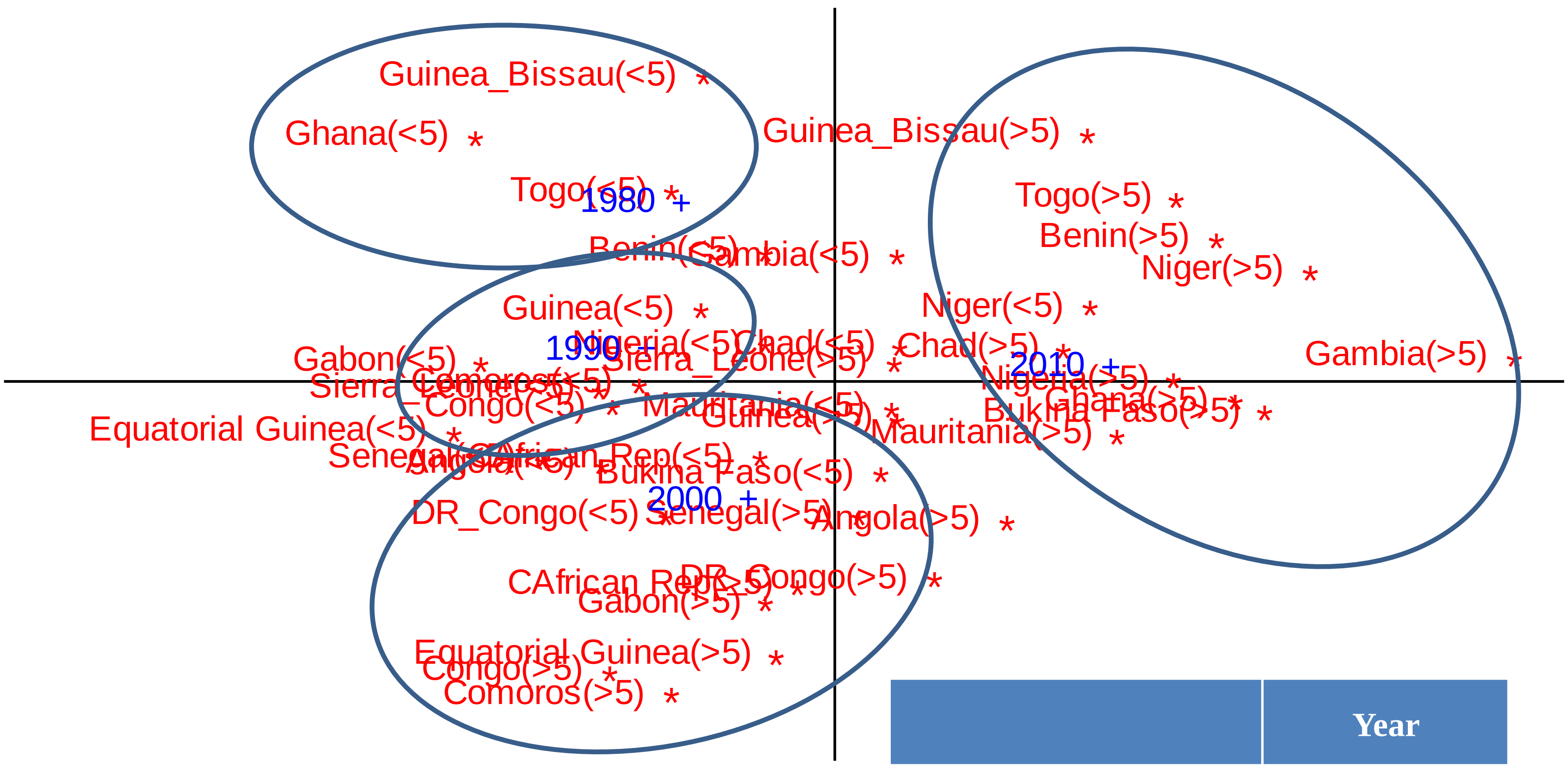
so that we can obtain a graphical depiction of the association.

We shall describe the two most popular approaches which involve recoding a multi-way contingency table into a two-way form by considering

- *stacking* one of the variables,
- the *indicator matrix* form of the data
- the *Burt matrix* form of the data

Multiple Correspondence Analysis

Stacking (by Year)

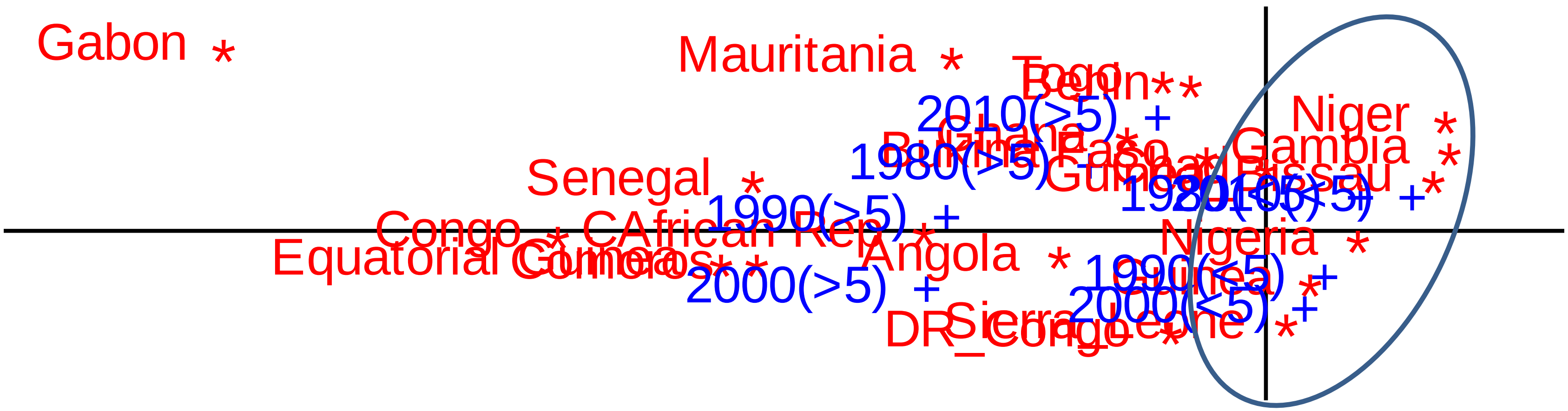


This plot reflects 97.97% of the total inertia (association) of

	Year
Countries (< 5 yrs)	Table 1
Countries (≥ 5 year)	Table 2

Multiple Correspondence Analysis

Stacking (by Country)



Countries with relatively higher deaths amongst infants than those aged at least 5 years:

- Niger
- Gambia
- Nigeria
- Guinea-Bissau
- Sierra-Leone
- DR Congo

	Year (<5)	Year (≥5)
Countries	Table 1	Table 2

Expected number of deaths in Gambon . . .

	1980	1990	2000	2010
< 5 years	183	397	465	272
≥ 5 years	292	862	1494	1080

. . . relatively (compared with other countries) much higher in older group

Indicator Matrix

Any sized contingency table can be depicted as an indicator matrix, denoted by Z.

Each row of the indicator matrix represents how each individual in the sample (n) is classified into the categories.

Z consists of only the elements 1 and 0; 1 where the individuals exhibits a particular characteristic, 0 where it doesn't.

For our malaria data – *Country* x *Year* x *Age* - Z is formed by concatenating three sub-matrices (one for each variable) such that

$$Z = [\, Z_I \;\; Z_J \;\; Z_K \,]$$

Z: n x 26; Z_I: n x 20; Z_J: n x 4; Z_K: n x 2

Maximum number dimensions in the subspace

$$\min(n, 26) - 1$$

Here, for the two sets of data, n = 2,186,470

Memory issues in R ☹

The Burt Matrix

For the coding of a two-way contingency table, the Burt matrix consists of the concatenation of the original contingency table with diagonal matrices consisting of the row and column marginal frequencies.

If

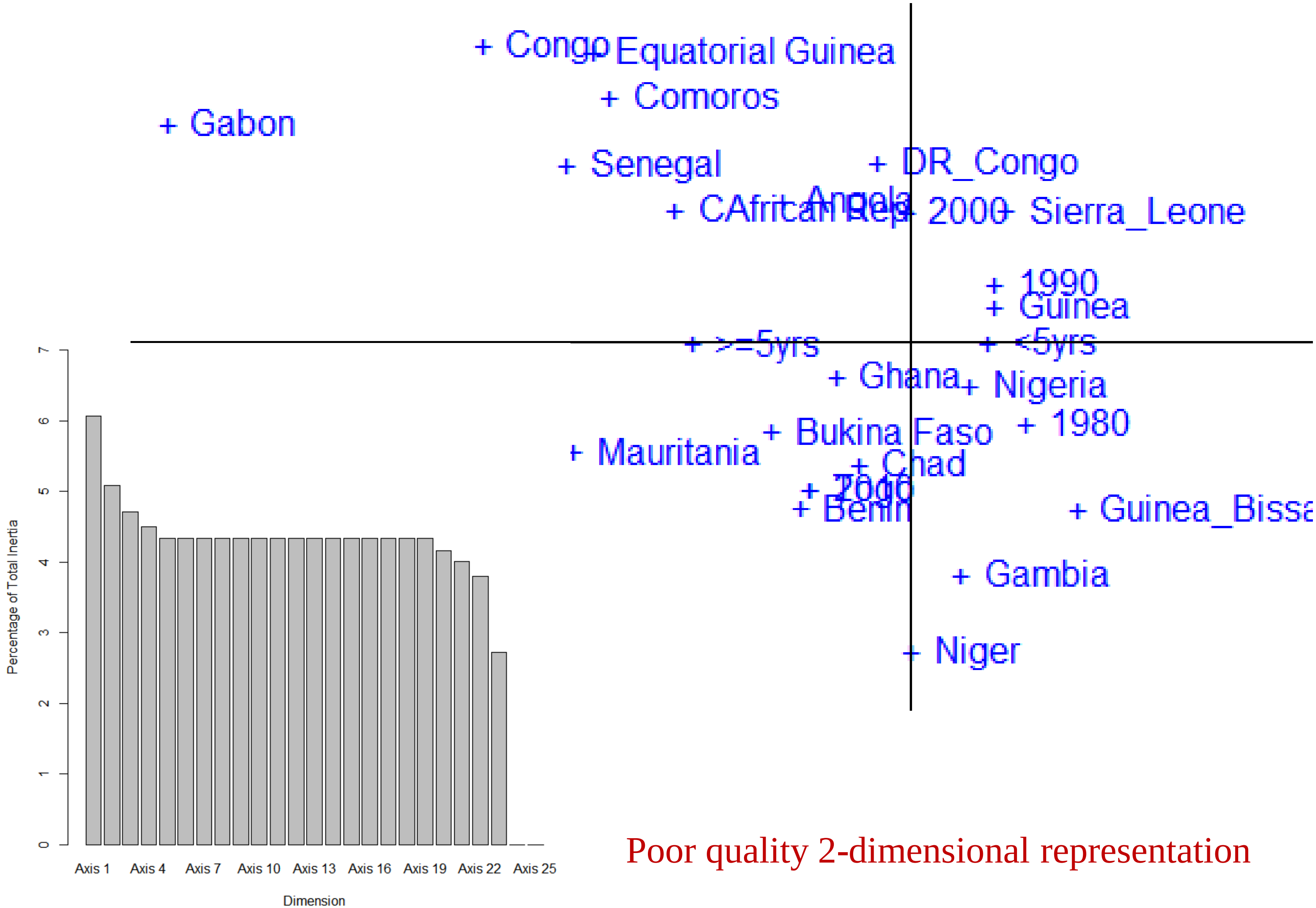
- D_I denotes the diagonal matrix of relative frequencies for *Country*,
- D_J is the diagonal matrix of relative frequencies for *Year*, and
- D_K is the diagonal matrix of relative frequencies for *Age*,

$$B = \begin{pmatrix} D_I & N_{IJ} & N_{IK} \\ N_{IJ}^T & D_J & N_{JK} \\ N_{IK}^T & N_{JK}^T & D_K \end{pmatrix}_{26 \times 26}$$

where, for example, N_{IJ} is the two way table for *Country* x *Year* (aggregating across *Age*)

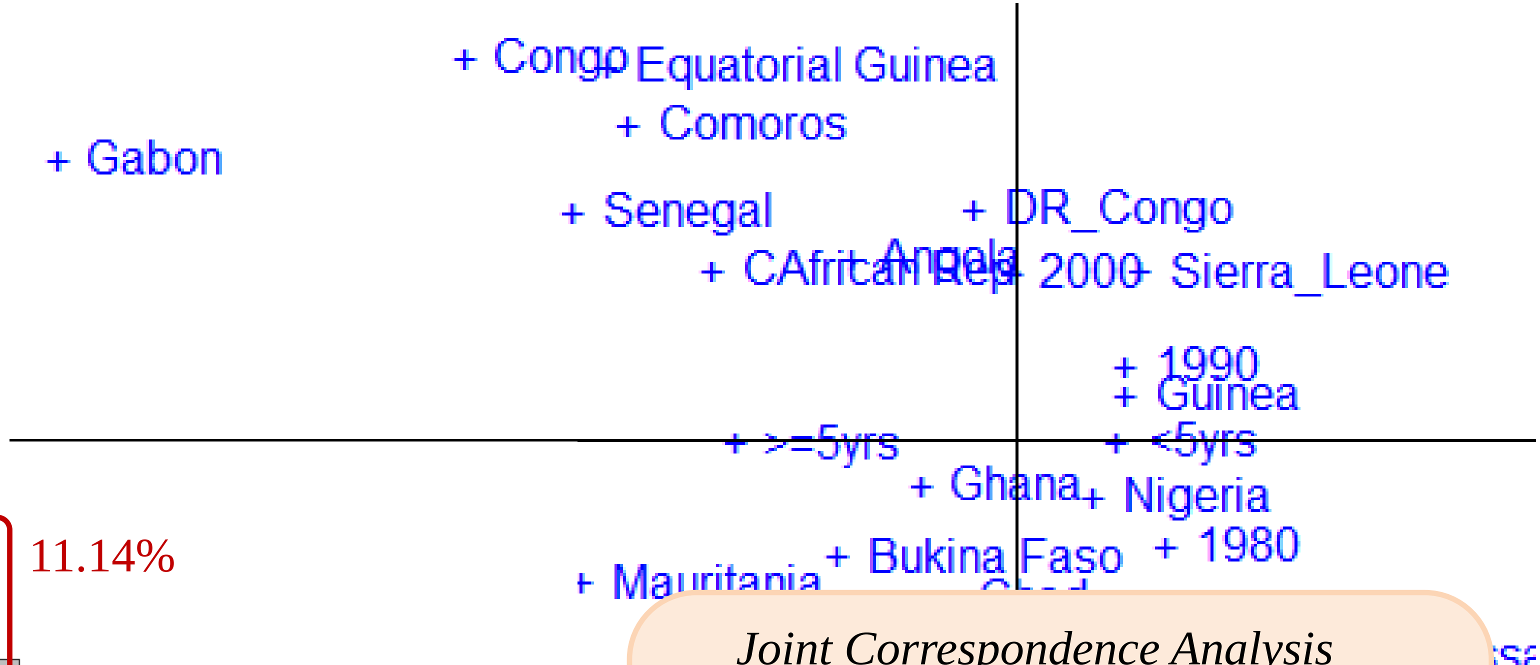
$$N_{IJ} = Z_I^T Z_J$$

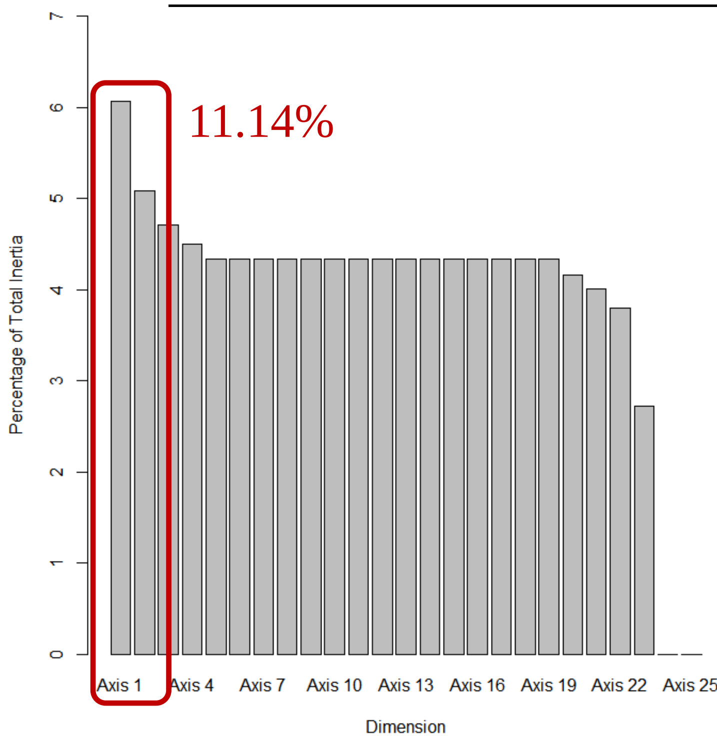
The Burt Matrix



Poor quality 2-dimensional representation

The Burt Matrix





Joint Correspondence Analysis

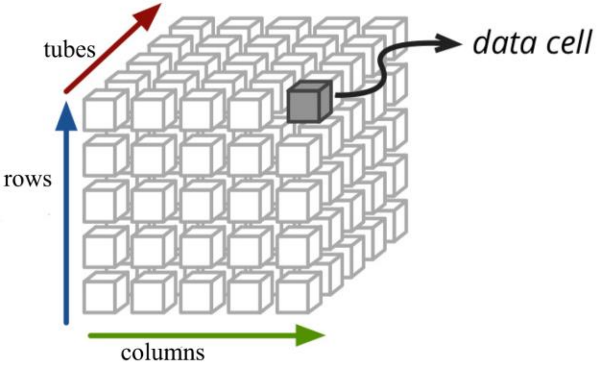
$$\mathbf{B} = \begin{pmatrix} \mathbf{D}_I & \mathbf{N}_{IJ} & \mathbf{N}_{IK} \\ \mathbf{N}_{IJ}^T & \mathbf{D}_J & \mathbf{N}_{JK} \\ \mathbf{N}_{IK}^T & \mathbf{N}_{JK}^T & \mathbf{D}_K \end{pmatrix}_{26 \times 26}$$

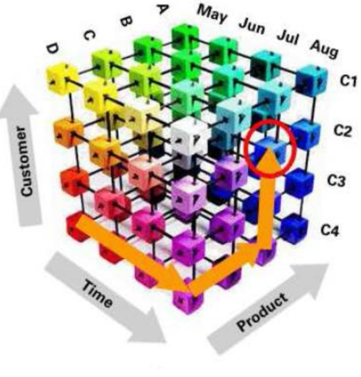
Poor quality 2-dimensional representation

Variations of Correspondence Analysis

Pieter Kroonenberg (Leiden University, The Netherlands), amongst others, criticised the use of recoding multi-way data into a two-way form since

- no information about each of the pair-wise interactions can be obtained
- no information about the multiple interactions between the variables can be obtained
- the measure of total inertia invariably exclusively involves Pearson's chi-squared statistic





To reflect varying association structures other measures of association can be considered.

Other issues with multiple correspondence analysis identified by Greenacre (1990)

Variations of Correspondence Analysis

By considering Pearson's chi-squared statistic we are treating the variables to be *symmetrically* associated. That is, both variables are treated as a predictor variable

When we have

- a predictor categorical variable
- a response categorical variable

the association is described as being *asymmetric*. Therefore the most appropriate measure of association is the Goodman-Kruskal tau index (Goodman and Kruskal, 1954, p. 759)

When considering this index, we *have non-symmetrical correspondence analysis*

(D'Ambra and Lauro, 1989, 1992; Kroonenberg and Lombardo, 1999; Lombardo, Beh and D'Ambra, 2007)

Variations of Correspondence Analysis

When variables consist of ordered categories, this ordered structure provides important additional information about the association between the variables.

There have been a variety of techniques proposed to reflect the case where a categorical variable is ordinal. See, for example

Parsa and Smith (1993), Ritov and Gilula (1983),
Schriever (1983), Yang and Huh (1999)

However, Nishisato (2007, p. 237) says of these such as techniques

“. . . We wonder if it useful or even meaningful to impose an order constrain on categories . . . More unfortunately than fortunately, it is a generally accepted view that if the categories are ordered then the weights given to them must be ordered. Why is the view so popular? . . . Frankly speaking, it is a silly and harmful belief”

That’s because they “force” coordinates to be ordered in a correspondence plot

Alternatively, use orthogonal polynomials → *generalised correlations*

Other Issues

- In this presentation we have looked at the traditional approach to correspondence analysis – the analysis of counts (from a contingency table)

Other Types of Data

- Square (symmetric) contingency tables – same variables but at two different time periods or locations
- Sparse data – large cell counts and small cell counts
- Ranked data – eg ranking 10 treatments from 1 to 10 (no ties)
- Ratings and preferences – eg "doubling" (Greenacre, 1984)
- Proximity (distance) data – between objects, cities, etc

Theoretical Links

- Log-linear models
- Time series analysis
- Only tentative links with Bayesian analysis
- Cluster identification – eg, dendrograms, `mclust` algorithm in R

Other Issues

- Nearly all of the popular statistical packages allow the user to perform a correspondence analysis on their categorical variables:
  - JMP
  - Minitab
  - SAS
  - SPSS

- There are also freely downloadable programs
  - PAST (**PA**leontological **ST**atistics) from http://folk.uio.no/ohammer/past/
  - *DtmVic5.6+* (**D**ata and **t**ext **m**ining **V**isualization, **i**nference, **c**lassification) from www.dtmvic.com

Other Issues

In S-PLUS

There are also a host of Splus functions that have been made available in the past: Everitt (1994), Venables and Ripley (1999, pp. 342 – 344), Beh (2005)

In R

- The `CAvariants` package by Lombardo and Beh (2014)
- The `ca` package in the MASS library
- Nenadić & Greenacre's (2007) `ca` library
- de Leeuw and Mair's (2009) `anacor` library
- Murtagh's (2005, pg 18 – 20) `ca.r` function
- Baxter and Cool (2010) present R code with an archaeological flavour
- Chessel, Dufour and Thioulouse's (2004) `dudi.ca` function in the library ADE4 ("**D**ata **A**nalysis functions to analyse **E**cological and **E**nvironmental data in the framework of **E**uclidean **E**xploratory methods"

Many of these R libraries also allow the analyst to perform a variety of other techniques that belong to the correspondence analysis family.

Other Issues

- Non-symmetrical correspondence analysis
- Taxicab correspondence analysis
- Constrained correspondence analysis
- Linearly constrained correspondence analysis
- Cumulative correspondence analysis
- Detrended correspondence analysis
- Semi-supervised detrended correspondence analysis
- Canonical correspondence analysis
- Partial canonical correspondence analysis
- Discriminant correspondence analysis
- Multi-block discriminant correspondence analysis
- Internal correspondence analysis
- Intra-table correspondence analysis
- Weber correspondence analysis

- Canonical non-symmetrical correspondence analysis
- Constrained non-symmetrical correspondence analysis
- Taxicab non-symmetrical correspondence analysis
- Partial non-symmetrical correspondence analysis

- Generalised constrained multiple correspondence analysis
- Multiple taxicab correspondence analysis
- Partial multiple correspondence analysis

See Beh and Lombardo (2014) book for more details on the family – includes **more than 35 members**