Exploring the Changing Media Repertoires in South Africa

Submitted as a partial fulfillment toward and MSc in Data Science 2017

Department of Statistical Science

University of Cape Town

by

Hans-Peter Bakker

December 2017

Supervisor: Professor Ian Durbach

Abstract

Abstract to go here...

Contents

Chapter 1

Data Preparation

In this section the ....

1.1 Overview of the All Media and Products Survey

asdfa

1.2 AMPS 95

from code book:

"The objective of the All Media and Products Survey (AMPS) is to collect information on the readership of newspapers and magazines, television viewing, radio listening, cinema going, and on the acquisition, possession or use of a selection of products and services, together with extensive demographic data."

"SAMPLING: The universe from which the AMPS sample is drawn, comprises adults aged 16 years or older resident in private households, or hotels, residential hotels and similar accommodation in the Republic of South Africa. In the case of each racial group, certain areas were excluded from consideration, as containing no persons in a given group of negligible numbers of them. A multistage, stratified, quasi-probability design was employed. This study is based on a full annual sample. The data were collected by personal, in-home interviews."

"DATE OF RESEARCH: January - June 1995"

"Number of cases: 14643 = Blacks 6557 + Coloureds 1439 + Indians 765 + Whites 5882."

"The AMPS95 dataset has the following weight variables: 1) Population Weights – "popwght". 2) Household Weights (not to be used with personal demographics) – "hhwght". 3) Household Decision Maker Weights - to be used with personal demographics. These must be used for personal analyses on household items – only on a filtered of male heads of household and female housewives – "hdmwght". 4) Purchaser Weights – These must be used for the household products only on a filtered base of those wholly or partly responsible for household purchases – “purwght”.”

1.2.1 Data Preparation

Decision on what variable to use and how to construct them:

wanted to get combined dataset of demographic data that could be useful and media consumption data:

On demographic / psychographic data: not really any psychographic stuff in 95.

a) educatn:

qpd10: 1 2 3 4 5 6 7 8 9 10 11 12: 1028 1909 1106 5008 3278 285 697 610 332 235 120 35

educatn: 1 2 3 4 5 6 7 8: 1028 1909 1106 5008 3278 697 610 1007.

so:

1: no schooling

2: some primary school

3: primary school complete

4: some high school

5: matric

6: Technikon diploma/degree completed

7: University degree completed

8: Other (incl: artisan, secretarial, technical, professional)

to create ordered factor took 8 and fit it after matric (ie 6) then change 6 to 7 and 7 to eight

b) h_inc_g1

Based on factors and on the questionaire,

1 - 8 factors... household income...,

24,25,26,27,28, 29 = 8 (R9000 +)

21,22,23 = 7 (R6000 - R8999)

19,20 = 6 (R4000 - R5999)

17,18 = 5 (R2500-R3999)

14,15,16 = 4 (R1400 - R2499)

10,11,12,13 = 3 (R900 - R1399)

6,7,8,9 = 2 (R500 - R899)

1,2,3,4,5 = 1 (R1 - 499)

personal income:

Same categories as h_inc...

two anomolies|:

(no personal income category 30) ... will categorise as 1

(refused31)...will add NA

c) sex:

1: male

2: female

d) age: (NB...as it stands only 4 levels. last two joined. for 5 levels as below use qpd8

1: 16-24

2: 25-34

3: 35-49

4: 50-64

5: 65+

e) lang

looks like, main language spoken at home: not sure what is meant by African1 and African1

1: English

2: Afrikaans

3: African1

4: African2

e) race_1:

not sure what race_2 refers to...only “1”: 8086 cases

WBCI: 1,2,3,4:

f) province

1-9: provinces. no idea which no is which province. Can make assumptions based on most to least populated...? Or check later, better code books?

from 2002:

1 Western Cape

2 Northern Cape

3 Free State

4 Eastern Cape

5 Kwazulu-Natal

6 Mpumalanga

7 Limpopo

8 Gauteng

9 North-West

g) wrk_stat

1: Full-time

2: part-time

3: other

h) marital status

qpd1: 1: Single, 2: Maried or living together; 3: Widowed, 4: Divorced, 5: Seperated.

Appear to be no psychographic information in AMPS 1995

i) occupation - industry (qpd5) CURRENTLY NOT IN THE DATASET...!!

Agriculture 1

Commerce and Finance 2

Construction 3

Electricity, gas, water and sanitary 4

Manufacturing 5

Mining and Quarrying 6

Services-Government/Personal/Business 7

Transport, Storage and Communication 8

Other and Unemployed 9

metropole

FROM 2002 VALUE LABELS...ASSUMED...NEED TO CHECK FOR 95...

1 Cape Town

2 Cape Town Fringe Area

3. Port Elizabeth/Uitenhage

4 East London

5 Durban

6 Bloemfontein

7 Greater Johannesburg (Alex.,JHB,Sandton,Soweto,Randburg)

8 Reef (Urban Gauteng excl JHB,Pta,Vaal)

9 Pretoria (includes North-West Metro)

10 Kimberley

11 Pietermaritzburg

12 Soweto NB SOWETO HERE AND AT LEVEL 7

13 Vaal (De Deur,Meyerton,Sasolburg,Vanderbijlpark,Vereeniging)

Media consumption data: aim to identify degrees of engagement:

Aim was to identify degrees of engagement by both media type (print, tv, radio) and particular vehicles (beeld, 5FM, eTv etc.)

no questions pertaining to digital media usage in AMPS 1995

print

dailies, weekly newspapers, weekly magazines, fortnightly magazines, monthly magazines

B2: In an average week from Monday to Friday, 5 issues of each daily newspaper are published. How many different issues of... (mention daily newspaper), do you usually read or page through?

C2. In a 6 week period, 6 issues of each weekly newspaper are published. How many different issues of ...(MENTION WEEKLY NEWSPAPER) do you usually (read or page through)?

D2. In a 6 week period, 6 issues of each weekly magazines are published. How many different issues of...(Mention weekly magazines), do you usually read or page through?

E2. In a 12 week period, 6 ISSUES of each fortnightly magazines are published. How many different issues of..(MENTION FORTNIGHTLY MAGAZINE), do you read or page through?

F2. In a 6 months period, 6 ISSUES of each monthly magazine are published. How many different issues of....(MENTION MONTHLY MAGAZINE) do you usually read or page through?

Multiplied by thoroughness measure....

B6-F6. People read some publications very thoroughly but just page through others. Please tell me how thoroughly you usually read? Just give me(t) the number.

(NB... from later AMPS.. so no evidence that this is the case for 1995.... will make the assumption. (to reflect increased engagement, need to reverse these in the dataset)

1 Read right through it

2 Read most articles and page through the rest

3 Read some articles and page through the rest

4 Read only selected parts and ignore the rest

5 Page right through it

6 Just glance at parts of it

Variables: eg "Argus" ie by vehicle: "Engagement" Value: : measure of degree of engagement with that vehicle

Additional derived variables:

total_engagement_dailies: simply sum it up, so if a particular respondent "usually" reads all 5 issue of Beeld and all 3 issue of The Star: and the thoroughness measures are 2 and 3 respectively, the value for that item in total_engagement_dailies = (5 x 2) + (3 x 3) = 19.

total_engagement_print: simply the sum of all engagement values for that respondent.

radio

By radio listening we mean that you personally have listened to the radio it may be all of a programme or only part of it. It doesn't matter if it was your own radio or somebody where you listened to it.

G1. Here is a list of radio stations. Which of them, if any, have you personally listed to in the PAST 7 DAYS?

G2. And now I would like you to think about YESTERDAY. Which of these stations, if any did you personally listen to YESTERDAY/home or away from home, from the time you woke up in the morning till you went to sleep?

For a given radio station, coded 0 for not listened. 1 for yesterday or in the past 7 days and 2 for both yesterday and past 7 days. Rationale that if someone listened to a given station only yesterday or past seven days, it does not indicate regular listening of that station, while if someone listened to the station in the past seven days and yesterday, they are more likely to be considered regular listeners... so higher level of engagement.

FOR TIME... would have liked this info, but missing in the available dataset.

G4.You said you listened to .... (MENTION STATION), YESTERDAY/ON SATURDAY. Think about the WHOLE DAY, from the time you woke up in the morning, when you went to work, at work, when you came home until you went to sleep.

television

By watching TV we mean that you personally have watched all or part of a programme. It doesn't matter where you watched it.

H1. Here is a list of TV services. Which of these services, if any, have you personally watched in the PAST 7 DAYS?

H2. Now I would like you to think about YESTERDAY. Which, if any, of these TV Services did you personally watch YESTERDAY (used Saturday if yesterday was Sunday)

all media

dataset : "media95".Single dataset with media engagement data..

1.3 AMPS 2002

1.3.1 For demographics

1.3.1.1 age (ca44co38): NOT same as 1995

Original coding

1 15 - 19

2 20 - 24

3 25 - 34

4 35 - 44

5 45 - 49

6 50 - 54

7 55 - 64

8 65+

1.3.1.2 sex (ca46co51a)

1 = male

2 = female

1.3.1.3 education (ca46co48)

Original coding, except to create ordered factor I took 8 and fit it in after matric (ie 6) then changed 6 to 7 and 7 to eight (so 8 = University degree completed)

1: no schooling

2: some primary school

3: primary school complete

4: some high school

5: matric

6: Technikon diploma/degree completed

7: University degree completed

8: Other Post Matric

1.3.1.4 household income (ca46co50)

1 Up to R499

2 R500-899

3 R900-1 399

4 R1 400- 2 499

5 R2 500-3 999

6 R4 000-6 999

7 R7 000-11 999

8 R12 000+

1.3.1.5 race (ca46co51b)

1 White

2 Black

3 Coloured

4 Indian

1.3.1.6 province (ca46co56)

1 Western Cape

2 Northern Cape

3 Free State

4 Eastern Cape

5 Kwazulu-Natal

6 Mpumalanga

7 Limpopo

8 Gauteng

9 North-West

1.3.1.7 metro (ca46co57):

1 Cape Town

2 Cape Town Fringe Area

3 Port Elizabeth/Uitenhage

4 East London

5 Durban

6 Bloemfontein

7 Greater Johannesburg (Alex.,JHB,Sandton,Soweto,Randburg)

8 Reef (Urban Gauteng excl JHB,Pta,Vaal)

9 Pretoria (includes North-West Metro)

—

metro2 (ca46co58) becomes after adding 9:

10 Kimberley,

11 Pietermaritzburg

12 Soweto

13 Vaal (De Deur,Meyerton,Sasolburg,Vanderbijlpark,Vereeniging)

as in '95 needed to sort out double count of Soweto....

1.3.1.8 language (ca46co75): (see details: only showing 0-8)??????

0 Afrikaans

1 English

2 Zulu

3 Xhosa

4 N. Sotho

5 S. Sotho

6 Tswana

7 Tsonga

8 Venda

9 Swazi

10 Ndebele

11 Other

1.3.1.9 life stages (ca46co77):

1 At-home singles

2 Starting out singles

3 Mature singles

4 Young Couples

5 New parents

6 Mature parents

7 Single parents

8 Golden nests

9 Left alones

1.3.1.10 marital status (ca44co09):

1 Single

2 Married or living together

3 Widowed

4 Divorced

5 Separated

1.3.1.11 personal income (ca45co39; ca45co40; ca45co41; ca45co42)

combined these to give the following levels:

from (ca45co39) use codes as they are given

1 R1-R199

2 R200-R299

3 R300-R399

4 R400-R499

5 R500-R599

6 R600-R699

7 R700-R799

8 R800-R899

9 R900-R999

—-

from (ca45co40) (added 10 to each)

0 + 10 = 10 R1 000-R1 099

1 + 10 = 11 R1 100-R1 199

etc.., giving

12 R1 200-R1 399

13 R1 400-R1 599

14 R1 600-R1 999

15 R2 000-R2 499

16 R2 500-R2 999

17 R3 000-R3 999

18 R4 000-R4 999

19 R5 000-R5 999

—-

from (ca45co41) (added 20 to each)

0 + 20 = 20 R6 000-R6 999

1 +20 = 21 R7 000-R7 999

etc..giving

22 R8 000-R8 999

23 R9 000-R9 999

24 R10 000-R10 999

25 R11 000-R11 999

26 R12 000-R13 999

27 R14 000-R15 999

28 R16 000-R19 999

29 R20 000-R24 999

—

from (ca45co42): NB value text filw only shows 1 = Yes, 2 = No. But considering the questionaire and the fact that there is a 0-4 coding:

here, based on other documents will assume:

0 = R25 000 - R29 999, add 30 to become a coding of 30

1 = R30 000 - R39 999, add 30 to become 31

2 = R40 000 + , add 30 to become 32

3 = No personal income, change to coding 0

4 = Refused, change to code 60 (will sort this out later)

1.3.1.12 LSM (universal: 10 groups) (ca46co64):

change group 10 coding from 0 to 10

0 Group 10

1 Group 1

2 Group 2

3 Group 3

4 Group 4

5 Group 5

6 Group 6

7 Group 7

8 Group 8

9 Group 9

1.3.2 For media

1.3.2.1 print:

applied same as in 1995 (issues multiplied by reversed thoroughness) . ended up with 158 vehicles and 2 types (newspapers and magazines).

74 newspapers and 84 magazines (vehicles)

1.3.2.2 radio:

basically similar to '95. in terms of engagement measures: 1: past four weeks; 2: past 7 days; 3: yesterday

Problem: some important stations in the questionaire (RSG and SAFM) but do not appear in the dataset... possibly lost in translation?

Ended with 28 "vehicles"

1.3.2.3 TV:

similar to '95 and radio in terms of engagement measures: 1: past four weeks; 2: past 7 days; 3: yesterday

Ended up with 10 tv vehicles

1.3.2.4 internet:

(first available set after 1995 that includes something on this...)

here first considered "accessed internet" in the past 4 weeks (1=yes, 2 = no); past 7 days (1=yes, 2 = no); and yesterday (1=yes, 2 = no). Note the text file on this is incorrect. Has two 7days. but 'ca38co52 'actually refers to yesterday by the questionaire...

first converted all "no = 2" and "NA" to 0 and then as with radio and tv. simply added these three variables to give first degree of engagement; ie 0, 1, 2, 3:

next, considered time: 1 Up to 30 min 3850 2 31 min - 60 min 3 61 min - 179 min 4 180 min - 360 min 5 361 + min. and changed all NA's to 0.

lastly extracted all “to” that relate to news and information services (ie not banking, email, purchasing, chatting,), specifically:

ca38co41_1 Accessed internet To do banking transactions X

ca38co41_2 Accessed internet To send/receive e-mail X

ca38co41_3 Accessed internet To read a magazine or newspaper on-line

ca38co41_4 Accessed internet To listen to radio (any station(s)) on-line

ca38co41_5 Accessed internet To actually make a purchase X

ca38co41_6 Accessed internet To obtain news

ca38co41_7 Accessed internet To chat on-line X

ca38co41_8 Accessed internet To obtain information about travel

ca38co41_9 Accessed internet To obtain information about sport

ca38co42_0 Accessed internet To obtain information about business and financial matters

ca38co42_1 Accessed internet To obtain information about health, fitness and medical matter

ca38co42_2 Accessed internet To obtain information about entertainment

ca38co42_3 Accessed internet To obtain information about cars/motoring

ca38co42_4 Accessed internet To obtain information about computing and technology

ca38co42_5 Accessed internet To source information on other products/services I may want to

ca38co42_6 Accessed internet To obtain other kinds of information

ca38co42_7 Accessed internet for other X

After some thought : maybe better to use the "whats" (add them up and multiply by the time spent in past seven days) as an engagement figure (did this due to unexpected low correlation with first try).

—-

Finally created two datasets (media_type_02 and media_vehicles_02) to use in exploration. latter uses on the “what” of internet set.

1.4 AMPS 2012

1.4.1 For demographics

1.4.1.1 age (personal: ca56co34): NOT same as 1995

1.4.1.2 sex (demographics: ca91co51a)

1 = male

2 = female

1.4.1.3 education (demographics: ca91co48)

1.4.1.4 come (demographics: ca91co50)

1.4.1.5 race (demographics: ca91co51b) NB codes changed from 2002. Need to standardise before comparing...

1 Black

2 Coloured

3 Indian

4 White

1.4.1.6 province (demographics: ca91co56)

1 Western Cape

2 Northern Cape

3 Free State

4 Eastern Cape

5 Kwazulu-Natal

6 Mpumalanga

7 Limpopo

8 Gauteng

9 North-West

1.4.1.7 metro (demographics: ca46co57):

metro1 (demographics: ca91co57)

1 Cape Town

2 Cape Town Fringe Area

3 Port Elizabeth/Uitenhage

4 East London

5 Durban

6 Bloemfontein

7 Greater Johannesburg (Alex.,JHB,Sandton,Soweto,Randburg)

8 Reef (Urban Gauteng excl JHB,Pta,Vaal)

9 Pretoria (includes North-West Metro)

—

metro2 (cdemographics: ca91co58) becomes after adding 9:

10 Kimberley,

11 Pietermaritzburg

12 Soweto

13 Vaal (De Deur,Meyerton,Sasolburg,Vanderbijlpark,Vereeniging)

as in '95 and '02 need to sort out double count of Soweto....

1.4.1.8 language (demographics: ca91co75)

1 Afrikaans

2 English

3 Zulu

4 Xhosa

5 N. Sotho

6 S. Sotho

7 Tswana

8 Tsonga

9 Venda

10 Swazi

11 Ndebele

12 Other

1.4.1.9 life stages (demographics: ca91co77): NB coding changed from '02

1 At-home singles

2 Young Independent Singles

3 Mature singles

4 Young Couples

5 Mature Couples

6 Young Family

7 Single Parent Family

8 Mature Family

1.4.1.10 marital status (personal: ca56co09):

1 Single

2 Married or living together

3 Widowed

4 Divorced

5 Separated

1.4.1.11 personal income NB.... amounts changed from ’02...to ’12 (assume adjustments made every year??)

combined these to give the following levels:

from (ca57co61) use codes as they are given

1 R 1 - 499

2 R 500 - 599

3 R 600 - 699

4 R 700 - 799

5 R 800 - 899

6 R 900 - 999

7 R 1 000 - 1 099

8 R 1 100 - 1 199

9 R 1 200 - 1 399

—-

from (ca57co62) (added 10 to each)

0 + 10 = 10 R 1400 - 1599

1 + 10 = 11 R 1 600 - 1 999

etc.., giving

10 R 1400 - 1599

11 R 1 600 - 1 999

12 R 2 000 - 2 499

13 R 2 500 - 2 999

14 R 3 000 - 3 999

15 R 4 000 - 4 999

16 R 5 000 - 5 999

17 R 6 000 - 6 999

18 R 7 000 - 7 999

19 R 8 000 - 8 999

—-

from (ca57co63) (added 20 to each)

0 + 20 = 20 R 9 000 - 9 999

1 + 20 = 21 R 10 000 - 10 999

etc..giving

20 R 9 000 - 9 999

21 R 10 000 - 10 999

22 R 11 000 - 11 999

23 R 12 000 - 13 999

24 R 14 000 - 15 999

25 R 16 000 - 19 999

26 R 20 000 - 24 999

27 R 25 000 - 29 999

28 R 30 000 - 39 999

29 R 40 000 - 49 999

—

from (ca57co64): based on questionarie

+30

30 R 50 000+

31 No personal income: change coding to 0

32 Refused (cange to code = 60....will sort out later)

1.4.1.12 LSM (universal: 10 groups) (lsm: ca91co64):

changed group 10 coding from 0 to 10

1 Group 1

2 Group 2

3 Group 3

4 Group 4

5 Group 5

6 Group 6

7 Group 7

8 Group 8

9 Group 9

10 Group 10

1.4.1.13 lifestyle (lsm: ca58co39 SAARF Lifestyle Groups Total Groups LSM Groups 1-10)

metadata file on values is different. Will accept the questionaire p10 categories (coding) add 1 to get rid of 0 to get

1: None

2: Cell Sophisticates

3: Sports

4: Gamers

5: Outdoors

6: Avid Readers

7: Traditionals

8: Cell Fundamentals

9: Homebodies

10:Bars & Betters

11: Showgoers

12: Gardeners

1.4.1.14 attitudes (lsm: ca67co10). NB check on documentation to understand how this is collected., Also may want to consider median groups information. Also check metadata file on attitude responses... maybe can do something with this.

added 1 to get rid of zero (NB here 5703 NAs)

1: None

2: Now Generation

3: Nation Builders

4: Distants

5: Rooted

6: Global Citizens

1.4.2 For media

1.4.2.1 print:

applied same as in 1995/2002 (issues multiplied by reversed thoroughness) . ended up with 172 vehicles and 2 types (newspapers and magazines).

54 newspapers and 118 magazines (vehicles)

1.4.2.2 radio:

basically similar to '95 and '02 in terms of engagement measures: 1: past four weeks; 2: past 7 days; 3: yesterday

Problem: quite different numbers of stations. Needed some cleaning up.

Ended with 97 “vehicles”

1.4.2.3 TV:

similar to ’95 & ’02 and radio in terms of engagement measures: 0 : nono; 1: past four weeks; 2: past 7 days; 3: yesterday

Ended up with 12 tv vehicles. NB, here could use tv viewing time...\

1.4.2.4 internet:

here as in ’02 first considered “accessed internet” in the past 12 months(1=yes, 2 = no), 4 weeks (1=yes, 2 = no); past 7 days (1=yes, 2 = no); and yesterday (1=yes, 2 = no).

Converted all “no = 2” and “NA” to 0 and then as with radio and tv. simply added these four variables to give first degree of engagement; ie 0, 1, 2, 3, 4:

no information on time spent on internet in ’12

In effort to get better idea of the “what” respondents do on the internet (note some similar others quite different to ’02)

did not consider non information/news/media uses (eg, bank transactions and emails) specifically: (NB here could include social media)

CELLPHONE SPECIFIC (could come back to this...for now, not included)

ca49co28 Cellphone activity - access the Internet or Web from your cellphone

ca49co29 Cellphone activity - facebook

ca49co30 Cellphone activity - twitter

ca49co31 Cellphone activity - other social networking

ca49co38 Cellphone activity - watch television on your cellphone

ca49co40 Cellphone activity - read a newspaper/magazine on your cellphone

ca49co41 Cellphone activity - listen to a radio station on your cellphone

ca49co43 Cellphone activity - subscribe to receive content or services on your cellphone

CELLPHONE OR COMPUTER

ca49co55 Use a computer or a cellphone to search (e.g. googling)

ca49co58 Use a computer or a cellphone for social networking

ca49co63 Use a computer or a cellphone for reading a newspaper/magazine on-line

ca49co64 Use a computer or a cellphone to read/access current news/articles on-line

ca49co69 Use a computer or a cellphone to watch television on-line

ca49co71 Use a computer or a cellphone to listen to the radio on-line

Coding:

1 I do this activity using a computer

2 I do this activity using a cellphone

3 I dont do this activity at all

4 I do this activity using both computer and cellphone

Considered that my interest is in using the internet...not so much what my device is?? in the context of my study, does it really matter?

So, # change NA and 3 = 0; 1,2,4 = 1

After some thought... also considered (although quite different in '02) : to get better "engagement" figure...added these "what its used for" and multiplied by the first engagement figure (ie 0-4) on the basis that someone who accessed the web yesterday is generally going to be more engaged and the number of different activities are a further indication of .... need to think about this. Is this a reasonable measure of relative engagement with the medium....(multiplying internet_level1 with sum of internet_level2)

—-

Finally created two datasets (media_type_12 and media_vehicles_12) to use in later exploration. latter uses on the "what" of internet set.

Chapter 2

Exploring the Data

2.1 Exploring AMPS 1995

After discussion with Ian: Consider a level one... total engagement per media type (newspapers, magazines, radio, tv); and try to identify relative combinations High and Low...Ie four combination 2... ie 16 levels?

2.1.1 For level one

...split total_engagement_print into same for magazines (weekly, fortnightly, monthly) and newspapers (weekly and daily)

scale the type data and consider kmeans (identified 6 levels)

a tree to interpret the 6 levels:

1: Mainly TV (low on all others)

2: TV, Mags, Newspapers (high on all three)

3: Low radio, tv, newspapers (low on all media)

4: Mainly radio (low newspapers, tv)

5: Mainly TV and Newspapers

6: Mainly TV and Radio

next want to consider prediction of these from demographic and other media data: check out R file for waht I did here... basically used rantree and then random forests ... before considered hclustering before settling on kmeans.

also factor analysis.. want to consider using approach used in amps 2002 ie principal components and the scores of PCs as relative values of cross media use... need to try and make better sense of trhis...

So what the fuck now

2.2 Exploring AMPS 2002

2.3 Exploring AMPS 2012

2.3.1 The 2012 Dataset

The 2012 dataset used for this exploration consists of responses from 25 108 participant (of which 15 162 reside in metropolitan areas, which constitutes the focus of this project). The respondents were based on a stratified random survey, with proportions informed by the StatsSA census. The following describes the variables that were considered in this exploration and, where relevant, the transformations that were effected are described:

- Respondent identities (questionaire numbers 'qn')
- Population weights ('pwgt') NOTE: For now, these have not been considered.I still need to consider how or if to use these (advice ?)
- Demographic Variables (columns 3:16). Not all the demographic variables extracted in for the exploratory set were used. For example, gender ('sex') appeared to be a poor differentiator of media use; and household income ('hh_inc') was preferred to personal income due to the possibility that many young people may not yet be earning but their media consumption is afforded by household income. The variables used in this exploration are:
  - age brackets ('age': ordered factors)

    Aggregated levels:

    1: 15 - 24

    2: 25 - 44

    3: 45 - 54

    4: 55+
  - education level ('edu': ordered factors)

    Aggregated levels:

    1. <matric (1,2,3,4)

    2. matric (5)

    3. >matric (6,7,8)
  - household income level ('hh_inc': ordered factors)

    Aggregated levels:

    1 <R2500 (1,2,3,4)

    2 R2 500 - R6 999 (5,6)

3 R7 000-11 999 (7)

4 >=R12 000 (8)

- race ('race': unordered factors)

  1 Black

  2 Coloured

  3 Indian

  4 White

- metropolitan area ('metro': unordered factors)

  0: Not Metropolitan - These respondents were extracted to ensure a metropolitan focus due to the fact that media options in rural areas are likely to skewed.

  1: Cape Town

  2: Cape Town Fringe Area

  3: Port Elizabeth/Uitenhage

  4: East London

  5: Durban

  6: Bloemfontein

  7: Greater Johannesburg (Alex.,JHB,Sandton,Randburg)

  8: Reef (Urban Gauteng excl JHB,Pta,Vaal)

  9: Pretoria (includes North-West Metro)

  10: Kimberley,

  11: Pietermaritzburg

  12: Soweto

  13: Vaal (De Deur,Meyerton,Sasolburg,Vanderbijlpark,Vereeniging)

- home language for full dataset ('lang': unordered factors):

  1: Afrikaans

  2: English

  3: Zulu

  4: Xhosa

  5: N. Sotho

  6: S. Sotho

  7: Tswana

  8: Tsonga

  9 :Venda

  10: Swazi

  11: Ndebele

  12: Other

- home language for cape town only ('lang_ct': unordered factors):

  1: Afrikaans

  2: English

  3: Xhosa

  4: Other

- home language for Johannesburg only ('lang_jhb': unordered factors): NOTE: To be completed

- lifestages ('lifestages: unordered factors)

  1: At-home singles

  2: Young Independent Singles

  3: Mature singles

  4: Young Couples

  5: Mature Couples

  6: Young Family

  7: Single Parent Family

  8: Mature Family

- marital status ('mar_status': unordered factors)

  1: Single

  2: Married or living together

  3: Widowed

  4: Divorced

  5: Separated

- living standards measure ('lsm': ordered factors)

  Aggregated groups

  1 Groups 1& 2

  2 Groups 3 & 4

  3 Groups 5 & 6

  4 Groups 7 & 8

  5 Groups 9 & 10

- lifestyle ('lifestyle': unordered factors)

  1: None

  2: Cell Sophisticates

  3: Sports

  4: Gamers

  5: Outdoors

    6: Avid Readers
    7: Traditionals
    8: Cell Fundamentals
    9: Homebodies
    10:Bars & Betters
    11: Showgoers
    12: Gardeners

  - attitudes ('attitudes': unordered factors)

    1: None
    2: Now Generation
    3: Nation Builders
    4: Distants_Survivors
    5: Distants_Established
    6: Rooted
    7: Global Citizens

- Aggregated media 'type' engagement levels consists of scaled values reflecting relative engagement for 5 media types (columns 17:21).

  - newspapers:
    - How many issues of a particular print vehicle does the respondent read in a given issue period. For example, dailies would be 0-5 and montly magazines would be considered over a six week period (ie 0-6).
    - How thoroughly the newspapers is normally read: A scale of 1-6 in growing thoroughness from: 1 = "just glance at parts of it" to; 6 = "Read right through it".
    - A per vehicle engagement scale was created by multiplying the number of issues a respondent normally reads in any given issue period by how thoroughly that product is read. The scale will differ depending on the issue periods.
    - For a type-level engagement value, an aggregate over all the newspapers was created by summing the vehicle engagement values (no of issues x thoroughness) and standardising (mean = 0, sd = 1) the sum.
  - magazines:
    - The same as in newspapers above
  - tv:
    - Relative engagement was determinted by how recently a respondent personally watched a particular channel. Resulting in: 0 = "not at all"; 1 = "In the past four weeks"; 2 = "in the past 7 days"; and 3 = "yesterday".

    - An aggregate value over all tv was created by summing the per channel (vehicle) values and standardising (mean = 0, sd = 1).
  - radio:
    - The same as in tv
  - internet:
    - Analogous to radio and tv, a first level of engagement was determined by considering the recency of accessing the internet. In particular 0 = “not”, 1 = “in the past 12 months”, 2 = “in the past four weeks”, 3 = “in the past 7 days”, 4 = “yesterday”.
    - The focus of this project is primarily on media for the purposes of news, information and entertainment. Accordingly the use of a computer or cellphone for example accessing email, banking, dating or shopping was not considered. Accessing the internet for purposes of search, social networking, print media sites, listening to the radio, accessing news, or watching tv were extracted.
    - To create a relative engagement level for each of these purposes, the recency value (0-4) was muliplied by 0/1 for each of the purposes.
    - An aggregate value for all the internet purposes was created by summing the columns for the six relative engagement values. The final vector of total engagement value was standardised (mean = 0, sd = 1).

- Variablies reflecting relative engagement values per media vehicle (columns 22:309): For print, these values would reflect the number of issues multiplied by thoroughness for each print vehicle; for electronic (radio & tv) the values would reflect recency of viewing or listening to the individual channel/station as a proxy for engagement; and for each of the six purposes of using the internet, recency would again serve as a proxy of relative engagement. Examples of vehicles include:
  - Beeld
  - Alex FM
  - SABC 1
  - MNet (main)
  - internet for search (int_search)
  - internet for news (int_news)
  - etc...

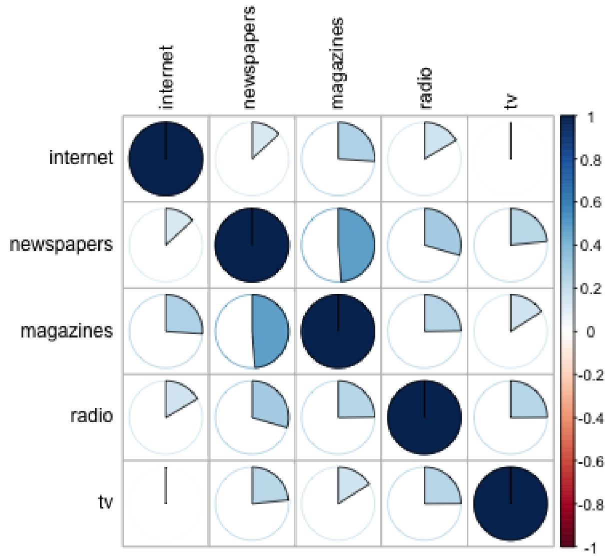
2.3.2 Exploration of Media Type

In this section relative engagement of the five media types (newspapers, magazines, tv, radio, internet) were considered to investigate possible latent structures that could group the respondents by media type.

2.3.2.1 Correlation by Media Type

To gain some sense of the relationship between the five types, correlations were considered in Figure 2.1 below. The correlation matrix shows relatively strong positive correlation between newspapers and magazines, generally stronger cross media correlations for magazines, but very weak correlation between tv and internet engagement. Possible reasons for these values will be considered in more detail below.

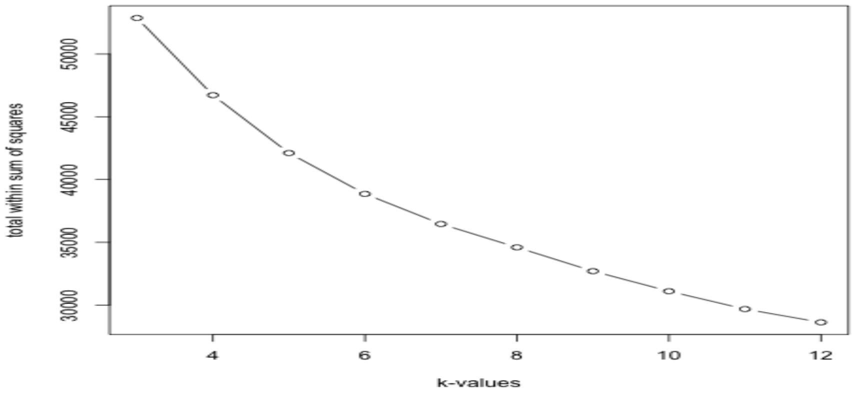
Figure 2.1: Correlation Matrix of Media Types (2012)



2.3.2.2 Clustering of Media Type

Although hierarchical clustering in dendograms was first considered, the results were messy and unhelpful. *Kmeans* proved to be more helpful. Figure 2.2 shows declining values of total within cluster sum of squares (wss) for increasing values of k. Although there does not appear to be a clear candidate, a consideration of cluster balance in terms of quantity per cluster resulted in a selection of 5 clusters.

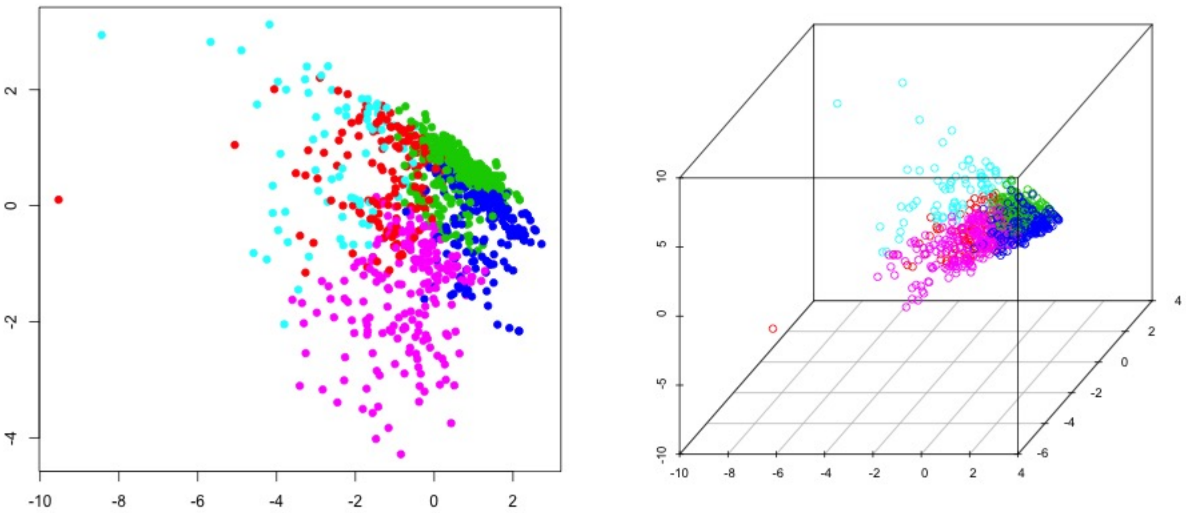
Figure 2.2: Plot of within sum of squares for different values of k



The five centers (clusters) were added to the dataset as factors. To consider the separation of these clusters, Multidimensional Scaling (MDS) was applied to a random subset of 1000 respondents. The plots in figure 2.3 illustrate reasonable seperation of clusters. Although the 3D version on the scaling does not appear to be an improvement, dynamic capabilities of R's 'rgl' library made it is possible to show reasonably good cluster separation.

Additional Self Organising Maps were considered, but these did not prove to be more helpful than the MDS visualisations already applied. Using documentation supplied by Saarf on the principal component methods used to identify its own Media Groups, a similar approach was applied to the media type engagement variables. This did not prove particularly helpful since in most cases second, third and more components could not simply be ignored. It was not clear what the Media Groups used as initial media type values.

Figure 2.3: 2D and 3D plots of Multidimensional Scaling of 5 clusters



In order to confirm the relative value of these clusters, some predictions were considered.

After randomly separating the full dataset into training (70%) and test (30%) sets, a simple random forest (RF) as well as linear discriminant analysis (LDA) were applied with the cluster as outcome categories and media type variables as predictors. Given that these were used to construct the cluster categories, one would expect relatively good predictions. The simple RF showed an accuracy of ~97% on the test sample, while the LDA proved only slightly less accurate at ~91% on the test sample.

To explore the strength of relationship between clusters and the demographic variables, similar predictions were run, but this time with only demographic variables as predictors against the media type cluster categories. The RF showed an accuracy of ~44% on the test sample, while the LDA reflected a slightly improved test sample accuracy of ~48%. Although this may not appear as impressive when compared with the previous predictions, it does give some cause to believe that the clustering into five groups may have value when compared with simple random selections.

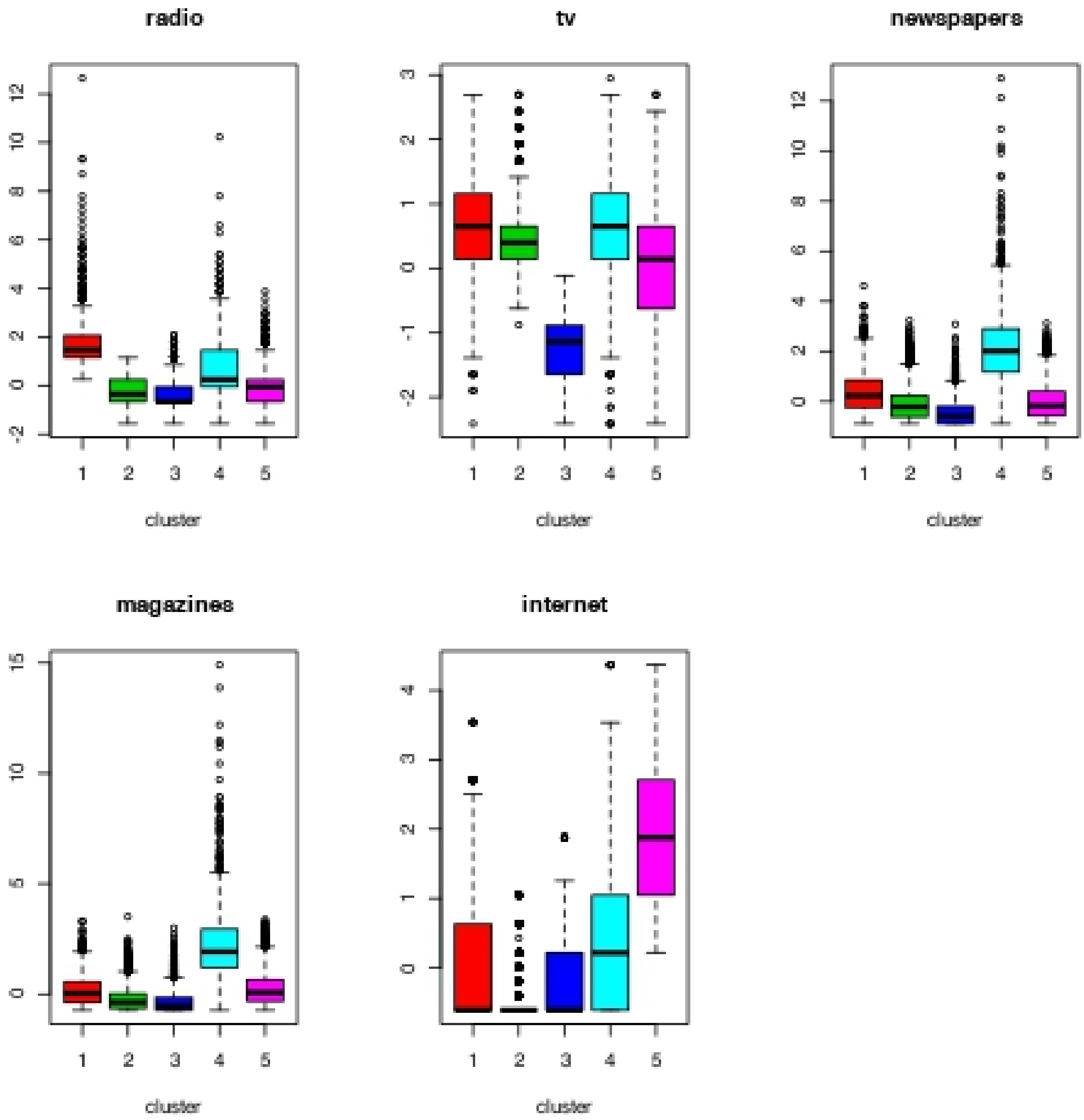
2.3.2.3 Qualitative Descriptions of Type

Following the esablishment of these 5 clusters of media types, an attempt was made to describe them qualitatively in terms of media type use as well as demographic profiles for each of the clusters. From figure 2.4, the following deductions shown in table 2.1are made:

Table 2.1: Interpreting Media Type Clusters

Cluster	Media-Type Descriptor	Comments
1 (Red)	"Radio and TV above all else"	Highest engagement on radio
2 (Green)	"TV is all"	Low engagement on all (especially internet) with the
3 (Dark Blue)	"Minimal All Media"	Lowest engagement on all media types
4 (Light Blue)	"Maximum All Media"	Relatively high engagement on all media types, mai
5 (Pink)	"Big on Internet and some TV"	Highest engagement on internet, otherwise only TV

Figure 2.4: Boxplots per Media Type Clusters of Relative Engagement



An attempt was also made to identify either systematic or significant demographic associations with the 5 media-type clusters by considering figures 2.5 and 2.6below. The results are reflected in table 2.2.

Table 2.2: Interpreting Media Type Profiles

CLUSTER	MEDIA-TYPE	DEMOGRAPHIC PROFILE
1 (Red)	“Radio and TV above all else”	mainly black, matric, 15-44, LSM 5-8
2 (Green)	“TV is all”	mainly coloured, <matric, 55+, LSM 5-8, slightly m
3 (Dark Blue)	“Minimal All Media”	mainly white, <matric, LSM 3-4
4 (Light Blue)	“Maximum All Media”	all races, >matric, 25-54, LSM 7-10, >12000
5 (Pink)	“Big on Internet and some TV”	mainly white, matric and >matric, 15-44, LSM 9-10

Figure 2.5: Plots of Race, Education, LSM vs Media-Type Clusters

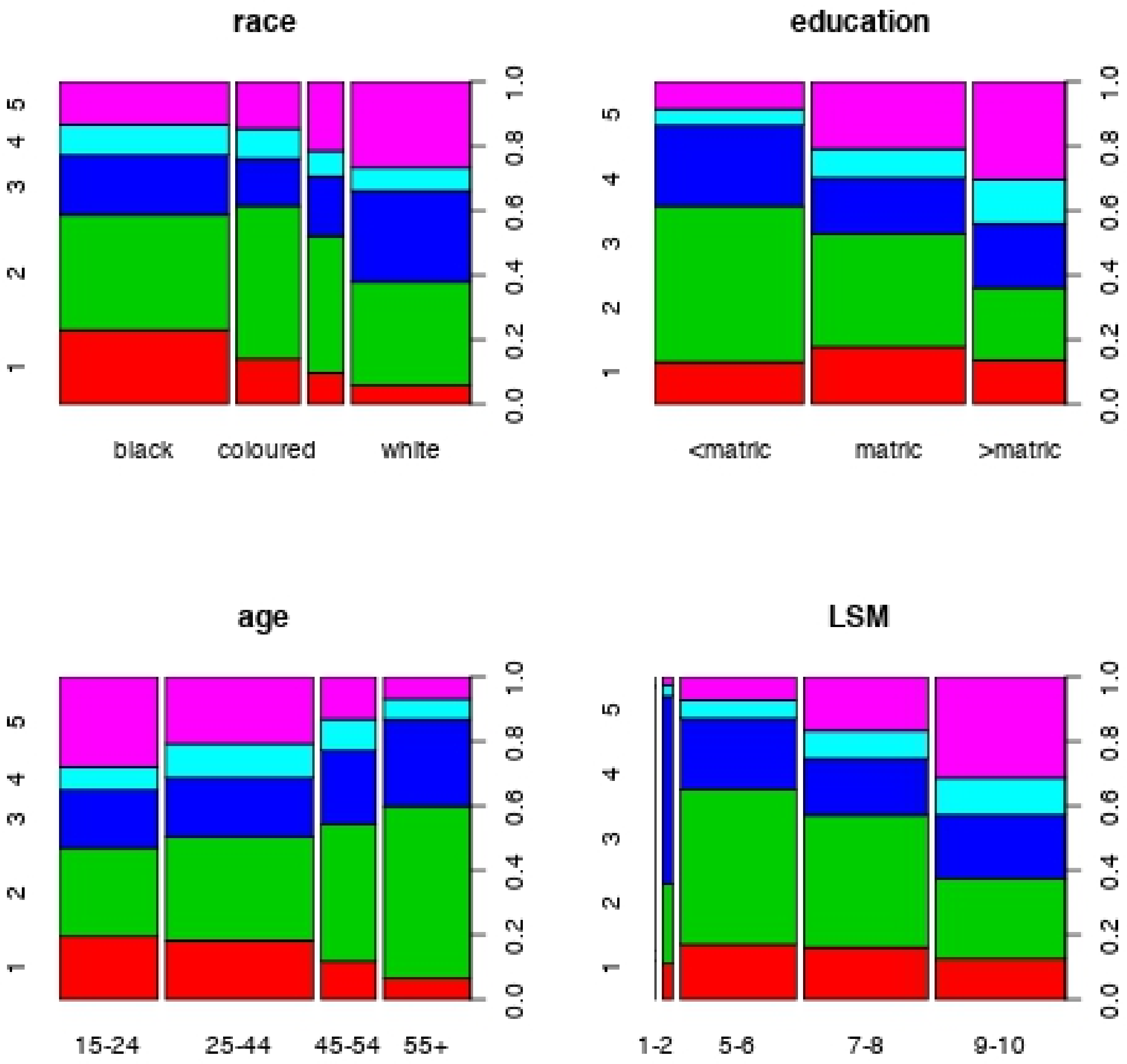
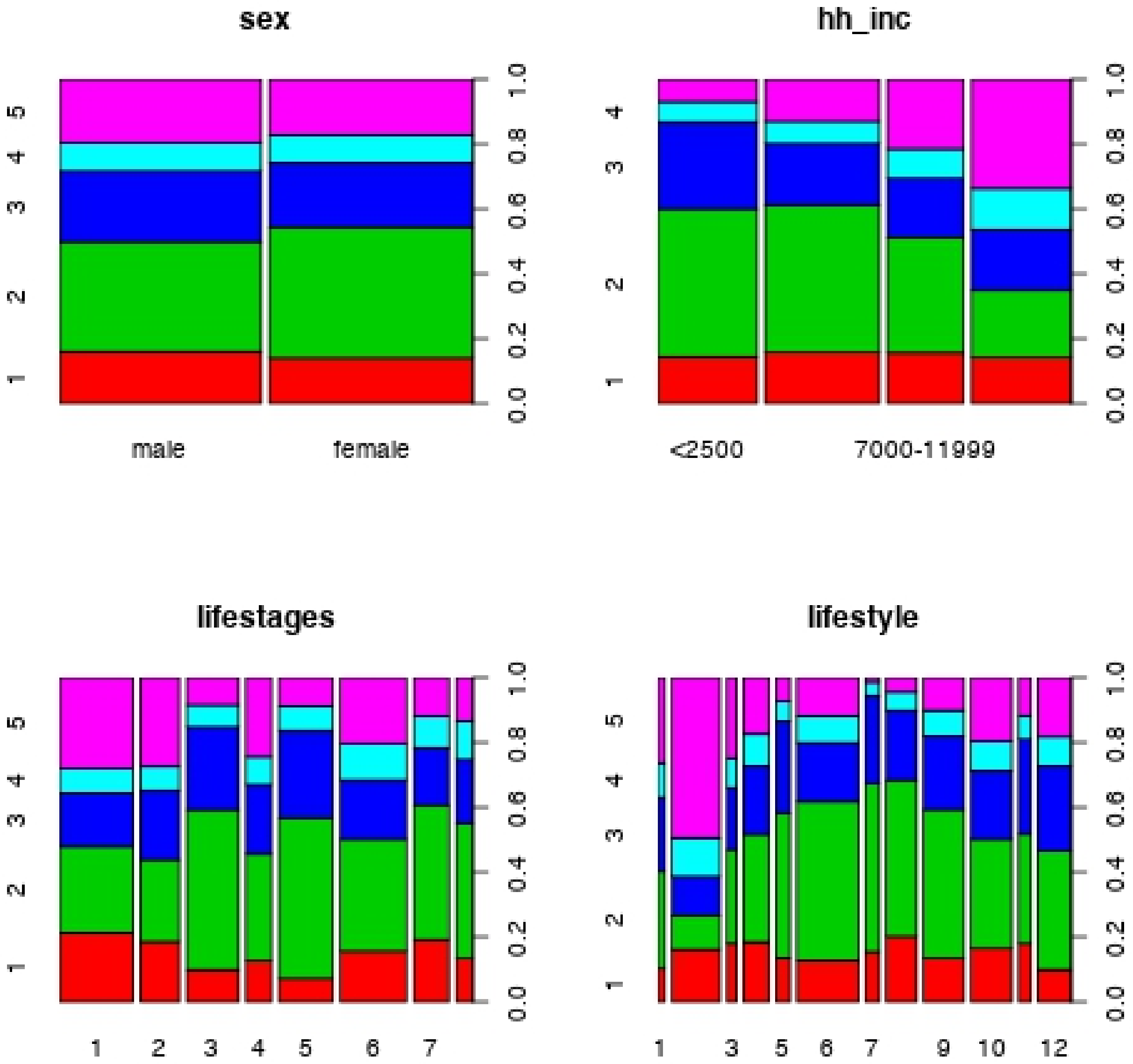


Figure 2.6: Plots of Sex, Household Income, Lifestages and Lifestyles vs Media-Type Clusters



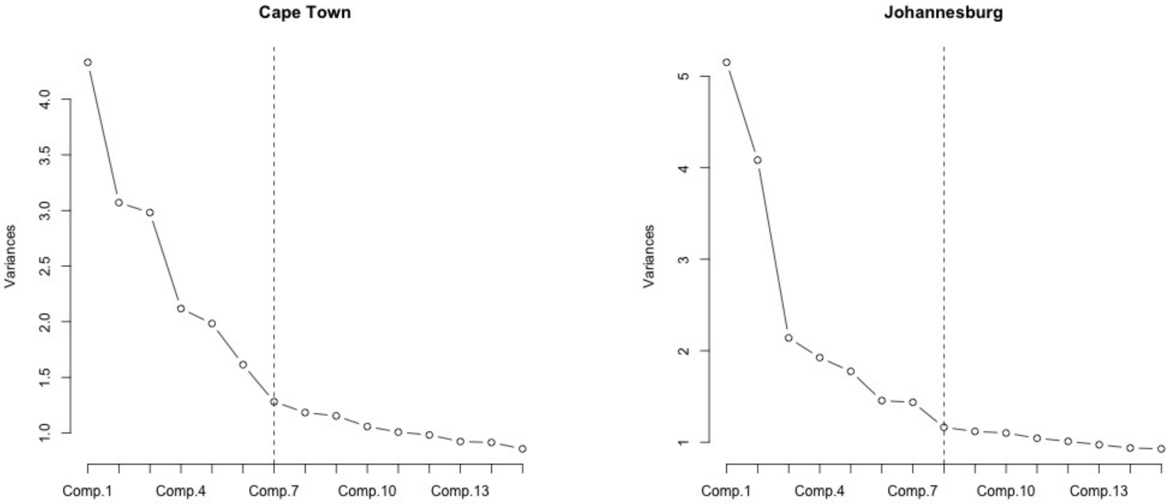
2.3.3 Exploration of Media Vehicles for Cape Town and Johannesburg

Since specific media vehicle consumption depends largely on any particular metropolis, the dataset was subsetted into two datasets for Cape Town (metro = 1) and in Johannesburg (metro = 7). Furthermore, only those media vehicles with a reasonable penetration in the two cities were included. Accordingly, only variables with more than 10% non-zero values were included to create a Cape Town dataset of 38 media vehicles and a Johannesburg set of 43 vehicles. To simplify interpretation, the coding of the language variable was changed for Cape Town to reflect the dominant situation in that city (1 = Afrikaans; 2 = English; 3 =

Xhosa; 4 = Other).

After the data was scaled (mean = 0, sd = 1) principal components analyses (PCA) were done to get some idea of latent structure in the dataset. Figure 2.7 below shows a relatively clear "elbow" at seven components for Cape Town, reflecting a cumulative variance proportion of ~46%. The screeplot for Johannesburg shows an "elbow at eight components with a cumulative variance proportion of ~45%. The number of components served as a guideline to identify the number of factors to extract in a factor analysis on the same variables.

Figure 2.7: Scree Plot of Principal Components for Cape Town & Johannesburg 2012



A factor analysis for seven factors for Cape Town resulted in a cumulative variance explained by the seven factors of ~35% and for Johannesburg's eight factors explained ~34% of the variance.

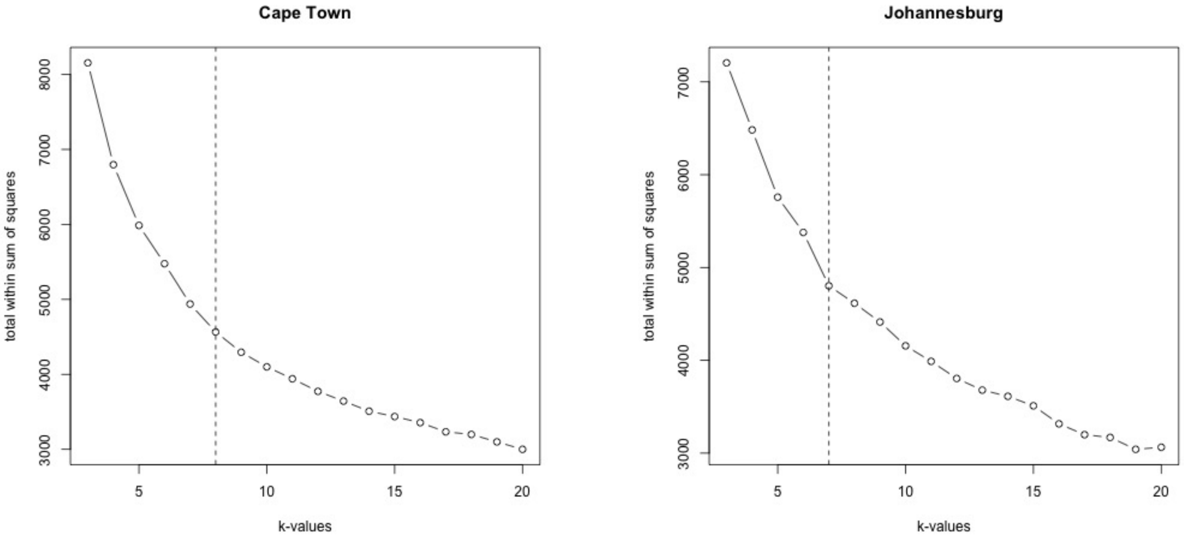
The loadings of the media vehicle variables on the seven factors were considered in order to identify the top 10 media vehicles that dominate each of the seven factors for Cape Town and eight for Johannesburg. Table 2.3 below identify the media vehicles with their loadings on each of the factors per city.

To identify media groupings *kmeans* analysis was done on the two sets of factor scores to ensure that as much of the information contained in the scores was used in determing groups that are based on proximity with regard to their factor scores. To identify an optimum number of centers, various k-values were considered in terms of their total within sum of squares (wss). From figure 2.8 below, inflexion points could be considered at eight centers for Cape Town and at seven for Johannesburg.

Table 2.3: Top Media Vehicles Loadings per Factor

CAPE TOWN						JOHANNESBURG					
FACT	VEHICLE	LOAD	FACT	VEHICLE	LOAD	FACT	VEHICLE	LOAD	FACT	VEHICLE	LOAD
1	SABC_3	0.77	5	HUISgenoot	0.71	1	DRUM	0.54	5	int_news	0.97
	e_tv	0.67		DIE_BURGER	0.69		SundayWorld	0.53		int_print	0.7
	SABC_2	0.67		Rapport_Sun	0.61		METRO_FM	0.53		int_search	0.24
	SABC_1	0.66		Son	0.43		Sowetan	0.51		int_social	0.18
	Son	0.14		Radio_Tygerberg	0.31		TRUE_LOVE	0.5		DSTV	0.07
	tvplus	0.13		SABC_2	0.18		Sunday_SUN	0.48		Mens_Health	0.06
	DAILY_Voice	0.13		Kfm	0.17		DAILY_SUN	0.47		Mail_Guardian	0.06
	Cape_Argus	0.13		M_Net_Main	0.12		KAYA_FM	0.46		X5FM	0.05
	YOU	0.09		SABC_3	0.09		Soweto_TV	0.41		Auto_Trader	0.05
	Radio_Tygerberg	0.09		DSTV	0.08		Move	0.41		Highveld_Stereo	0.05
2	int_search	0.89	6	int_print	0.83	2	SABC_2	0.73	6	KICKOFF	0.7
	int_social	0.85		int_news	0.8		SABC_3	0.69		SOCCER_LADUMA	0.7
	int_news	0.38		int_social	0.2		SABC_1	0.64		DAILY_SUN	0.31
	X5FM	0.34		int_search	0.19		e_tv	0.58		Sowetan	0.23
	int_print	0.29		M_Net_Main	0.13		DAILY_SUN	0.21		Sunday_SUN	0.21
	Kfm	0.22		METRO_FM	0.12		METRO_FM	0.2		SundayWorld	0.21
	Good_Hope	0.2		X5FM	0.1		Soweto_TV	0.17		YFM	0.18
	M_Net_Main	0.19		Sunday_Times	0.09		KAYA_FM	0.15		Mens_Health	0.15
	DSTV	0.19		Mens_Health	0.07		YFM	0.11		car	0.13
	Mens_Health	0.13		DSTV	0.06		Ukhozi_FM	0.1		City_Press_Sun	0.09
3	Umhlobo_Wenene_FM	0.65	7	DAILY_Voice	0.67	3	Sunday_Times	0.56	7	DSTV	0.71
	DRUM	0.58		Son	0.56		The_Star	0.49		Premium_Compact_dish	0.47
	METRO_FM	0.57		Heart_FM	0.49		Mail_Guardian	0.38		M_Net_Main	0.46
	DAILY_SUN	0.56		Good_Hope	0.29		The_Times	0.36		Soweto_TV	0.19
	KICKOFF	0.49		Edgars_CLUB	0.14		Talk_Radio_702	0.36		X5FM	0.19
	Jet_Club	0.34		people	0.14		City_Press_Sun	0.33		Highveld_Stereo	0.16
	SABC_1	0.33		SABC_3	0.14		Mens_Health	0.32		METRO_FM	0.13
	Sunday_Times	0.19		e_tv	0.13		Getaway	0.28		KAYA_FM	0.13
	DAILY_Voice	0.16		Voice_of_the_Cape	0.13		SundayWorld	0.26		Sunday_Times	0.11
	tvplus	0.1		Premium_Compact_dish	0.13		int_search	0.24		YFM	0.11
4	Cape_Argus	0.51				4	int_social	0.85	8	YOU	0.52
	YOU	0.5					int_search	0.74		people	0.46
	Sunday_Times	0.47					X5FM	0.26		FAIRLADY	0.4
	CAPE_TIMES	0.41					int_print	0.24		COSMOPOLITAN	0.32
	DSTV	0.36					M_Net_Main	0.18		Highveld_Stereo	0.32
	Premium_Compact_dish	0.35					YFM	0.17		DRUM	0.31
	FAIRLADY	0.32					int_news	0.17		Move	0.28
	Weekend_Argus_Sat	0.32					Highveld_Stereo	0.15		TRUE_LOVE	0.25
	M_Net_Main	0.31					Jacaranda	0.13		The_Citizen	0.16
	people	0.28					METRO_FM	0.11		Sunday_Times	0.15

Figure 2.8: *Kmeans* Total Within Sum of Squares for Various Centers (k-values)



2.3.3.1 Profiling Media Groups and Identifying Media Repertoires

The eight media groups for Cape Town and seven for Johannesburg were considered against relative engagement of the five media-types, shown in figure 2.9, as well against various demographic variables, shown in figures 2.10, 2.11, and 2.12 below. Table 2.4 below summarises an interpretation of these plots for the eight Cape Town media groups, while table ... summarises an interpretation for the seven Johannesburg media groups

Table 2.4: Summary Profiles of Media Groups for Cape Town

MEDIA GROUP	MEDIA-TYPE	DOMINANT DEMOGRAPHICS	SUMMARY
1 (black)	no internet, otherwise average media, cluster 1/2 then 4	45-54, <matric, <6999, coloured, Afrikaans, lsm 5-8	older, coloured, low income, low education, TV is big but also other media
2 (red)	no internet, lowest print, some tv & radio, cluster 3	mainly younger, more male, <matric, <2500, black, Xhosa, lsm 1-6	younger, black, low income, low education, min media
3 (green)	highest for all media, internet medium cluster 4 then 1	15-44,<6999, black, Xhosa, lsm 1-8	younger, black, low to mid income, max media
4 (dark blue)	higher print, low internet, average radio & tv, cluster 4	+55, >7000, white, Afrikaans, lsm 7-10	older, white, afrikaans, higer income, max media user
5 (light blue)	second highest internet, low print, average radio & tv, cluster 5	15-44, more female, =>matric, >7000, English/Afr, lsm 7-10	younger, higher education, higher income, big on internet
6 (pink)	highest internet, cluster 5	15-44, more male, >matric, >= 12000, mainly white, English, lsm 7-10	younger, white, english, high income, big on internet
7 (yellow)	no internet, low print, some radio & tv, cluster 2	+55, more female, <matric, <12000, coloured and indian, Afr/Eng, lsm 5-8	older, low education, mid to low income, coloured/indian, TV dominates
8 (grey)	low tv, average radio, print, higher internet, cluster 3	>=matric, >= 12000, white/indian, English, lsm 9-10	white/indian, higher education and high income, min media user

Table 2.5: Summary Profiles of Media Groups for Johannesburg

MEDIA GROUP	MEDIA-TYPE	DOMINANT DEMOGRAPHICS	SUMMARY
1 (red)	no internet, otherwise average media, mainly cluster 2	55+, more female, lower education (<matric), low income (mostly <2500), other than white, lower lsm (esp 3-6)	older, female, low education, low income, mainly TV
2 (green)	highest internet, otherwise low to average other media, mainly cluster 5	25-44, more male, higher education (>matric), higher income (>=12000), more white, lsm 9-10	younger, male, higher education, higher income, more white, lsm 9-10, big on internet
3 (dark blue)	high all media, higherston radio and magazines, mainly cluster 4	15-44, more female, matric, all income, black, lsm 5-8	younger, female, all income, black, lsm 5-8, max media user
4 (light blue)	average radio, low print, second highest internet, mainly cluster 5	younger (15-24), better education >matric, higher income >12000, higher lsm (mainly 9-10)	young, well educated, high income, big on internet
5 (pink)	highest on newspapers, pretty average otherwise, mainly cluster 4 with a bit of cluster 1	14-44, largely male, lower income <7000, black, lsm 5-8,	younger, largely male, lower income, black, all media
6 (yellow)	tv high, low internet, average otherwise, mainly cluster 2 followed closely by cluster 1.	more older (+55), slightly more female, lower education <=matric, income >7000, mainly coloured and indian, higher lsm 7-10,	older, lower education, higher income, mainly coloured and indian, mainly TV and radio
7 (grey)	lowest on tv, low all media, some internet, dominant cluster 3	45+, slightly more educated, >12000, white, higher lsm (esp 9-10)	older, better educated, higher income, white, lsm 9-10, min media user

Figure 2.9: Plots of Media Groups vs Relative Engagement by Media Type

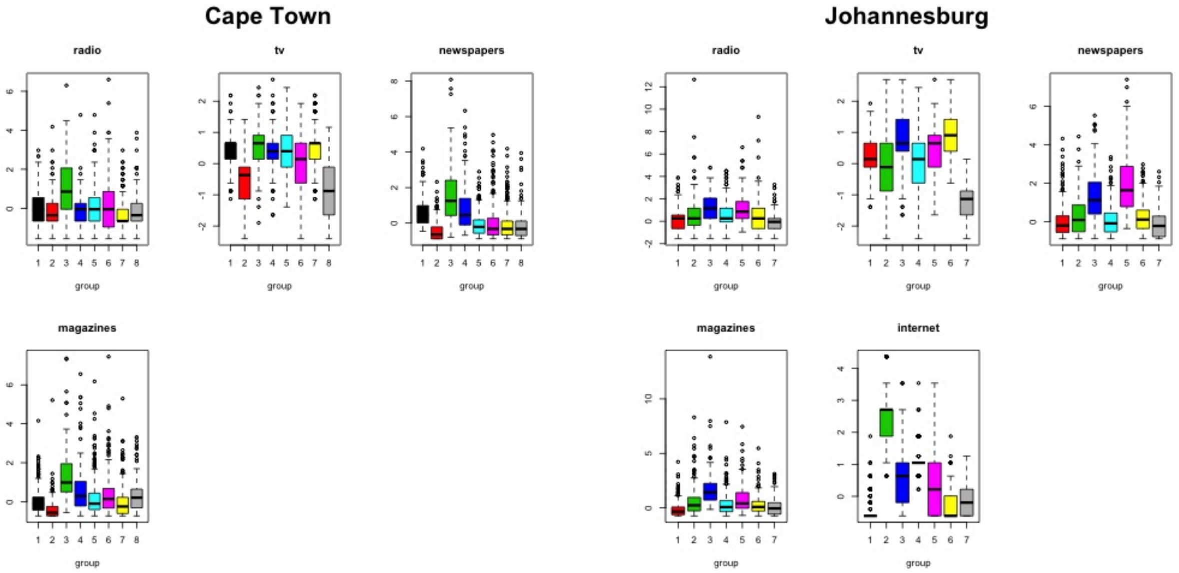


Figure 2.10: Media Group vs Age, Sex, Household Income and Education

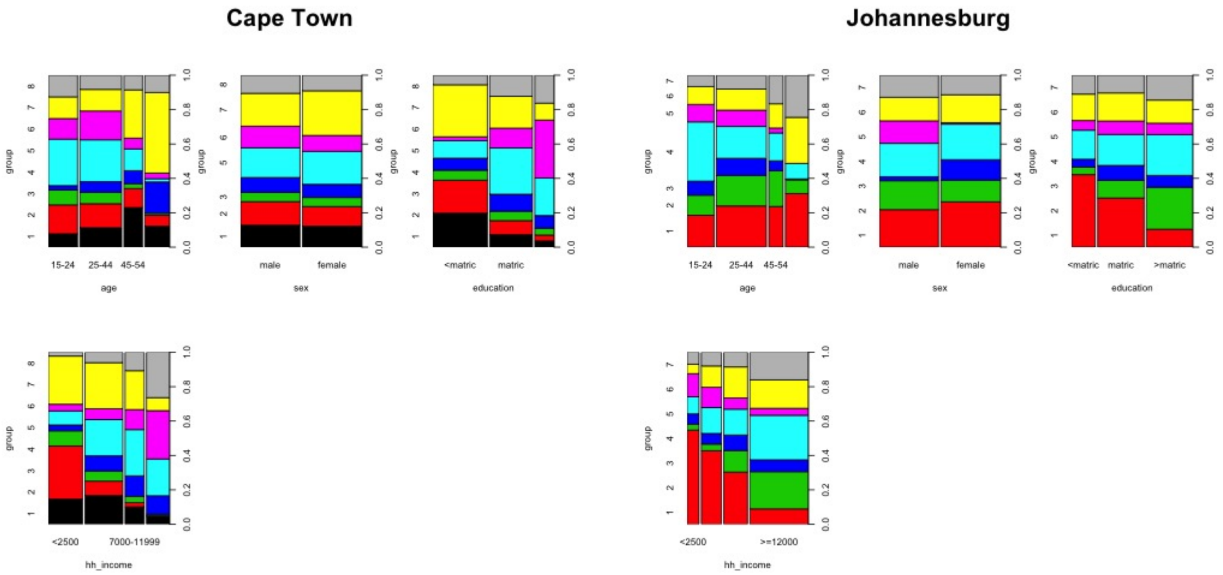


Figure 2.11: Media Group vs Race, Language, LSM, and Media-Type Cluster

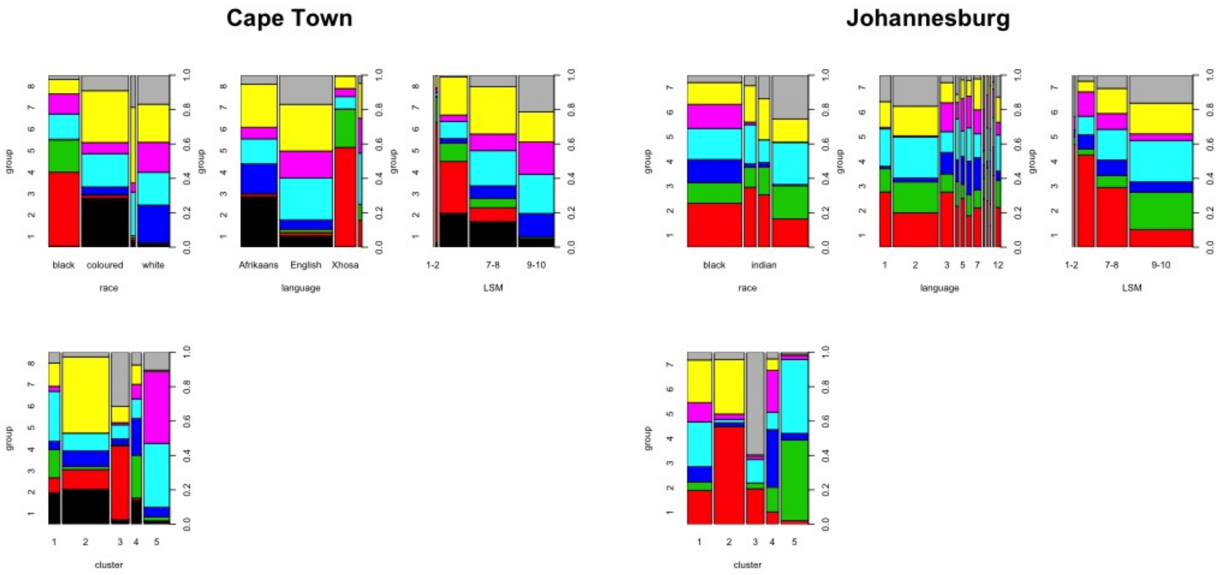
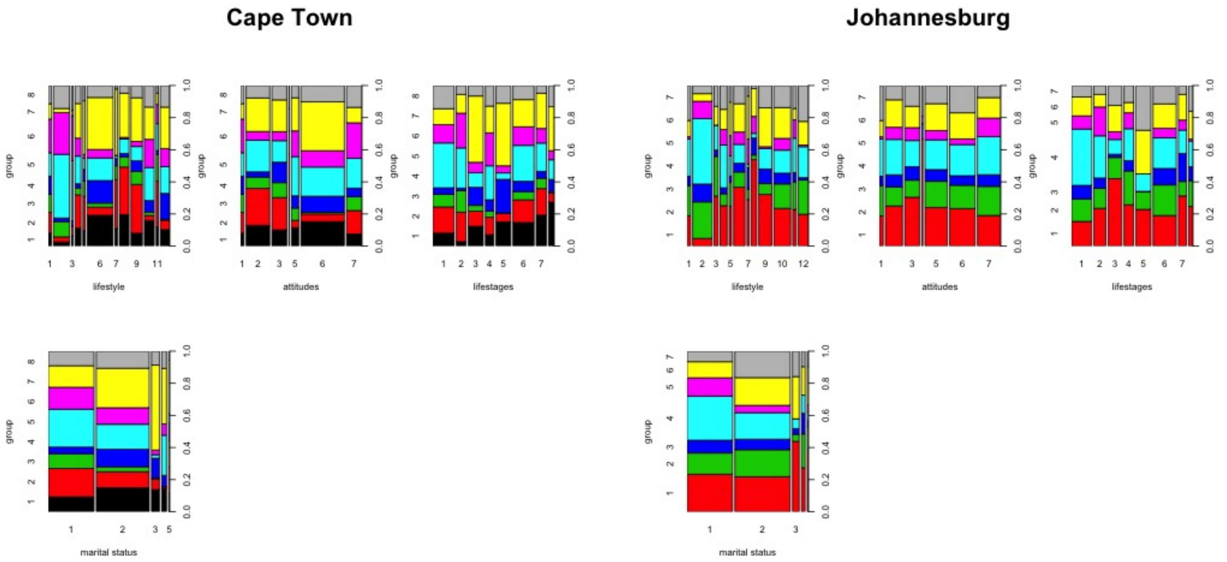


Figure 2.12: Media Groups vs Lifestyle, Attitude, Lifestages, and Marital Status



To consider the link between the media groups and factors, the coefficients of the *kmeans* centroids on each of the factors was considered. These are reflected in tables 2.6 and 2.7 below.

Table 2.6: Media Group Centroid Coefficients on Factors for Cape Town

CAPE TOWN							
Media Group	**Factor1**	**Factor2**	**Factor3**	**Factor4**	**Factor5**	**Factor6**	**Factor7**
1	0.3011	-0.6077	-0.1455	-0.409	0.4388	0.0116	1.3899
2	-0.403	-0.746	0.7479	-0.692	-0.3988	-0.2508	-0.687
3	0.2208	-0.0212	2.5364	0.7676	-0.1448	-0.0848	-0.048
4	0.1027	-0.0441	-0.3262	-0.0031	2.1927	-0.2045	-0.4414
5	0.4202	1.197	-0.1493	-0.0174	-0.2783	-0.5822	0.0859
6	-0.3758	0.7715	-0.0564	0.1415	-0.1584	2.3043	-0.2182
7	0.5216	-0.7248	-0.4798	0.2401	-0.3693	-0.1496	-0.1816
8	-1.6369	0.3625	-0.4443	0.3131	-0.2415	-0.6191	0

Table 2.7: Media Group Centroid Coefficients on Factors for Johannesburg

JOHANNESBURG								
Media Group	**Factor1**	**Factor2**	**Factor3**	**Factor4**	**Factor5**	**Factor6**	**Factor7**	**Factor8**
1	-0.2223	0.554	-0.3366	-0.6293	-0.2349	-0.1483	-0.702	-0.0756
2	-0.1906	-0.1867	0.2863	0.4003	2.2094	-0.1838	0.0366	-0.0966
3	1.7438	0.0842	0.1409	0.1526	-0.2411	-0.4701	-0.0477	1.2369
4	-0.1825	-0.0717	0.0103	1.2487	-0.6918	-0.1976	0.1289	-0.1178
5	0.8923	0.0989	0.3513	0.0239	-0.1255	2.5374	-0.2321	-0.2613
6	-0.0665	0.3443	0.0611	-0.7195	-0.3113	-0.181	0.9061	-0.0452

In summary, a table that combines the salient information also indicating dominant media repertoires is shown in table 2.4 below.

Table 2.8: Media Repertoires by Media Group for Cape Town by Dominant Factors (loadings > 0.3)

MEDIA GROUP	SUMMARY PROFILE	MEDIA REPERTOIRES		
1	older, coloured, low income, low education, TV is big but also other media	**FACTOR 7 (1.4)**	**FACTOR 5 (0.4)**	**FACTOR 1 (0.3)**
		DAILY_Voice Son Heart_FM	HUISgenoot DIE_BURGER Rapport_Sun Son Radio_Tygerberg	SABC_3 e_tv SABC_2 SABC_1
2	younger, black, low income, low education, min media	**FACTOR 3 (0.7)**		
		Umhlobo_Wenene_FM DRUM METRO_FM DAILY_SUN KICKOFF Jet_Club SABC_1		
3	younger, black, low to mid income, max media	**FACTOR 3 (2.5)**	**FACTOR 4 (0.8)**	**FACTOR 1 (0.2)**
		Umhlobo_Wenene_FM DRUM METRO_FM DAILY_SUN KICKOFF Jet_Club SABC_1	Cape_Argus YOU Sunday_Times CAPE_TIMES DSTV Premium_Compact_dish FAIRLADY Weekend_Argus_Sat M_Net_Main	SABC_3 e_tv SABC_2 SABC_1
4	older, white, afrikaans, higer income, max media user	**FACTOR 5 (2.2)**	**FACTOR 1 (0.1)**	
		HUISgenoot DIE_BURGER Rapport_Sun Son Radio_Tygerberg	SABC_3 e_tv SABC_2 SABC_1	
5	younger, higher education, higher income, big on internet	**FACTOR 2 (1.2)**	**FACTOR 1 (0.4)**	**FACTOR 7 (0.1)**
		int_search int_social int_news X5FM	SABC_3 e_tv SABC_2 SABC_1	DAILY_Voice Son Heart_FM
6	younger, white, english, high income, big on internet	**FACTOR 6 (2.3)**	**FACTOR 2 (0.8)**	**FACTOR 4 (0.1)**
		int_print int_news	int_search int_social int_news X5FM	Cape_Argus YOU Sunday_Times CAPE_TIMES DSTV Premium_Compact_dish FAIRLADY

So this is where I could put a caption or what

MEDIA GROUP	SUMMARY PROFILE	MEDIA REPERTOIRES		
1	older, coloured, low income, low education, TV is big but also other media	**FACTOR 7 (1.4)**	**FACTOR 5 (0.4)**	**FACTOR 1 (0.3)**
		DAILY_Voice Son Heart_FM	HUISgenoot DIE_BURGER Rapport_Sun Son Radio_Tygerberg	SABC_3 e_tv SABC_2 SABC_1
2	younger, black, low income, low education, min media	**FACTOR 3 (0.7)**		
		Umhlobo_Wenene_FM DRUM METRO_FM DAILY_SUN KICKOFF Jet_Club SABC_1		
3	younger, black, low to mid income, max media	**FACTOR 3 (2.5)**	**FACTOR 4 (0.8)**	**FACTOR 1 (0.2)**
		Umhlobo_Wenene_FM DRUM METRO_FM DAILY_SUN KICKOFF Jet_Club SABC_1	Cape_Argus YOU Sunday_Times CAPE_TIMES DSTV Premium_Compact_dish FAIRLADY Weekend_Argus_Sat M_Net_Main	SABC_3 e_tv SABC_2 SABC_1
4	older, white, afrikaans, higer income, max media user	**FACTOR 5 (2.2)**	**FACTOR 1 (0.1)**	

		HUISgenoot DIE_BURGER Rapport_Sun Son Radio_Tygerberg	SABC_3 e_tv SABC_2 SABC_1	
5	younger, higher education, higher income, big on internet	**FACTOR 2 (1.2)**	**FACTOR 1 (0.4)**	**FACTOR 7 (0.1)**
		int_search int_social int_news X5FM	SABC_3 e_tv SABC_2 SABC_1	DAILY_Voice Son Heart_FM
6	younger, white, english, high income, big on internet	**FACTOR 6 (2.3)**	**FACTOR 2 (0.8)**	**FACTOR 4 (0.1)**
		int_print int_news	int_search int_social int_news X5FM	Cape_Argus YOU Sunday_Times CAPE_TIMES DSTV Premium_Compact_dish FAIRLADY Weekend_Argus_Sat M_Net_Main
7	older, low education, mid to low income, coloured/indian, TV dominates	**FACTOR 1 (0.5),**	**FACTOR 4 (0.2)**	
		SABC_3 e_tv SABC_2 SABC_1	Cape_Argus YOU Sunday_Times CAPE_TIMES DSTV Premium_Compact_dish FAIRLADY Weekend_Argus_Sat M_Net_Main	
8	white/indian, higher education and high income, min media user	**FACTOR 2 (0.4)**	**FACTOR 4 (0.3)**	
		int_search int_social int_news X5FM	Cape_Argus YOU Sunday_Times CAPE_TIMES	

			DSTV Premium_Compact_dish FAIRLADY Weekend_Argus_Sat M_Net_Main	

2.3.3.2 Predictions and Canonical Correlations

In this sections the value of the categories will be considered by applying various predictions and considering canonical correlations

comments: next want to do some predictive runs; then focus on Jhb; then some comment on cohorts and how I would expect to apply to LDA

Chapter 3

uction

The media environment, a fundamental area of interest to marketers, has undergone dramatic changes that in already changing the nature of the discipline. The contextual framework in which both the intended MSc in Data Science and a future PhD will be situated is the media environment in South Africa.

The primary focus of a master's half-dissertation would be on establishing the principal media repertoires of South Africans as they pertained in 2016. The work on the masters half-thesis will link up with a proposed PhD-thesis by providing a basis and a foundation to consider both a longitudinal approach to media repertoires as well as multi-method research approaches to gaining a deeper understanding of how the repertoire groupings consume (or use) different media types.

The main benefit to this work and its contribution to the body of knowledge would be a comprehensive mapping of the new cross-media environment evolving in South Africa and the establishment of baseline research that can be utilised to monitor and track the media environment in the country. The research will also serve to contribute to the global question of how to make sense of the new media world and therefore how to respond as marketers.

In the context of the generally speculative nature of this document (see the Preamble above) particular emphasis has been placed in chapter 4 on describing the broad contextual overview of the important and dramatically changing media environment in which both the MSc half-thesis and the yet-to-be confirmed focus of a PhD thesis would be situated. This is followed by very basic, and by necessity preliminary, outlines for the proposed Masters and PhD studies in chapters 5 and 6 respectively. These outlines need to be read in conjunction with the contextual background chapter 4.

Due to the fact that the MSc thesis at least would be in the field of Data Science, more emphasis has also been placed on illustrating this focus on the data analytic methodologies by including preliminary work on factor analytic methods that would be included in the masters half thesis in section 5.4.

Chapter 4

Broad Contextual Overview

This section serves to provide a broad contextual overview that describes the researcher's interest in order to situate the proposed research for both the MSc half-thesis and the yet-to-be-articulated PhD. In section 4.1 it first considers the importance of media, in particular its role in the developing world, which sparked this researchers interest in the field; and also lead to an awareness of the importance of data-based analyses in the field.

In particular section 4.2 reviews the dramatic shifts impacting on the current media environment and section 4.3 considers some approaches to research in the current media context, which has become a very complex and an increasingly data-driven field.

4.1 The Importance of Media

A sound, functioning, strong media sector is important for societies and countries to build an open society. Nobel laureate and economist Amartya Sen observes that famine has never occurred in a democracy with a free press: "intimations of mass starvation are impossible to hide where journalists freely give voice to public criticism and warn of impending crises" (*Toward Economic Sustainability of the Media in Developing Countries* (Anon, 2007)).

Susman-Pena (2012), citing research over several years that show the coexistence of a sound media sector on the one hand and good developmental outcomes on the other, posits that media *matters*. She defines a *healthy* media sector as one that is "free, independent, produces high quality information, reaches all or most of the population, offers diverse perspectives, and provides the information people need to be able to make decisions and to be able to hold their government to account" (Susman-Pena, 2012).

Peters (2010, citing Jakubowicz and Sukosd, 2008), argues that media are an especially important focus of attention, "given the role they are often assumed to have in creating national identity and contributing to an energized democratic society". According to Conrad (2014), the notion that an independent media is the foundation of a functional democracy has been argued by liberal theorists, including John Locke, John Madison, John Milton, and

John Stuart Mill, for decades. Democratic government relies on the ability of its citizens to make informed decisions and this requires access to information that is accurate, more often than not implying a "free and independent media" (*Toward Economic Sustainability of the Media in Developing Countries* (Anon, 2007)).

An effective media sector is also important to hold public officials accountable and in so doing help in the exposing of corruption (*Toward Economic Sustainability of the Media in Developing Countries* (Anon, 2007)). Susman-Pena (2012) lists *governance institutions* as an important outcome of a healthy media sector. More specifically, media oversight that provides information about government activities, decision-making and budgeting can ensure government accountability and "expose vice or incompetence". (Peters, 2010, citing Paul Starr, 2009; Adsera, Boix and Payne, 2003), argues that "corruption is more likely to flourish when those in power have less reason to fear exposure" and that a strong negative correlation exists between corruption and free circulation of newspapers in a country.

A healthy media environment also assists in promoting economic growth by disseminating information and ensuring transparency (*Toward Economic Sustainability of the Media in Developing Countries* (Anon, 2007)). Susman-Pena (2012) argues that the support and promotion of a healthy media sector is critical to "grow economies, alleviate poverty, and improve lives". A healthy media sector creates an information flow that is the lifeblood of a working and efficient society. In fact an environment that fails to support a free press, such as a regulatory environment that does not support the freedom of journalists to do their work and one that discourages independence and plurality, will also fail in achieving these outcomes (*Toward Economic Sustainability of the Media in Developing Countries* (Anon, 2007)). And "by increasing people's knowledge about their own and other societies, the media may strengthen bonds and common understandings among people" (Susman-Pena, 2012).

In what Susman-Pena (2012) describes as "a remarkable commitment to supporting independent media, together with access to information, and freedom of expression" President Obama, in his May 2011 speech on the Middle East, declared:

> "Real reform does not come at the ballot box alone. Through our efforts we must support those basic rights to speak your mind and access information. We will support open access to the Internet, and the right of journalists to be heard–whether it's a big news organization or a lone blogger. In the 21st century, information is power, the truth cannot be hidden, and the legitimacy of governments will ultimately depend on active and informed citizens."

These arguments about the importance of media are in fact arguments for a particular kind of media and by extension for a particular kind of journalism. Anderson et al. (2012, p. 3) describe this by holding that not media as much as *journalism* matters: "Journalism exposes corruption, draws attention to injustice, holds politicians and businesses accountable for their promises and duties. It informs citizens and consumers, helps organize public opinion, explains complex issues and clarifies essential disagreements. Journalism plays an irreplaceable role in both democratic politics and market economies."

Anderson et al. (2012, p. 3) qualify this argument by stating that not all journalism matters, since "much of what is produced today is simply entertainment or diversion" . What matters has variously been referred to as *hard news, accountability journalism*, or *the iron core of news.* Rather than trying to list or define the elements that separate hard news from what they refer to as "fluff", they adopt Lord Northcliffe's litmus test:

> "News is something someone somewhere doesn't want printed. Everything else is advertising" (Anderson et al., 2012, p. 3).

And state that:

> "Hard news is what distinguishes journalism from just another commercial activity. There will always be a public appetite for reporting on baseball, movie stars, gardening and cooking, but it's of no great moment for the country if all of that work were taken over by amateurs or done by machine. What is of great moment is reporting on important and true stories that can change society. " (Anderson et al., 2012, p. 3).

4.2 The Current Media Environment

The digitization of media content and the expansion of the Internet has led to a large increase in the media sources available to consumers. In this proliferation of media, one of the main research questions has been how audiences respond to this explosion of choice. For example, does it lead to people consuming a steady and consistent diet of their preferred news genre or do they expand their consumption to a wider, more diverse range of sources (Xu et al., 2014, citing Gentzkow and Shapiro, 2011; Ksiazek, Malthouse and Webster, 2010)?

According to Xu et al. (2014, p. 100, citing Hollander, 2008; Iyengar and Hahn, 2009; Kziazek, Malthouse and Webster, 2010) a key finding on this question by various researchers is that consumers have responded to the increase in media sources using a strategy of "selective exposure" in which they consume more of similar news from a small number of news providers rather than consuming a greater variety from a larger number of news providers.

According to Schröder (2015, pp. 60-61), previous patterns and routines of news consumption, including trust in the news, are being transformed by a "nexus of innovative technologies and the revolutionizing journalistic processes they afford. This nexus results in news media content which sets societal agendas and frames cultural issues in new ways for news audiences, irrespective of whether they get their news directly from social media platforms or not."

Schröder (2015, p. 61, citing Bjur et al., 2013) argues that audiences are inherently cross-media and in the digital age, “emerging patterns of cross-media use are far more seamless and blurred, hybrid and complex than they used to be”.

Edgerly (2015, p. 1) states that in a world where we have moved beyond a limited number of television channels, radio stations and print news outlet to an environment in which we make selection choices “amid hundreds of television channels, smart phone technologies, and virtually unlimited news options available online”, the most defining characteristic of the current media environment is *media choice.*

While the “low-choice” media environment was characterised by a “consistent approach to news presentation and style”, the “high-choice” media environment is characterised by *diversity,* reflected by ideologically driven news and the blurring of news and entertainment (Edgerly, 2015, p. 2). Furthermore, Edgerly (2015, pp. 2-3, citing Baym, 2010; Shoemaker and Reese, 1996) identifies the “interplay of ownership influence, pressure to fill a large news hold, and increased competition” as resulting in “news content that shies away from traditional notions of neutrality and objectivity”.

Edgerly (2015, pp. 3-4) also identifies soft news, daytime and late-night talk shows, and news satire as examples of a rise of “hybrid media” that “blur the line between news and entertainment” and a expresses concern that “individuals are turning away from news altogether, or the news they do select is too entertainment oriented”. Edgerly (2015, p. 4, citing Postman, 1986) cautions that the popularity of hybrid media combined with declines in traditional forms of news may indicate that we are “amusing ourselves to death”.

The drop in news exposure among younger people is particularly evident. In this regard the Pew Research Center for the People and the Press (2010) found that only 23% of 18 to 29 year-olds regularly read a newspaper, compared to 55% of people over 65 - a pattern also evident in audiences of network evening news. The declining levels of news exposure among younger people may not be an indication that they are “tuned out” or “fleeing” from news, but that they are consuming “a different set of news content”, for example from various online and social media sources and from entertainment news-driven television shows, such as the *Daily Show.* The concern is that the younger generations are replacing traditional forms of news with lower quality ones and that the declines in political knowledge and participation among this cohort reflects a change in news diets. “The overarching concern is not that these new types of news are inherently bad, but that the exclusive use of *only* ideologically driven news, or *only* media that mix news and entertainment is the cause for worry” (Edgerly, 2015, p. 4, citing Mindich, 2005; Patterson, 2008).

The explosion of media choice has also resulted in audiences drifting away from mainstream media (Turcotte et al., 2015, p. 520, citing Prior, 2007; Stroud, 2011), exacerbated by “a steady decline of public trust in the institution of news” (Turcotte et al., 2015, pp. 520-521, citing Gronke and Cook, 2007; Pew Research, 2012). Turcotte et al. (2015, p. 521, citing Gronke and Cook, 2007; Ladd, 2011) hold that “scholars have observed aggregate level declines of public trust in the news over the last few decades, transforming a once revered news profession to a subject of disdain and dissatisfaction". However, in spite of this aggregate decline of trust, not all news outlets are similarly impacted. For example, according to Turcotte et al. (2015, p. 521, citing Arceneaux, Johnson and Murphy, 2012; Gronke and Cook, 2007), "the public is more trusting of their preferred outlets for news and more trusting of local news outlets" and a 2014 survey released by Public Policy Polling found that Fox News was considered both the most and least trusted news source, suggesting that news outlet credibility varies according to partisan predispositions. The current era of expanded media choice, has led to people choosing news sources that are in agreement with their ideology - a selective exposure that may have made people too trusting of their preferred outlets while distrusting news sources that don't agree with their ideological leanings (Turcotte et al., 2015, pp. 521, citing Arceneaux et al., 2012). While demographics and political knowledge can also play a role in determining which news outlets one perceives as credible (Turcotte et al., 2015, pp. 522, citing Stroud and Lee, 2013), media trust has been conceptualised in several different ways, including trust in content, trust in journalists or those responsible for delivering the news, and trust in media ownership Turcotte et al. (2015, pp. 522, citing Williams, 2012). Trust in news can also directly influence political behaviour. As media distrust grows, the voting public becomes more dependent on partisan cues to determine their vote; and it is more likely to abandon mainstream news. Furthermore the dropping levels of media trust "fosters a heightened perception that the current political climate is a polarizing one" Turcotte et al. (2015, pp. 522, citing Ladd, 2011; Ladd, 2013).

Edgerly (2015, p. 1, citing Mindich, 2005; Prior, 2007; and others) argues that the expansion of media options and therefore of media choice has led to a concern that it makes it easy to avoid news content altogether and that the "fragmented-nature" of news exposure can lead to the gravitation of audiences toward "sources that reinforce their ideological viewpoints and are of lower quality". She identifies a concern that ideological news "makes it easier for people to consume only content that agrees with their political views" Edgerly (2015, p. 3) and in reference to the red and blue divide signifying the Republican and Democratic political poles in the United States, she argues that "despite all the colourful options the news media landscape offers, some audiences only see 'red media' or 'blue media"' Edgerly (2015, p. 3).

(Xu et al., 2014, p. 100, citing Ksiazek, Malthouse, and Webster, 2010; Stroud, 2008) also argues that due to the "increasing proliferation of news outlets, consumers with a particular political preference will be more likely to consume from news outlets that match their own value beliefs. This behavior results in a penchant for ideological segregation."

In a study on cross-media usage, Schröder (2015, p. 71) asks if the established mixtures and levels of news media qualify a given audience as "well-informed, resourceful and competent citizens". Starting with a question of what the media repertoire of a *competent* citizen would look like, he holds that traditionally daily newspaper readership would have been considered a *"sine qua non* of informed citizenship" and that the decline in newspaper readership would therefore indicate "democratic deficit... that could pose a serious challenge to the democratic health of a country". However this view is challenged by the changing face and growing consumption of online news services, many of which emanate from former print offerings, and the growing role of mobile platforms and social media that may indicate high levels of *digital literacies*, which in turn are increasingly being considered a "prerequisite of democratic citizenship, as well as civic agency" Schröder (2015, p. 71, citing Lund et al., 2014; Curren et al., 2009).

In such a complex media environment while most news mediums struggle for audiences some social media sites are showing potential for growth as carriers of news. Facebook for example has become one of the fastest growing tools for gathering news with more than half its users consuming news on the site and 78% of its users reporting exposure to news while using Facebook for social and other reasons (Turcotte et al., 2015, pp. 521, citing Pew Research, 2014). Also in a media environment with so much choice, "one extremely important way [individuals] decide what to pay attention to is through recommendations that reach them through their online social networks" Turcotte et al. (2015, pp. 523, citing Mutz and Young, 2011), suggesting that opinion-leaders play increasingly important roles in facilitating exposure to news; and 'friends' can play the role, traditionally reserved for journalists and editors, as information gate keepers, vetting the significance and relevance of news content and therefore help shape public agendas (Turcotte et al., 2015, pp. 524).

Xu et al. (2014, p. 98, citing comScore, 2012) argues that it is critical for advertisers to understand any changes in news consumption behaviour since, after email and texting, accessing news in the 2012 study was found to be the most popular mobile data activity in the U.S. The digitization of news has also fundamentally reshaped the news industry. For example in the U.S. newspaper advertising revenues fell 47% from 2005 to 2009 as online advertising spending climbed to more than $100-billion in 2012 (Xu et al., 2014, p. 97).

The proliferation of new media outlets, resulting in growing numbers of people meeting their news needs through multiple outlets, has raised the problem for marketers of how best to target and reach consumers in such a multichannel environment. While online tracking technologies have made it possible to track online consumers, marketers still need to consider the placement of advertising on multiple outlets in order to reach consumers effectively. The emergence of these "disruptive channels" has made it imperative for marketers to monitor changes in consumers' media and news consumption behaviours (Xu et al., 2014, p. 97, citing Ahonen, 2011; Athey, Calvano and Gans, 2013).

> "If you wanted to sum up the past decade of the news ecosystem in a single phrase, it might be this: Everybody suddenly got a lot more freedom. The newsmakers, the advertisers, the startups, and, especially, the people formerly known as the audience have all been given new freedom to communicate, narrowly and broadly, outside the old strictures of the broadcast and publishing models." (Anderson et al., 2012, p. 1)

4.3 Research of the Media Environment

Schröder (2015, pp. 60-61) argues that it is important to monitor “on a continuous basis precisely what the landscape of news looks like: what technological platforms and formats are receding and emerging, and which are dominant, as well as how people are accessing, navigating in, and making sense of the landscape of news.”

Various attempts to “develop and operationalise” new conceptual frameworks for “mapping and explaining the cross-media practices of audiences”, suggests that cross-media consumption can be researched from three perspectives: firstly in terms of *functional differentiation* that considers the extent to which different media complement and co-exist with each other, for example when one medium specialises in fulfilling certain types of needs in order to differentiate itself from its rivals; secondly, research that is concerned with the building of *media repertoires*, i.e. how audiences create “personal constellations of media”, reflecting a variety of media technologies, media genres and media brands or products, which jointly fulfill their everyday needs for information, diversion and sociability; and, thirdly research that considers *location ensembles* that adopts a “situational perspective” to study how media belong to or transcend specific socio-spatial contexts (Schröder, 2015, p. 62, citing Bjur, 2013).

Another challenge in audience research is situating media research in the debate about “mediatization“. Schröder (2015, p.62, citing Hepp, 2013) holds that we have been “stepping into the era of mediatization, in which the role of the media across the range of social institutions and everyday life has grown in quantitative as well and quantitative terms”. He identifies a plurality of “cultures of mediatization”, arguing that “the processes and ’logics’ of mediatization should not just be explored at the level of social institutions, but also in the everyday processes through which people encounter, acquire and make sense of the media in their dual appearance as technologies and multimodal discourses”. For example one such a ’logic’ is *audience logic*, which can be operationalised using a notion of *worthwhileness*, described using seven factors or dimensions that determine why some news media and not others are chosen to become parts of an individual’s news media repertoire (Schröder, 2015, p. 63).

The challenges of studying where and how, “the places and spaces”, news media are being consumed has also become much more difficult. In the 1980’s it was possible to sent observers into homes in order the map the uses - including concomitant social interactions - of various traditional media types in ethnographic or grounded theory-based research. However, today’s multi-platform media world requires alternative research methods. For example various *multi-method* approaches have been attempted in which media use is observed using techniques involving a triangulating mixture of questionnaire-based surveys, focus groups and ethnological observation of participants’ online practices, including methods described as *virtual shadowing* of users activities (Turcotte et al., 2015, pp. 524, citing Vittadini and Pasquali, 2014; Jensen and Sorensen, 2014).

Identifying *media repertoires*, as first outlined in the second research perspective described by Schröder (2015, p. 62, citing Bjur, 2013) above, and also mapping trends and changes in such repertoires may serve as a valuable starting point to gain a deeper understanding of the media environment on which to build further explorations of media usage that could serve marketers and other media analysts. In fact, Edgerly (2015, p. 4) encourages an approach grounded in work on media repertoires, arguing that "we can learn a lot about audiences by examining what combinations of media they choose over others".

According to Edgerly (2015, p. 4), such a *repertoire* approach to media exposure was first developed by Heeter in 1985 to describe channel-watching routines of television users. She identifies two main lines of research pertaining to repertoires: one focussing on repertoires within a single medium; another considering repertoires across media. Furthermore, to account for the ability to sample from many different types of news in repertoire research, news exposure is conceptualised as a "complex pattern of news use rather than a single media selection."

Edgerly (2015, pp. 1-2, citing Hasebrink and Popp, 2006, and others) describes a news repertoire approach to research in a cross-media environment as "identifying distinct ways that media users combine news across a wide array of media platforms and content". She holds that this approach is "less about exposure to a single news source, and more about the subset of news sources that people consume in tandem." And, "as such a repertoire approach provides a window into the decision-making strategies of audiences who are faced with increased media options". Using a national survey of the media usage of U.S. adults she identifies the existence of six distinct news repertoires, finding that while some are clearly ideologically based, spanning multiple media platforms, others have repertoires that function at a media level and others still show respondents who consume both politically conservative and liberal news. Furthermore the six repertoires show distinct audience groups in terms of media engagement and participation as well as socio-demographic profilesEdgerly (2015, p. 2).

Schröder (2015, pp. 70-71), using both quantitative and qualitative methodologies on a relatively small sample in Denmark, identified seven types of news consumers, who all used a mixture of traditional and new sources of news.

Chapter 5

Outline of a Masters half-thesis in Data Science

5.1 Context

The specific context in which this research will be situated is in the South Africa media environment, which has not been immune to the factors that have impacted on the media environment as outlined in chapter 4 above.

Of particular importance to this thesis is the fact that in South Africa a national survey of media (and product) usage, the All Media and Products Survey (AMPS), conducted under the auspices of the South African Audience Research Foundation (SAARF), has been conducted at least once a year from 1974 until 2016. The surveys, which are based on stratified random sampling to reflect the demography of South Africa, involved around 20 000 participants in each of two surveys a year.

No information about media repertoire research as it pertains to South Africa could be found.

5.2 Aim of the Research

Utilising data from AMPS determine the dominant media repertoires in South Africa for 2016 and link these repertoires with demographic and psychographic profiles of respondents.

5.3 Methods

The analytic methods for this thesis will be primarily the use of various factor analytic techniques, explored more fully in section 5.4 below, applied to AMPS 2016 data. The factor analytic methods will aim to identify dominant latent factors in media usage that can be considered groups with different media repertoires. The intended outcome is a description of the principal media repertoires that exist in the South African market and to identify the demographic profiles of the groups that represent these repertoires.

5.4 Preliminary Review of Data Analytic Methods

5.4.1 Measurement

Researchers are often interested in variables that cannot be directly observed, such as intelligence or attitudes and perceptions to a product or service. These unobservable variables are described as *latent* variables, *factors* or *constructs* and researchers try to get information about them through observable variables, such as the response to a scaled question, these variables are also referred to as *measured*, *indicator*, or *manifest* variables. Factor analyses and Structural Equation Modeling are techniques designed to reduce the number of observable variables into a smaller number of latent variables by considering the covariation of the observed variables (Schreiber et al., 2006, p. 323).

Research aimed at measuring opinions, attitudes and perceptions typically make use of statements and ask participants to express their opinions or attitudes to the statements on a number of categories using verbal descriptions of scales that are arranged in increasing levels of intensity. One such technique, and the one used in the dataset used for this project, is called a *Likert* scale, which makes use of descriptions for each scale point and can range from three to seven points (Berndt & Petzer, 2011, p. 189).

The use of parametric statistical techniques using Likert scale data is controversial. The debate can be distilled as representing two competing views: on the one hand that Likert scales represent *ordinal* level data and hence they must be analysed using non-parametric statistics; and on the other, that Likert *scales* as opposed to individual Likert items or *interval* in character and can therefore be analysed with the more powerful statistical techniques that apply to parametric data (Carifio & Perla, 2008, p. 1150).

Carifio & Perla (2008, p. 1150) maintain that "a great deal of empirical evidence" should resolve this debate. They refer to Monte Carlo studies of the F-test that were performed by Glass *et al.* (1972) that showed the F-test to be "extremely robust" to violations of its assumptions as evidence that the F-test applied to ordinal data produces unbiased results. They also refer to various studies (Carifio, 1976; 1978) on the nature of Likert *scales* (particularly if they comprise at least eight items) as opposed to individual Likert items that have shown the "Likert response format produces empirically interval data and, in fact, can approximate ratio data, in theory and actuality".

5.4.2 Principal Components Analysis

5.4.2.1 Introduction

Kline (1994, p. 28) states that factor analysis without understanding is an "unmitigated evil" and argues that once the calculation of principal components has been understood, the nature of factor analysis becomes self-evident. Although his reference was the context of the social sciences, it could also apply to the management sciences. For this reason Principal Components Analysis (PCA) will be considered in some detail.

Section 5.4.2.2 will begin with a description and the aims of PCA before outlining a more detailed mathematical derivation of the process of identifying principal components (section 5.4.2.3). Section 5.4.2.3 describes some arguments for selecting and interpreting the principal components and their loadings.

5.4.2.2 Description and Purpose

PCA is used to identify underlying dimensions of multivariate data by describing a set of new variables, that are fewer than the original set, but yet explain most of the variance in the original sample (Grimm & Yarnold, 2004, pp. 99-100). If the original variables are nearly uncorrelated, then it would make no sense to consider PCA. According to Radloff (2015), PCA entails the finding of an orthogonal transformation of an original set of correlated variables to a new, reduced set of variables that is entirely uncorrelated. This new set of variables are called the "principal components".

PCA results in a list of components that are ordered by the proportion of the variance that they account for and that it's advantage lies in the fact that "no other method of extracting factors could yield factors, at any stage of extraction, which explained more variance than components" (Kline, 1994, p. 38).

Each principal component is a linear combination of the original variables and a measure of the amount of information conveyed by each of the principal components is contained in its variance. The principal components are arranged, by construction, into descending order such that the first component represents most of the explained variance and the last, the least. The first principal component can therefore be considered the most informative, and the last the least informative (Afifi et al., 2011, p. 357).

Grimm & Yarnold (2004, p.101) use a geometric description to underline the key role of the first principal component:

> "...think of the principal component as a line that passes through a swarm of data points ... plotted in a multidimensional space ... (where) the number of dimensions of this space is equal to the number of variables. In that space, each observation lies at some distance – measured perpendicularly – from the line reflecting the first principal component ... The first principal component (line)

> is associated with the minimum sum of squared error terms for the sample of observations . . . Formally, the linear function, or principal component, is referred to as an *eigenvector ...* in this case, the first *eigenvector* . Also, the amount of the total variance that is explained by an *eigenvector* is known as the *eigenvalue*" (and the first *eigenvector* is associated with the largest *eigenvalue*).

A second principal component (or second eigenvector) is also a linear function of the original variables, but maximises the amount of remaining explained variance. It is calculated in the same way as the first. Since the two principal components are perpendicular (orthogonal), none of the variance explained by the second could have been explained by the first (Grimm & Yarnold, 2004, pp. 101-102).

Although the focus of this project is on dimension reduction, a more comprehensive list of reasons for using PCA is outlined in Afifi et al. (2011, pp. 357-358), namely:

a) To reduce the dimensions of a problem without losing much of the information contained in the data. Here only the first few components would be selected for further analysis. This technique is attractive since the components are not intercorrelated and therefore allows for simpler analyses than working with complex interrelationships;

b) the principal components can also be used as an effective test for underlying normality. If the principal components demonstrate a normal distribution, the originating variables can be assumed to have a normal distribution;

c) another use of PCA is to search for outliers: a histogram of each of the principal components can effectively identify specific subjects with measurements that are very large or very small;

d) as a means of overcoming multi-collinearity, PCA can be considered a "step toward factor analysis"; and

e) PCA can be used as an exploratory technique. Most of the examples of factor analyses in Kline (1994), use PCA as an initial exploration of the number of factors to consider in subsequent factor analyses.

A key assumption of PCA is that the total variance of any given variable reflects the sum of explained and error variance (Grimm & Yarnold, 2004); and although this statistical technique does not require a multivariate normal assumption of the original variables, inferences can be drawn from the sample principal components when the population is multivariate normally distributed (Radloff, 2015).

5.4.2.3 Deriving the Principal Components

Basic Terminology and Principles This section draws on Afifi et al. (2011, pp. 359-363). Consider the case of $n$observations of just two variables, X_1, X_2.

PCA leads to the creation of two entirely new variables C_1 and C_2 , called the *principal components*, that can be written as linear combinations of X_1 and X_2, i.e.

$$\begin{aligned} C_1 &= b_{11}X_1 + b_{12}X_2 \\ C_2 &= b_{21}X_1 + b_{22}X_2 \end{aligned}$$

The coefficients, b_{ij}, are determined on the basis of three requirements or restrictions:

1. The variance of C_1 is as large as possible;
2. The values of C_1 and C_2 are uncorrelated;
3. For each component, the sum of the squares of the coefficients must equal unity in order to attain orthogonal eigenvectors, i.e. $b_{11}^2 + b_{12}^2 = b_{21}^2 + b_{22}^2 = 1$

Afifi et al. (2011, p.361) demonstrate PCA graphically, showing that it amounts to rotating the original X_1 and X_2 axes to new C_1 and C_2 axes.

As shown in section 5.4.2.3, variances of C_1 and C_2 are the eigenvalues (alternatively, characteristic roots, latent roots, or proper values) and the sum of all the eigenvalues are equal to the sum of the variances from the original dataset. This relationship is retained in general, that is the total variance is preserved when rotating the original dataset to its principal components.

The set of coefficients of the linear combinations, b_{ij} are described as the orthonormal eigenvectors of the variance-covariance matrix.

Deriving Principal Components

Introduction According to Afifi et al. (2011), the mathematics of determining the principal component coefficients b_{ij} (the eigenvectors) and their eigenvalues was first derived by Hotelling (1933) and is outlined in work published by Kline (1994) in which he demonstrates an iterative process in which a trial vector is tested against a set of criterion values. The extent of the divergence is used to modify the first trial vector to produce a second. This process is continued until further iterations produce the same results (Kline, 1994, p.30).

The following section provides a more general derivation of the characteristic equations that produce the required eigenvalues and eigenvectors.

General Derivation This section draws primarily from Radloff (2015).

Consider a matrix of independent random variables $\mathbf{X} = [X_1, X_2, \cdots, X_p]'$ of length p with an expected value $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, which is positive-semi definite.

The aim of PCA is to identify a new matrix of variables, $\mathbf{C} = [C_1, C_2 \cdots C_p]'$ whose columns are uncorrelated, each of which is a linear combination of the original components of $\mathbf{X}$ and whose variances are arranged in decreasing order, thus

$$C_j = b_{1j}X_1 + b_{2j}X_2 + \ldots + b_{pj}X_p$$

or, in matrix notation, $C_j = \mathbf{b}_j'\mathbf{X}$, where $\mathbf{b}_j = [b_{1j}, ..., b_{pj}]'$ is a vector of "constraints", defined in such a way that $\mathbf{b}_j'\mathbf{b}_j = 1$ and $\mathbf{b}_j'\mathbf{b}_i = 0$ in order to achieve the desired orthogonal transformation.

To find the first component we want to choose $\mathbf{b}_1$ in such a way as to maximise the variance of C_1, but subject to the normalising constraint $\mathbf{b}_j'\mathbf{b}_j = 1$ and that the first $r \leq p$ eigenvalues of $Var(\mathbf{X})$ =$\boldsymbol{\Sigma}$ are distinct.

That is, we want to maximise $Var(C_1) = Var\left(\mathbf{b}_1'\mathbf{X}\right) = \mathbf{b}_1'\boldsymbol{\Sigma}\mathbf{b}_1$.

Using the Lagrangian expression with λ as the Lagrange multiplier to maximise the $Var\left(C_1\right)$, we consider the expression:

$$H_1 = \mathbf{b}_1'\boldsymbol{\Sigma}\mathbf{b}_1 - \lambda(\mathbf{b}_1'\mathbf{b}_1 - 1)$$

Differentiating and equating to zero, gives

$$\frac{\partial H_1}{\partial \mathbf{b}_1} = 2\boldsymbol{\Sigma}\mathbf{b}_1 - 2\lambda\mathbf{b}_1$$

and hence

$$\left(\boldsymbol{\Sigma} - \lambda\mathbf{I}\right)\mathbf{b}_1 = 0 \tag{5.1}$$

This equation 5.1 gives the important relationship, or *characteristic equations*, $\boldsymbol{\Sigma}\mathbf{b}_1 = \lambda\mathbf{b}_1$, showing that $\mathbf{b}_1$ is an eigenvector of $\boldsymbol{\Sigma}$, corresponding to the eigenvalue λ.

Let the r distinct eigenvalues of $\boldsymbol{\Sigma}$ be denoted by $\lambda_{1,} > \lambda_2 > \ldots > \lambda_r \geq 0$, arranged in descending order and consider $Var\left(\mathbf{b}_1'\mathbf{X}\right) = \mathbf{b}_1'(\boldsymbol{\Sigma}\mathbf{b}_1)$= $\mathbf{b}_1'\lambda\mathbf{b}_1$= λ_1. Since the aim is to maximise the variance of C_1, the first principal component, it is equated with the largest eigenvalue, namely λ_1.

It follows that the first principal component, $\mathbf{b}_1$, would be the orthogonal eigenvector of $\boldsymbol{\Sigma}$ that corresponds to the eigenvalue λ_1.

To determine the second principal component, namely $C_2 = \mathbf{b}_2'\mathbf{X}$, the argument is extended as follows. In addition to setting a constraint of $\mathbf{b}_2'\mathbf{b}_2 = 1$, a second constraint is imposed, namely that C_1 and C_2 should also be orthogonal, i.e. uncorrelated. To achieve this we require the covariance between the first two components to be equal to zero, thus

$$\begin{aligned} Cov\left(C_2, C_1\right) &= Cov\left(\mathbf{b}_2'\mathbf{X}, \mathbf{b}_1'\mathbf{X}\right) \\ &= E\left(\mathbf{b}_2'\left(\mathbf{X}-\boldsymbol{\mu}\right)\left(\mathbf{X}-\boldsymbol{\mu}\right)'\mathbf{b}_1\right) \\ &= \mathbf{b}_2'\boldsymbol{\Sigma}\mathbf{b}_1 \\ &= \mathbf{b}_2'\lambda_1\mathbf{b}_1 = \lambda_1\mathbf{b}_2'\mathbf{b}_1 = 0 \end{aligned}$$

To maximise the variance of $C_2 = Var(\mathbf{b}_2'\mathbf{X}) = \mathbf{b}_2'\boldsymbol{\Sigma}\mathbf{b}_2$, subject to these two constraints two Lagrange multipliers denoted λ and δ are introduced.

The resulting Lagrangian expression is

$$H_2 = \mathbf{b}_2'\boldsymbol{\Sigma}\mathbf{b}_2 - \lambda(\mathbf{b}_2'\mathbf{b}_2 - 1) - \delta\mathbf{b}_2'\mathbf{b}_1$$

Differentiating and equating to zero, gives

$$\frac{\partial H_1}{\partial \mathbf{b}_2} = 2\boldsymbol{\Sigma}\mathbf{b}_2 - 2\lambda\mathbf{b}_2 - \delta\mathbf{b}_1 = 0 \tag{5.2}$$

And from the *characteristic equation*, equation 5.1 above, i.e., $\left(\boldsymbol{\Sigma} - \lambda\mathbf{I}\right)\mathbf{b}_1 = 0$, it follows that $\boldsymbol{\Sigma}\mathbf{b_1}$ =$\lambda\mathbf{b}_1$ or equally $\boldsymbol{\Sigma}\mathbf{b_2}$ =$\lambda\mathbf{b}_2$.

Thus, equation 5.2 becomes

$$2\lambda\mathbf{b}_2 - 2\lambda\mathbf{b}_2 - \delta\mathbf{b}_1 = 0$$

implying that $\delta = 0$.

Pre-multiplying equation 5.2 by $\mathbf{b}_1'$ gives

$$2\mathbf{b}_1'\boldsymbol{\Sigma}\mathbf{b}_2 - 2\lambda\mathbf{b}_1'\mathbf{b}_2 = 0$$

Resulting in the *characteristic equation*s

$$\boldsymbol{\Sigma}\mathbf{b}_2 - \lambda\mathbf{b}_2 = 0$$

$$(\boldsymbol{\Sigma} - \lambda\mathbf{I})\mathbf{b}_2 = 0$$

We choose λ to be the second largest eigenvalue, λ_2, in order to satisfy the requirement of descending variances of the principal components.

For the r^{th} principal component we seek the vector $\mathbf{b}_r$ such that $\mathbf{b}_r\mathbf{X}$ will have a maximum variance subject to the condition that $\mathbf{b}_r'\mathbf{b}_r = 1$ and $\mathbf{b}_r'\mathbf{X}$ needs to be uncorrelated with the previously determined uncorrelated principal components $\mathbf{b}_j'\mathbf{X}$ for $j = 1, 2, \ldots, r-1$.

The quantity to be maximised would be $Var(C_r) = \mathbf{b}_r'\boldsymbol{\Sigma}\mathbf{b}$, subject to the constraints $\mathbf{b}_r'\mathbf{b}_r = 1$ and $\mathbf{b}_r'\mathbf{b}_j = 0$ for $j = 1, 2, ..., r-1$.

The relevant Lagrangian expression is

$$H_r = \mathbf{b}_r' \mathbf{\Sigma} \mathbf{b}_r - \lambda(\mathbf{b}_r' \mathbf{b}_r - 1) - \sum_{j=1}^{r-1} \delta_j \mathbf{b}_r' \mathbf{b}_j$$

where $\lambda, \delta_1, \delta_2, ..., \delta_{r-1}$ are the Lagrange multipliers.

Extending the argument used above in deriving the first and second component, differentiating and equating to zero gives,

$$\frac{\partial H_r}{\partial \mathbf{b}_r} = 2\mathbf{\Sigma} \mathbf{b}_r - 2\lambda \mathbf{b}_r - \sum_{j=1}^{r-1} \delta_j \mathbf{b}_j = 0$$

Pre-multiplying this expression by $\mathbf{b}_k'$ with $k < r$, yields

$$2\mathbf{b}_k' \mathbf{\Sigma} \mathbf{b}_r - 2\lambda \mathbf{b}_k' \mathbf{b}_r - \sum_{j=1}^{r-1} \delta_j \mathbf{b}_k' \mathbf{b}_j = 0$$

As demonstrated above, $\delta_k = 0$ for $k = 1, 2, \ldots, r-1$.

Considering the orthogonal nature of the vectors, this reduces to $\mathbf{\Sigma} \mathbf{b}_r = \lambda \mathbf{b}_r$ or the equivalent expression $(\mathbf{\Sigma} - \lambda \mathbf{I})\, \mathbf{b}_r = \mathbf{0}$, to produce the *characteristic equations* for the r^{th} or general case.

The first r principal components of $\mathbf{x}$ are therefore $\mathbf{b}_1'\mathbf{X}, \mathbf{b}_2'\mathbf{X}, \ldots, \mathbf{b}_r'\mathbf{X}$, with $\mathbf{b}_j$ the ortho-normalised eigenvectors of $Var(\mathbf{X}) = \mathbf{\Sigma}$ corresponding to the j^{th} largest eigenvalue of $\mathbf{\Sigma}$.

In cases where the eigenvalues are equal there is no unique way of choosing the corresponding eigenvectors, but as long as they are orthogonal, the argument will remain valid. That is.....

From the above, that is if the covariance matrix $Var(\mathbf{X}) = \mathbf{\Sigma}$ is positive definite with p distinct eigenvalues, there exists a unique orthogonal matrix $\mathbf{B}$ such that $\mathbf{C} = \mathbf{B}'\mathbf{X}$ represents a vector of the principal components of the original dataset $\mathbf{X}$. The columns of $\mathbf{B}$ constitute the ortho-normalised eigenvectors of $\mathbf{\Sigma}$, arranged in such a way that the first column is the ortho-normalised eigenvector of the largest $e$igenvalue and the last column the eigenvector for the smallest eigenvalue.

$Var(\mathbf{C})$ is a diagonal variance-covariance matrix, with the eigenvalues of $\mathbf{X}$, representing the variances of the components of $\mathbf{c}$ on the diagonal and zero elsewhere, reflecting the orthogonal nature of the components, since

$$\begin{aligned} Var\left(\mathbf{C}\right) &= Var\left(\mathbf{B}'\mathbf{X}\right) \\ &= \mathbf{B}'Var\left(\mathbf{X}\right)\mathbf{B} \\ &= \mathbf{B}'\mathbf{\Sigma}\mathbf{B} \\ &= \mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix} \end{aligned}$$

We can express $Var(\mathbf{C}) = \mathbf{\Lambda} = \mathbf{B}'\mathbf{\Sigma}\mathbf{B}$ giving the relationship between the covariance matrix of $\mathbf{X}$ and the corresponding principal components, namely $\mathbf{\Sigma} = \mathbf{B}\mathbf{\Lambda}\mathbf{B}'$.

Since

$$trace(\mathbf{\Lambda}) = trace\left(\mathbf{B}'\mathbf{\Sigma}\mathbf{B}\right) = trace\left(\mathbf{\Sigma}\mathbf{B}'\mathbf{B}\right) = trace\left(\mathbf{\Sigma}\right) = \sum_{j=1}^{p} Var\left(X_j\right)$$

We have shown that the sum of the variances of the original variables is equal to the sum of the variables of the principal components, meaning that all the variance in the original dataset is retained in the principal components. It also means that one can divide any of the variances of the elements of $\mathbf{c}$ by the sum of the variances of all the elements to determine the relative proportion of that variance (eigenvalue*)* and therefore the relative importance of a particular principal component, i.e.,

$$\frac{\lambda_i}{\sum_{j=1}^{j=p} \lambda_j}$$

There is often also value in determining the correlation coefficients between the principal components and the original variables (Radloff, 2015, p. 102).

Principal Components Derived from Standardised Variables It is common to calculate the principal components from a set of variables that have first been standardised, i.e.,

$$Z_i = \frac{X_i - \bar{X}}{S_i}$$

where $\bar{X}$ is the mean and S_i the standard deviation of X_i.

This would result in a sum of variance of Z_i to be equal to 1. In effect this means that one is finding the principal components from the correlation matrix of $\mathbf{X}$ rather than from the covariance matrix $\mathbf{\Sigma}$ as was demonstrated above.

The derivations would be the same, replacing $\mathbf{\Sigma}$ with the correlation matrix of $\mathbf{X}$, but the values of the eigenvectors and eigenvalues will be different and since the elements of the diagonal of $\mathbf{\Lambda}$ will all be 1, the sum of the diagonals will be equal to the rank p of the matrix.

Finally, $\mathbf{\Sigma}$, the population covariance matrix, is generally unknown and requires the estimation of $\mathbf{\Sigma}$ by $\hat{\mathbf{\Sigma}}$, which is estimated from the sample data (Radloff, 2015, p. 102).

Selection and Interpretation of Components

Selection of Components A key question relates to how many components to select for subsequent analysis since one of the reasons for doing PCA is to reduce the sample space to fewer variables, making it possible for the dependent variable to be regressed using the principal components rather than the original variables (Afifi et al., 2011, p. 363).

The question is which components should be used without losing too much of the information contained in the original dataset? According to Radloff (2015, p. 105), there are at least two alternative criteria to be considered, the choice of which depends on the purpose of the analysis:

- To delete those components that are relatively unimportant as predictors of the original independent variables. That is, delete the principal components with the smallest variance (or eigenvalues).
- To delete those components that are relatively unimportant as predictors of the dependent variable in the problem. That is, delete the principal components with the smallest absolute correlation coefficient with the dependent variable. In this case it's important to understand that the dependent variable need not be highly correlated with the principal components that have high eigenvalues in order for the "explanatory power of the complete principal components regression to be high".

Afifi et al. (2011, pp. 367-368), argue that "ideally, we wish to obtain a small number of principal components, say two or three, which explain a large percentage of the total variance, (for example) 80% or more", but that since this ideal is rare in practice, a compromise would be to choose as few components as possible to explain "a reasonable percentage" of the total variance.

They describe a "rule-of-thumb" according to which only those principal components explaining at least $100/p$ percent of the total variance are selected, where p denotes the number of variables(Afifi et al., 2011).

Furthermore, Afifi et al. (2011, pp. 367-368) argue that it is important to realise that the eigenvalues represent "estimated" variances of the principal components and can show large sample variations and therefore that "arbitrary cut-off points and small differences should not be taken too seriously". They conclude their discussion on the selection of components by stating that the selection of the number of principal components should be made on the "basis of some underlying theory".

Grimm & Yarnold (2004, p. 103) describe different types of "stopping rules" to decide on the selection of components :

- The percentage of variance criterion: Specifying *a priori* that components will be included until "some absolute percentage of the total variance has been explained";
- The *a priori* criterion: In cases where one is trying to replicate a previous study, one knows in advance how many eigenvectors to extract. This approach would also be appropriate when one has a "theoretically motivated" idea about the appropriate number of eigenvectors to extract.
- Kaiser's (1960) stopping rule: Extract only eigenvectors with eigenvalues of at least 1, that is the equivalent of the variance of a single standardised variable.
- The "scree test": Catell (1966) proposed a graphic procedure, described by Afifi et al. (2011, p. 364) as the "elbow rule", that consists of plotting the eigenvalues for successive components. These eigenvalues usually drop quickly after the first few before stabilising to a more gradual decline. The components (eigenvectors) in the steep decline are retained while those in the gradual decline are not (Grimm & Yarnold, 2004, pp. 103-104).

Afifi et al. (2011, p. 364) describe rules that approach the subject from the opposite perspective, namely to discard principal components with the smallest variances. In this regard one such a rule is to discard all components that have a variance less than $70/p$ percent of the total variance, i.e., as opposed to the $100/p$ rule-of-thumb described above for selecting components. Another rule is to discard any components that explain only small proportions of the variance, for example less than 5%, since they may simply reflect random variations in the data.

Citing Stevens's (1986) summary of research on the accuracy of these stopping rules, Grimm & Yarnold (2004, p. 104) conclude that Kaiser's (1960) stopping rule should be used when there are fewer than 30 variables and where the communalities are at least 60%; alternatively they argue that the "scree test" should be used in applications for which there are at least 200 observations and where the communalities are reasonably large.

Afifi et al. (2011, p. 363) argue that none of the stopping rules work well in all circumstances and that they should be used to provide guidance only. The authors conclude that the selection of the number of components should, if possible, be driven by the underlying theory and the purpose of the analysis and that since PCA is often used as an initial exploratory technique and hence none of the "rules" should be taken too seriously and that one should retain as many components as one can either interpret or are useful in future analyses.

Interpretation of Components Once the number of components has been decided, how to interpret and use them become the next questions. In interpreting the components, one needs to examine the coefficients b_{ij} defining each of the components C_i and note that a high coefficient of a principal component on a given variable is an indication of high correlation between that variable and its principal component (Afifi et al., 2011, p. 368).

Grimm & Yarnold (2004, pp. 105-106) discuss rotating the axes as part of the interpretation of principal components, but this appears to be more commonly the practice in factor analysis. Accordingly it will be discussed in more detail in chapter 5.4.3.

Interpretation can be facilitated by standardising the original variables before conducting PCA, that is subtracting the mean and dividing by the standard deviation (Afifi et al., 2011, p. 366), thus results in an analysis of the correlation matrix rather that the covariance matrix. The interpretation is made simpler for the following two reasons:

1. Using standardised variables means that the total variance would be equal to the number of variables p. Therefore the proportion of variance explained by any particular principal component is simply the corresponding eigenvalue divided by p.

2. The correlation between any particular principal component and any particular variable is simplified to $r_{ij} = b_{ij}(Var(C_i)^{\frac{1}{2}}$. This provides a measure of the degree of dependence of a given principal component on each of the standardised variables. This correlation is called the *factor loading*, which will also be discussed in greater detail in chapter 5.4.3.

It is important to note that the coefficients obtained from standardised data are quite different from those obtained from raw (original) data. According to Afifi et al. (2011, p. 366), there is no easy way to convert results obtained from one input matrix to those obtained from the other. If the covariance matrix is used and the scale of one of the variables is changed, the results of the principal components analysis will also change. They argue, however that "the majority of researcher prefer to use the correlation matrix because it compensates for the units of measurement of the different variables. But, if it is used, then all interpretations must be made in terms of the standardised variables". This dependence of the principal components on the scales of measurement is considered one of the drawbacks of PCA(Radloff, 2015, p. 106).

For the results of a PCA analysis to be reliable Grimm & Yarnold (2004, p. 100) argue that the sample should be more than five times as large as the number of variables, described as the "subjects-to-variables ratio" (STV). Citing Gorsuch (1983), Grimm & Yarnold (2004) argue than any analysis should contain a minimum of 100 observations, irrespective of the STV ratio.

Kline (1994, p. 39) states that although it is often done, it would be wrong to interpret the first principal component as indicating the presence of a general factor, i.e., the first principal component, which will have large positive loadings on most of the variables when most of the correlations in the input matrix are positive. He argues that such a result is "an artifact of the method" and should not be interpreted as indicating the presence of a general factor in the data. Furthermore, after the first principal component, subsequent components are usually bipolar, that is they have both positive and negative loadings. Kline (1994, p. 39) holds it is important to understand that this bipolarity is also an “artifact of the method” and should not be interpreted as reflecting anything in the original data.

In summary, Kline (1994, p. 39) states that the fact that PCA “produces an arbitrary general factor followed by bipolar factors makes interpretation of results difficult” and that “factors with many positive and negative loadings are also hard to interpret”. As a result, methods of simplifying PCA, such as the rotation of axes (see Chapter 5.4.3) have been developed

5.4.3 Exploratory Factor Analysis

5.4.3.1 Introduction

This section will build on the understanding, derivation and application of Principal Components Analysis.

The main purpose of Exploratory Factor Analysis (EFA) and the similarities of and differences between PCA and EFA will be discussed in section 5.4.3.2. This will be followed my considering methods of deriving factors in section 5.4.3.3 which will be concluded by a description of the maximum likelihood method of factor extraction in section 5.4.3.3. Goodness of fit tests will be considered in section 5.4.3.3, followed by brief discussions on rotation and the interpretation of factors in sections 5.4.3.4 and 5.4.3.6 respectively.

5.4.3.2 Description and Purpose

> The "essential purpose of Exploratory Factor Analysis is to describe, if possible, the covariance relationships among many variables in terms of a few underlying, but unobservable, random quantities called factors. Basically, the factor model is motivated by the following argument: Suppose variables can be grouped by their correlations. That is, suppose all variables within a particular group are highly correlated among themselves, but have relatively small correlations with variables in a different group. Then it is conceivable that each group of variables represents a single underlying construct, or factor, that is responsible for the observed correlations (Johnson & Wichern, 2002, p. 477)."

Exploratory Factor Analysis (EFA) is closely related to Principal Components Analysis (PCA). As in PCA, EFA is a dimension-reduction technique that aims to identify a small set of readily interpreted eigenvectors that explain most of the variation in a given dataset. While these eigenvectors are called "components" in PCA, they are called "factors" in EFA (Grimm & Yarnold, 2004, p. 106).

Afifi et al. (2011, pp. 379-380) describe PCA and EFA as being similar in that they are both techniques for examining and exploring the interrelationships among multiple variables and neither of these techniques make a distinction between dependent and independent variables. They differ, however, in their main objectives: In PCA this is to select a number of components that explain as much of the variation in the data as possible, while in EFA the factors are chosen mainly to explain the interrelationship among the original variables.

According to Kline (1994, p. 36), the terms "components" and "factors" are often used interchangeably. Citing Harman (1976), he argues that although there is a real difference that should be understood, the difference between components and factors become trivial as sample sizes increase. Kline (1994, p. 36) defines a factor as "a linear combination of variables, any combination, constitutes a factor" and that "what is required in factor analysis is a combination of variables so weighted as to account for the variance in the correlations." He defines *factor loadings* as the "correlations of the variables with the factor, the weighted combination of variables which best explains the variance."

Kline (1994, p. 36) argues that components are "real factors" in that they are derived directly from the data, while the common factors from EFA are "hypothetical" since they are estimated from the data and, while PCA explains all the variance in the data, factor analysis does not necessarily do so. According to Kline (1994), this can be an advantage since it is unlikely that factors can explain all the variance in a dataset because "all correlations contain error, the full account of principal components must be contaminated by error."

According to Grimm & Yarnold (2004, p. 107), the difference between EFA and PCA relate to their assumptions. While the key assumption in PCA is that the total variance in a given sample is described by the sum of the explained and the error variances; in EFA the total variance is made up of the sum of three different kinds of variances, namely the *common variance*, the *specific variance* and the *error variance.*

The *common variance* is a reference to that portion of the total variance that is shared with other variables in the analysis; the *specific variance* refers to that portion of the variance that does not correlate with other variables; and the *error variance* - as in PCA - describes "inherently unreliable random variation" (Grimm & Yarnold, 2004, p. 107).

While PCA assists in determining underlying dimensions, EFA could be used to not only determine the number of dimensions underlying the data, but also the meaning of each of these dimensions, and how these dimensions interrelate with each other (Grimm & Yarnold, 2004, p. 99).

5.4.3.3 Derivation of an Orthogonal Factor Model

This section first describes some of the basic terminology and principles of EFA before providing some detail on the derivation of the Principal Components and Iterative methods of deriving factors in sections 5.4.3.3 and 5.4.3.3 respectively. It then describes the Maximum Likelihood method and a test for a model's goodness of fit in section 5.4.3.3.

Basic Terminology and Principles Consider, as in section 5.4.2.3 above, a set of independent random variables, $\mathbf{X} = [X_1, X_2, ..., X_p]$

The aim of factor analysis is to represent each of these variables as a linear combination of a smaller set of "common factors" plus a specific factor that is unique to each variable.

According to Afifi et al. (2011, pp. 381-383), this *factor model* is described as

$$\begin{aligned}
X_1 &= \lambda_{11}F_1 + \lambda_{12}F_2 + ... + \lambda_{1m}F_m + e_1 \\
X_2 &= \lambda_{21}F_1 + \lambda_{22}F_2 + ... + \lambda_{2m}F_m + e_2 \\
&\vdots \\
X_p &= \lambda_{p1}F_1 + \lambda_{p2}F_2 + ... + \lambda_{pm}F_m + e_p
\end{aligned}$$

where:

1. m represents the number of common factors, ideally smaller than the number of variables in the original dataset p.
2. $F_1, F_2, ..., F_m$ are the common factors. They are assumed to have zero means and unit variances.
3. λ_{ij} denote the coefficients of F_j and are called the *factor loadings* of the i^{th} variable on the j^{th} common factor.
4. $e_1, e_2, ...e_p$ are the unique or specific factors, each relating to one of the original set of variables.

The factor model in matrix notation can be described as

$$\mathbf{X} = \mathbf{\Lambda F} + \boldsymbol{\epsilon}$$

This factor model breaks each response variable X_i into two parts: one due to the common factors and another due to the unique factor. This also breaks the variance of X_i into two parts: the *communality*, or that part of the variance that can be attributed to the common factors and the *specificity*, that part of the variance that is due to the unique factor e_i . Since the data has been standardised, the sum of these two variance parts is equal to 1. The *communality* is denoted by h_i^2 and the *specificity* by u_i^2 and therefore $Var(X_i) = h_i^2 + u_i^2 = 1$ Factor analysis is concerned with finding estimates of the factor loadings, λ_{ij}, and the communalities, h_i^2. According to Afifi et al. (2011), there are many ways in which these estimates can be determined in processes that involve an initial extraction of factors; followed by rotation to generate new factors that assist in the interpretation.

To extract the initial factors, the principal components (5.4.3.3), the iterative (5.4.3.3) and the maximum likelihood (5.4.3.3) methods will be considered, before describing rotation in section 5.4.3.4.

In all the methods the number m of common factors is a required input, which is ideally known *apriori*. If it is not known, a default option set in most computer programmes sets the number of common factors equal to the number of eigenvalues greater than 1 Afifi et al. (2011). Alternatively, as demonstrated by Kline (1994), the use of PCA as a preliminary investigative technique can be used to establish a sense of the underlying number of common factors to use in subsequent EFA. According to Afifi et al. (2011), Gorsuch (1983) can be considered for a review of several other methods for choosing m numerically.

Principal Components Analysis Method The following section draws mainly on Afifi et al. (2011, pp. 383-386) and is described here as an intuitively accessible explanation of factor analysis and clearly demonstrates how it relates to PCA.

In this method the first m principal components are modified to fit the factor model described above, since not only are they orthogonal vectors, but they would also explain most of the variance in the data, making them attractive options to use as common factors.

Recall the linear equations that constitute the principal components in section 5.4.2.3:

$$\begin{aligned} C_1 &= b_{11}X_1 + b_{12}X_2 + ... + b_{1p}X_p \\ &\vdots \\ C_p &= b_{p1}X_1 + b_{p2}X_2 + ... + b_{pp}X_p \end{aligned}$$

Or, in terms of linear algebraic terms:

$$\mathbf{C} = \mathbf{BX}$$

Consider the common factor F_j, which given the assumption of unit variance, can be expressed as $F_j = C_j/\sqrt{Var(C_j)}$, i.e. dividing a particular principal component by its standard deviation. Using this relationship the set of linear equations describing the principal components can be inverted to a set of linear equations describing the standardised response variables as

$$\begin{aligned} X_1 &= b_{11}C_1 + b_{21}C_2 + ... + b_{p1}C_p \\ &\vdots \\ X_p &= b_{1p}C_1 + b_{2p}C_2 + ... + b_{pp}C_p \end{aligned}$$

or

$$\mathbf{X} = \mathbf{B'C}$$

Given the relationship $F_j = C_j/\sqrt{Var(C_j)}$ between common factors and principal components, the i^{th} equation can be expressed as

$$X_i = b_{1i}F_1\left(\sqrt{Var(C_1)}\right) + b_{2i}F_2\left(\sqrt{Var(C_2)}\right) + ... + b_{pi}F_p\left(\sqrt{Var(C_p)}\right)$$

Let the factor loadings $\lambda_{ij} = b_{ji}\left(\sqrt{Var(C_j)}\right)$ be defined for the first m components while combining the last $(p - m)$ terms into a single term, e_i. That is, for the first m components the transformation of the principal components model into a factor model can be described as

$$X_i = \lambda_{i1}F_1 + \lambda_{i2}F_2 + ... + \lambda_{im}F_m + e_i \tag{5.3}$$

and e_i can be expressed as

$$e_i = b_{m+1,i}F_{m+1}\left(\sqrt{Var(C_{m+1})}\right) + ... + b_{pi}F_p\left(\sqrt{Var(C_p)}\right)$$

When the original X_i variables are standardised to Z_i the factor loadings λ_{ij} represent the correlation between Z_i and F_j.

The matrix of factor loadings is sometimes called the "pattern matrix", but when the factor loadings are correlations as in this derivation, it is referred to as the "factor structure matrix" (Afifi et al., 2011, p. 385).

Finally, it will be shown that the communality of X_i is given by $h_i^2 = \lambda_{i1}^2 + \lambda_{i2}^2 + ... + \lambda_{im}^2$

Consider equation 5.3 and recall that the communality of X_i is defined as that part of the variance of X_i that can be attributed to the common factors, that is excluding the specificity, or that part of the variance attributable to the unique factor e_i, hence the communality can be considered to be described as

$$h_i^2 = Var(\lambda_{i1}F_1 + \lambda_{i2}F_2 + ... + \lambda_{im}F_m)$$

Since the common factors F_i are constructed to be orthogonal, their covariances would equal zero; and since they are assumed to have unit variance, this expression of communality h_i^2 becomes

$$h_i^2 = \lambda_{i1}^2 + \lambda_{i2}^2 + ... + \lambda_{im}^2$$

Iterated Components Method The principal source for this description of the Iterated Components Method, which is also called the "principal factor analysis" or "principal axes factoring approach", is Afifi et al. (2011, pp. 386-390). As described above, the communality of a particular variable is that part of its variance that is associated with the common factors. In this method, the communalities are used instead of the original variances. Accordingly the unit variances of the input correlation matrix are replaced by estimates for the communalities. Only the variance associated with the common factors is considered as opposed to factoring the total variance contained in the response data. The communalities are not known *a priori* and therefore initial estimates need to be made. Afifi et al. (2011, p. 386) recommend using the default option in any given computer programme since "the resulting factor solution is usually little affected by the initial communality estimates". Kline (1994, p. 36) argues that this method of initial factor extraction is identical to PCA except for the replacement of the 1's in the diagonals as.

The steps in carrying out the "iterated factor extraction" are:

1. Find initial estimates for the communalities. The default option in most computer programmes would suffice;
2. Substitute these estimates for the diagonal elements in the correlation matrix, thus replacing the 1's in the diagonal;
3. Extract m principal components from the modified matrix;
4. Multiply the principal components' coefficients by the standard deviations of the respective principal components (as described in 5.4.3.3 above) to get factor loadings*;*
5. Calculate new communalities (also as described in section 5.4.3.3 above) from the factor loadings;
6. Replace the first communalities with these calculated ones and repeat steps 3, 4 and 5;
7. Continue this iteration until the communalities stabilise.

The factor loadings extracted by this method depend on the number of factors to be extracted. This is not the case in the principal component method.

It is possible to get negative variances and eigenvalues by this method since the standard variances of 1 are replaced sometimes by communalities considerably lower than one. These negative values should not be used in the analysis.

The Maximum Likelihood Method

Introduction If one is confident that the factor model is valid and that the variables have a multivariate normal distribution, then the maximum likelihood method (ML) should be considered. This method makes it possible to consider tests of hypotheses and/or determine confidence intervals ((Afifi et al., 2011, p. 390), citing Gorsuch, 1983). Another advantage of ML is that the estimates of the factor loadings are invariant to changes in scale of the original variables.

The ML method/procedure is also what is used in Confirmatory Factor Analysis (CFA) (see chapter 5.4.4) in which the constraints on the outcome of a factor analysis are specified *a priori* (Afifi et al. (2011, p. 390), citing Long, 1983; Bartholomew & Knott, 1999). According to Kline (1994, pp. 49-50), ML factor analysis differs from PCA and other methods of factor analysis in the following ways.

1. ML computes a set of factors, each of which in turn explains as much variance as possible of the population correlation matrix, as estimated by the sample correlation matrix. This as opposed to PCA or previously described methods which explain as much variation as possible in the sample matrix only. ML is considered a statistical method since inferences are made from a sample about a population. The consequence

of this is that large samples are even more essential than in PCA or other factor analytic techniques.

2. ML is particularly relevant (suited) to CFA, although it can also be used simply as a technique for dimension reduction.

3. ML as a method of variable reduction is usually used to search for factors, but according to Kline (1994, p. 50), when "test reliabilities and thus *communalities* are high, the difference between ML factor analysis and PCA becomes trivial".

4. Kline (1994, p. 50) holds that the strongest argument for using ML analysis is that it has statistical tests for the significance of each factor as it is extracted.

5. The maths of ML methods are more complex. According to Nunnaly (1978), cited in Kline (1994, p. 50), one needs a "solid grounding in calculus, higher algebra and matrix algebra" to understand them.

A Mathematical Model for Factor Structure The following draws on Morrison (2005, pp. 318-321). Consider the following factor model

$$\begin{aligned} X_1 &= \lambda_{11}F_1 + \ldots + \lambda_{1m}F_m + e_1 \\ &\vdots \\ X_p &= \lambda_{p1}F_1 + \ldots + \lambda_{pm}F_m + e_p \end{aligned}$$

where $X_i's$ represents observable random variables from a nonsingular distribution. For the sake of simplicity and since the focus is only on covariance, a population mean of zero will be assumed.

In this derivation, F_j represents the common factor variates and λ_{ij} denotes the factor coefficients of the j^{th} factor in the determination of the i^{th} response variable; or alternatively, the factor loading of the i^{th} response on the j^{th} common factor. The e_i denotes the i^{th} specific factor variate.

Let

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}, \mathbf{F} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix} \text{ and } \boldsymbol{\epsilon} = \begin{bmatrix} e_1 \\ \vdots \\ e_p \end{bmatrix}$$

and let

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1m} \\ \cdots & \cdots & \cdots \\ \lambda_{p1} & \cdots & \lambda_{pm} \end{bmatrix}.$$

The factor model can therefore be written as

$$\mathbf{X} = \mathbf{\Lambda F} + \boldsymbol{\epsilon}$$

Let the m common factor variates in $\mathbf{Y}$ be independently distributed with zero means and unit variances and assume that the elements of $\boldsymbol{\epsilon}$ are also independently distributed with mean zero and variances represented by the $p \times p$ matrix $\boldsymbol{\Psi}$, where its diagonals are made up by ψ_i, the specific variance or *specificity* of the i^{th} response variable. That is,

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_1 & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \psi_p \end{bmatrix}$$

The variance of the i^{th} response variable is,

$$Var(X_i) = Var(\lambda_{ij}F_j + \ldots + \lambda_{im}F_m + e_i)$$

Given the unity variance property of the factor variates, F_i , the definition of $Var(e_i) = \psi_i$ and the independence of the factor variates and $\boldsymbol{\epsilon}$, this can be written as

$$\sigma_i^2 = \lambda_{i1}^2 + ... + \lambda_{im}^2 + \psi_i$$

Similarly, the covariance of the i^{th} and j^{th} response variables can be expressed as

$$\sigma_{ij} = \lambda_{i1}\lambda_{j1} + ... + \lambda_{im}\lambda_{jm}$$

The population variance-covariance matrix for the response variables can therefore be written as:

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \cdots & \cdots & \sigma_{2m} \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1m} \\ \cdots & \cdots & \cdots \\ \lambda_{p1} & \cdots & \lambda_{pm} \end{bmatrix} \times \begin{bmatrix} \lambda_{11} & \cdots & \lambda_{p1} \\ \cdots & \cdots & \cdots \\ \lambda_{1m} & \cdots & \lambda_{pm} \end{bmatrix} + \begin{bmatrix} \psi_1 & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \psi_p \end{bmatrix}$$

Alternatively, in matrix form

$$\mathbf{\Sigma} = \mathbf{\Lambda\Lambda}' + \mathbf{\Psi} \tag{5.4}$$

The diagonal elements of $\mathbf{\Lambda\Lambda}'$, i.e.., $\sigma_i^2 - \psi_i = \sum_{j=1}^{m} \lambda_{ij}^2$, are called the *communalities* of the response variables and the parameter λ_{ij} represents the covariance of the i^{th} response and the j^{th} common factor.

This can be shown by considering $Cov(\mathbf{X}, \mathbf{F}') = E\left[(\mathbf{\Lambda F} + \boldsymbol{\epsilon})\,\mathbf{Y}'\right] = \mathbf{\Lambda}$, due to the independence of the factor variates F_i and the assumption of zero means in the specific factor variates, e_i.

In this matrix factorisation, we do not impose a condition that the sums of squares of the loadings decrease.

Now, let $\mathbf{T}$ represent an $m \times m$ orthogonal matrix. We can therefore represent equation 5.4.3.3 as

$$\mathbf{\Lambda T} \left(\mathbf{\Lambda T}\right)' + \mathbf{\Psi} = \mathbf{\Lambda T T' \Lambda'} + \mathbf{\Psi} = \mathbf{\Lambda \Lambda'} + \mathbf{\Psi} = \mathbf{\Sigma}$$

This demonstrates that although the elements of $\mathbf{\Lambda T}$ would be different to the original loadings matrix, $\mathbf{\Lambda}$, it produces the same population variance-covariance matrix, $\mathbf{\Sigma}$. From this it follows that one can choose infinitely many different orthogonal transformations, leading to the notion of finding an orthogonal transformation that leads to what Morrison (2005, p.321), citing Thurstone (1945), described as *simple structure.* Requirements for factor structure and rotations to how achieve them are discussed in section 5.4.3.4 below.

Estimating Model Parameters Now that the linear algebra model for factor structure has been defined, the question about how to estimate the model parameters $(\mathbf{\Lambda}, \mathbf{\Psi})$ will be considered

The fundamental assumptions of the derivation of these estimates are:

1. The raw data consists of n independent observations from a multi normal population with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma} = \mathbf{\Lambda \Lambda'} + \mathbf{\Psi}$, as derived in section 5.4.3.3 above.

2. $\mathbf{\Sigma}$, the variance-covariance matrix of $\mathbf{X}$ is of full rank with dimension $p \times p$.

3. The factor loading matrix $\mathbf{\Lambda}$ has m common factor columns, where m has been specified *a priori.*

4. The information in the sample covariance matrix $\mathbf{S}$ is considered to be sufficient to estimate the factor parameters

5. The fundamental likelihood function for $\mathbf{S}$ can best be described by the Wishard density, i.e. $f(\mathbf{S}) = C \mid \mathbf{S} \mid^{1/2(n-p-1)} \mid \mathbf{\Sigma} \mid^{-1/2n} \exp\left(-\frac{1}{2}n\mathrm{tr}\left\{\mathbf{\Sigma}^{-1}\mathbf{S}\right\}\right)$

The maximum likelihood estimators are determined by transforming the Wishard density and using equation 5.4.3.3 above to derive expressions for the relationships between the ML estimators. The solutions for the *characteristic equations* are then attained through an iterative procedure that can be initiated by making an initial estimate of $\boldsymbol{\psi}$, i.e. $\hat{\mathbf{\Psi}} = \mathrm{diag}\left(\mathbf{S} - \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}'\right)$. The process would continue until stability in the parameter estimates $(\hat{\mathbf{\Psi}}, \hat{\mathbf{\Lambda}})$ is achieved. Details of these derivations can be found in Morrison (2005, pp. 322-327).

Goodness of Fit of the Factor Model The ML method of deriving the loadings estimators allows for a formal test of the "goodness of fit" or adequacy of a given m-factor model.

The null hypothesis is described as:

$$H_o : \ \boldsymbol{\Sigma} = \boldsymbol{\Psi} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}' \tag{5.5}$$

where $\boldsymbol{\Lambda}$ represents the "loading matrix" of dimension $p \times m$ for the number of variables p and the number of common factors m; $\boldsymbol{\Sigma}$ is a $p \times p$ positive definite matrix representing the variance-covariance matrix of standardised variables; and $\boldsymbol{\Psi}$ represents the diagonal matrix containing the specific variances of the response variables. All of these represent the hypothesised population parameters.

Using a likelihood-ratio procedure, leads to the following χ^2 test statistic (Morrison, 2005, p. 327):

$$\chi^2 = \left[n - 1 - \frac{1}{6}(2p{+}5) - \frac{2}{3}m \right] \ln \left\{ \frac{\left| \hat{\boldsymbol{\Psi}} + \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' \right|}{|\mathbf{S}|} \right\} \tag{5.6}$$

Here $\hat{\boldsymbol{\Psi}}$ and $\hat{\boldsymbol{\Lambda}}$ are the solutions to the ML method developed in section 5.4.3.3 above and $\mathbf{S}$ represents the sample covariance matrix. According to Morrison (2005, p. 327), the coefficient was developed by Bartlett (1954) to improve convergence.

Morrison (2005, p. 327) argues if the hypothesis in 5.5 is true, and as n becomes larger, the test statistic 5.6 tends to a chi-squared distribution with v degrees of freedom, defined as:

$$v = \frac{1}{2}\left[(p-m)^2 - p - m \right]$$

The null hypothesis of having exactly m common factors would be rejected at the level of α if $\chi^2 \geq \chi^2_{\alpha;v}$. The same value would result for the covariance or the correlation matrix due to the *invariance* property of the estimated loadings and specific variances (Morrison, 2005, p. 327).

If $m = 0$, the null hypothesis would imply that the population covariance matrix is diagonal and the test statistic would be equivalent to testing the hypothesis for no correlation, as shown in Morrison (2005, p. 38).

In most cases one does not know the number of common factors m. According to Morrison (2005, p. 328), one would begin with a low estimate and then try successively larger numbers until either the hypothesis is no longer rejected or the procedure fails to converge.

It is important to note that the χ^2 statistics obtained in this procedure would not be independent, and as a result the true significance level of the test may be different from the nominal value used at each stage of the extraction. Also, according to Morrison (2005, p. 328), the degrees of freedom v for the test statistic needs to be positive, implying that the number of common factors cannot exceed the expression:

$$m < \frac{1}{2}(2p + 1 - \sqrt{8p + 1})$$

Citing Amemiya and Anderson (1990), Morrison (2005, p. 328) argues that the χ^2 goodness-of-fit statistic has a large-sample chi-squared distribution under the linear factor model if the common- and specific- factor variables are independent and that it is not necessary that those variables be multivariate normally distributed.

5.4.3.4 Rotation

Kline (1994, p. 61) posits that the results of the "first condensation" of factor analysis (or PCA) by any method, should not be interpreted as the final solution. Citing Steiger (1979) Grimm & Yarnold (2004, p. 105) ague that in most cases, the eigenvectors identified in PCA and factor analyses will be very difficult to interpret since there is no unique location for the eigenvectors - a dilemma they describe as "factor indeterminacy".

In this section the issues of rotation in order to achieve *simple structure* will be considered since, as Grimm & Yarnold (2004, p. 105) claim, when *simple structure* has been achieved, interpretation is often a "straightforward procedure."

Kline (1994, pp. 56-64) provides simple algebraic examples that show that the rotation of factors changes the factor loadings and therefore the meaning of factors, but that the different factor solutions are mathematically equivalent in that they explain the same amount of variance in each variable and as a result in the matrix as a whole; and that rotated factors also reproduce the original correlations. He also demonstrates that there are a virtual infinity of different ways in which factor axes can be rotated.

The properties of *simple structure,* as outlined by Thurstone (1974), cited in Grimm & Yarnold (2004, p.105) are:

1. Each variable should have at least one loading that is near zero on at least one of the vectors; and for situations with four or more eigenvectors, most of the variables should have loadings that are near zero on most of the eigenvectors.

2. For each eigenvector there should be at least as many variables with loadings that are near zero as there are eigenvectors.

3. For every pair of eigenvectors, there should be several variables that load on only one eigenvector. In general, variables should have a large loading on one eigenvector.

Grimm & Yarnold (2004, p. 105) describe different types of rotations, broadly distinguished by whether they are *orthogonal* or *oblique.*

The most commonly used orthogonal rotations include *varimax*, which aims to make as many values in each column of the factor loading coefficient table as close to zero as possible and *quartimax*, which aims to make as many values in each row of the table as close to zero as possible. Both try to achieve as much simple structure as possible without losing the independence between the eigenvectors.

In *oblique* rotations, independence is forfeited, leading to factors that may be correlated. Once simple structure has been achieved, the rotated coefficients need to be considered carefully to find the *central dimension* identified by the eigenvector Grimm & Yarnold (2004, p. 105).

For oblique rotations, the *promax* rotation is fast and conceptually simple. It tries to fit a target matrix which has a simple structure and requires two steps. The first, mostly using a varimax rotation, defines the target matrix by forcing the structure of the loadings to become bipolar. The second step involves a least square fit from the varimax solution to the target matrix. The results of oblique rotations are generally interpreted not by graphical means, but by considering the correlations between the rotated axis and the original variables, which are interpreted as loadings. (Abdi, ud, p. 6)

5.4.3.5 Factor Scores

Following the extraction and rotation of factors, it would be necessary to obtain aggregate scores that link items with the factors to allow for further analyses of the data. These scores can be obtained in various ways. The simplest is to add the values of the variables loading heavily on a given factor. Alternatively a *regression* procedure is used to determine factor scores according to which the inter-correlations among the X_i variables are combined with the factor loadings to produce *factor score coefficients.* These factor scores can then be used as data for further analyses. (Afifi et al., 2011, p. 396)

5.4.3.6 Interpretation

Once a set of initial factors are obtained, new factors are generated by rotating the axes, as described in section 5.4.3.4, in order to facilitate interpretation Afifi et al. 2011, p. 383.

One of questions raised in interpretation is what value of factor loading (coefficient) is required for a variable to be considered as a constituent of a given eigenvector? Grimm & Yarnold (2004, p. 106) hold that researchers typically consider variables with factor loadings of at least the absolute value of 0.3, that is the variable and the eigenvector share $(0.3)^2$ or 9% of their variance. However the practice of only considering factor loadings of greater than 0.3 ignores the sample size, which should be considered since the statistical significance of the correlation between a variable and an eigenvector depends on the sample size (Grimm & Yarnold, 2004, p. 106, citing Stevens (1986)).

Variables with negative factor loading coefficients are negatively correlated with the eigenvector. Eigenvectors that have positive and negative loadings are called *bipolar* eigenvectors Grimm & Yarnold (2004, p. 106).

When items are virtual paraphrases of each other, that is for example when a statement is simply rephrased, they will load as a single factor described as *bloated specifics* that look like factors but are really only specific variance. It is possible to discriminate between specific and common factors by the fact that the specific factors correlate with no other factors or external material (Kline, 1994, pp. 12, 112, citing Cattell (1978)).

Finally, when a component correlates with only one variable, this may indicate that the variable should be used as is and that it is not part of the latent variable structure of the data. This would not be ideal given the ordinal scale of the measurement instrument.

5.4.4 Confirmatory Factor Analysis

5.4.4.1 Introduction

This section will begin with a description and purpose of CFA before considering aspects related to the conducting of a CFA in section 5.4.4.3, which is followed by notes on the interpretation of CFA in section 5.4.4.5. An application of CFA is considering in section **??**. Since, according to Kline (1994, p. 80), CFA is "highly complex algebraically" and since the emphasis in this project is less on multiple regression techniques, the formal mathematical derivation of the methods employed in CFA will not be considered here.

5.4.4.2 Description and Purpose

CFA (Confirmatory Factor Analysis) is a type of structural equation modeling that deals specifically with measurement models; that is, the relationships between observed measures (indicators) and latent variables (factors). The goal is to establish the number and nature of factors that account for the variation and covariation among a set of indicators. As a "confirmatory technique it is theory driven" in which the planning of the analysis is driven by the theoretical relationships among the observed and unobserved variables (Schreiber et al., 2006, p. 323).

In CFA, much as with EFA, the observed data are assumed to be indicators of one or more underlying, latent constructs or factors. It similarly assumes that there are two main sources of variation in the response variables, namely variation due to common factors and variation from unique measurement error, which is in turn assumed to be a combination of variance specific to a given indicator and variance due to random error (Brown & Moore, ud; Bryant & Yarnold, 2004).

The difference between CFA and EFA is that while EFA aims to find the single underlying factor model that best fits the data, CFA can test more precise hypotheses. One can, for example, specify which items belong to which factors and how the factors relate to each other and it can be used to identify the model that offers to the best fit (Bryant & Yarnold, 2004, pp. 12, 109). According to Bryant & Yarnold (2004, p. 109, citing Bollen, 1989, Hayduk, 1987, Long, 1983), EFA "primarily represents a tool for theory building, whereas CFA represents a tool for theory testing".

Brown & Moore (ud, p. 2) argue that the two differ fundamentally by the number and nature of *a priori* specifications and restrictions made on the latent variable measurement model. EFA is used as an exploratory or descriptive technique to ascertain the number of common factors and to determine which measured variables are reasonable indicators of the various latent dimensions; while in CFA the number of factors, the pattern of indicator-factor loadings as well parameters such as those related to the independence or covariance of the factors and indicator unique variances are specified. This pre-specified factor solution is evaluated on the basis of how well it reproduces the sample covariance matrix of the measured variables (Brown & Moore, ud).

In practice CFA and EFA are often used together. For example, according to Brown & Moore (ud, p. 3), EFA is often used early in the process of scale development and construct validation while CFA is used later when the underlying structure has been established on prior empirical and theoretical grounds; and, according to Bryant & Yarnold (2004, p. 109), with multiple samples of sufficient size EFA can first be used to identify or discover a possible factor structure using data from one sample, while CFA could then be used to "simplify, refine and confirm" this basic model using the other samples' data. The two techniques can be viewed as opposite, yet complimentary, sides of a coin. Whereas EFA involves hindsight, CFA requires foresight" (Bryant & Yarnold, 2004, p. 109).

Bryant & Yarnold (2004, pp. 119-120) state that the most common single sample uses of CFA are to: a) see whether one particular factor model fits the data better than others, b) determine whether oblique factors fit the data better than orthogonal models and c) test hypotheses about differences in the strength of the relationship between pairs of latent factors. CFA can also be used effectively in multi-sample hypothesis testing, for example by testing whether the same factor structure holds across multiple groups, known as *simultaneous* CFA, which allows one to test hypotheses such as the invariance of factor loadings and unique error terms for a given model across independent samples (Bryant & Yarnold, 2004, p. 112, citing Alwin and Jackson, 1979, and others).

According to Brown & Moore (ud, p. 3), CFA is commonly used in the process of scale development to examine the latent structure of a test instrument by verifying the number of underlying dimensions (factors) of the instrument as well as the factor loadings, the pattern of item-factor relationships. CFA is also used to determine how a test should be scored, for example when the latent structure consists of two or more factors, the pattern of factor loadings will indicate which items load on which factors Brown & Moore (ud, p. 3). CFA is also routinely used to test the significance of a particular model or an hypothesised set of factors, including definitions about which variables would fit under which factor. It is used to compare factor models across different scenarios such as testing competing models with different estimates for the number of latent constructs (Bryant & Yarnold, 2004, p. 100).

Unlike EFA, which extracts factors from the data to maximise the common variance, CFA uses whatever model is specified to generate a predicted set of item interrelationships (Bryant & Yarnold, 2004, pp. 110-111) or, as (Schreiber et al., 2006, p. 323) describe, an estimate of a population covariance matrix that is compared with the observed sample covariance matrix, with the researcher aiming to minimise the difference between the observed and estimated matrices.

According to Brown & Moore (ud, p. 6), all CFA models contain the parameters of factor loadings, unique variances, and factor variances; where factor loadings reflect the regression slopes for predicting the indicators from the factors, unique variance is the variance in the indicator that is not accounted for by the factors, and the factor variance which is the dispersion of the sample responses on a particular latent variable or factor Brown & Moore (ud, p. 6).

5.4.4.3 Conducting CFA

A CFA requires an input matrix of either correlations or covariances of the original data as well as information about factor loadings, factor interrelationships, and measurement errors Bryant & Yarnold (2004, p.115).

With regard to the input matrix, standardisation of the raw data to produce correlation matrices is important when the observations reflect different units or scales of measurement or when data from separate groups are combined. If however the groups have significant differences in their means, simply combining them can result in "spurious correlations". It is generally better not to standardise data within groups when one is interested in exploring structural differences between groups since group differences in variability contained in co-variances can be obscured by using correlation matrices (Bryant & Yarnold, 2004, p. 115, citing Cunningham, 1978 and Joreskog and Sorbom, 1989).

The number of hypothesised latent factors and the pattern of item loadings that define each factor need to be stipulated. In deciding which factor loadings to include in a CFA model one should aim to develop "parsimonious" models in which the items load on as few factors as are necessary to provide a reasonable fit of the data (Bryant & Yarnold, 2004, p. 115). Brown & Moore (ud, p. 6) state that if "substantively justified", a CFA could also include error covariances that stipulate that two indicators co-vary for reasons other than the shared influence of the latent factors, for example due to *method effects*.

The nature of the relationship among the latent factors also need to be specified. These can either be independent (orthogonal), inter-correlated (oblique), or a combination of the two in which some are orthogonal and others oblique. By considering the chi-square values from models with orthogonal vs oblique factors, one can test for the hypothesis that the factors are interrelated (Bryant & Yarnold, 2004, pp. 115-116). The latent variables can also be defined as either *endogenous* or *exogenous* variables.

Brown & Moore (ud, p. 6) argue that when the CFA solution has two or more factors, a factor covariance is "almost always" specified to estimate the relationship between latent dimensions.

Unlike EFA, the nature of the relationships among the indicator unique variances can be modeled in CFA "provided that this specification is substantively justified and that the other identification requirements are met". When measurement error is specified to be random, (i.e. the unique variances are uncorrelated), the assumption is that the observed relationship between any two indicators loading on the same factor is due entirely to the shared influence of the latent variable (Brown & Moore, ud, p. 4). According to Bryant & Yarnold (2004, p.

116), researchers often allow for correlated measurement error in their CFA models, especially when these improve the fit of models that are already grounded in theory.

In practice, CFA is often “confined to the analysis of variance-covariance” structures. When that is the case, the parameters (factor loadings, error variances and covariances, and factor variances and covariances) are estimated to reproduce the input variance-covariance matrix. The analysis of covariance structures is based on the assumption that indicators are measured as deviations from their means, although the CFA model can be expanded to consider the analysis of mean structures (Brown & Moore, ud, p. 8).

In a CFA model parameters can be specified in three different ways: *free*, *fixed* or *constrained.* A *free* parameter is unknown and the analysis strives to find its optimal value that minimises the difference between the observed and predicted variance-covariance matrix, a *fixed* parameter is specified *a priori,* most often either 1 or 0, a *constrained* parameter is unknown, but while its not free to be any value, the values it can assume are restricted. The most common example of a constrained parameter is when it is restricted to be equal to another parameter (Brown & Moore, ud, p. 9).

The output of CFA and give parameter estimates as *completely standardised*, when both the latent variable and indicator are standardised and as a result the factor loading of an indicator that loads on only factor can be interpreted as the correlation between the indicator and the factor. Results can also be *unstandardised* reflecting the original input metrics or *partially standardised,* where either the indicator or latent variables are standardised (Brown & Moore, ud, p. 10).

To be able to estimate a CFA solution, the measurement model must be *identified*, which can occur if on the basis of the sample variance-covariance matrix a unique set of estimates for each parameter in the model can be obtained. Two issues that can inhibit identification are the *scaling* of the latent variables and *statistical identification.*

To understand scaling its important to realise that the units of measure of the latent variables need to be identified by the researcher. This is done in one of two ways: the most widely used method is the *marker indicator* approach according to which the unstandardised factor loading of one observed measure per factor is fixed to a value of 1; alternatively, the variance of the latent variable is fixed to a value of 1 (Brown & Moore, ud, p. 10). Bryant & Yarnold (2004, pp.117-118, citing Alwin and Jackson, 1979, 1980), suggests fixing the measurement scale for each latent factor by constraining either the factor variances to one, providing a standardised solution, or by fixing one loading on each factor (usually that of the highest loading indicator) to one .

Statistical identification relates to the fact that a CFA solution can only be estimated if the number of freely estimated parameters (i.e., factor loadings, uniquenesses, factor correlations) does not exceed the information in the input matrix (e.g., the number of sample variances and covariances). In this regard, a model is *over-identified* when the number of known elements exceeds the number of unknowns and it is *under-identified* when the opposite is the case.

The difference in the number of known and unknowns denote the model's *degrees of freedom* (df). An over-identified solution (i.e., when the df is positive), can result an output and accompanying goodness-of-fit evaluation, but an under-identified solution (when the df is negative) cannot be estimated Brown & Moore (ud, p. 11).

To overcome a problem of under-identification, one can specify or fix the values of some parameters to reduce the number of unknowns. In particular, for a model to be identified a minimum of k^2 elements must be fixed in which k is the number of latent factors in the model. On the other hand if a model is over-identified, meaning multiple estimates can be derived for free parameters one can increase the number of unknowns by freeing more parameters Bryant & Yarnold (2004, pp.117-118, citing Joreskog and Sorbom, 1989).

A CFA model can be properly identified, but still be *mis-specified,* which relates to whether a model is different from the structures that the observed data would support. Model mis-specification can result in solutions in which unique errors are negative, factor inter-correlations are more than one, or parameter estimates are very large. While trivial mis-specification, for example erroneously including a factor loading when the correct model has none, has little effect on goodness-of-fit indexes, more substantive mis-specification, for example leaving out an important factor loading, can dramatically lower the value of the goodness-of-fit indices (Bryant & Yarnold, 2004, p. 118, citing La Du and Tanaka, 1989 and Bagozzi and Yi, 1988).

Finally, an important difference between EFA and CFA is that in EFA all possible relationships (factor loadings) between the factors and indicators are freely estimated, allowing for *cross-loadings* where an indicator is predicted by more than one factor. In a CFA all cross-loadings are fixed to zero. As a result, while EFA models with two or more factors are subjected to rotations in order to achieve *simple structure* and therefore more accessible interpretations, in CFA, given the absence of cross-loadings, rotation does not apply Brown & Moore (ud, p. 12).

5.4.4.4 Estimation

The estimation in CFA (and, in general in SEM) involves a *fitting function* that operates to minimise the difference between the sample and the hypothesised variance-covariance matrices Brown & Moore (ud, p. 14). According to (Bryant & Yarnold, 2004, p. 116), CFA uses mainly three methods of estimation: Maximum Likelihood (ML), Generalised Least Squares (GLS) method, and the Unweighted Least Squares (ULS) method.

The ULS method has the advantage of being invariate to the scale of the variables and the ML and GLS methods provide an overall chi-square statistic for testing the model fit (Bryant & Yarnold, 2004, p. 116).

The most commonly used estimation method, and also the default option in most factor analysis software, for both CFA and SEM is the maximum likelihood method (ML), which aims to maximise the likelihood of the parameters given the observed data. See more details of the ML method as applied to the estimation of factors in section 5.4.3.3. ML methods estimate the parameters in a CFA model by an iterative process that begins with starting or initial values, which are either specified or generated by the software. Through a number of iterations these starting values are refined in an endeavour to minimise the difference between the sample variance-covariance matrix and that of the model-implied matrix until convergence results in stability (Brown & Moore, ud, p. 15).

ML assumes a large sample size, indicators that have been measured on continuous scales (i.e., approximate interval-level data), and multivariate normal distribution of the indicators as well as linear combinations of the indicators(Brown & Moore, ud, p. 15). Violations of normal distribution can distort goodness-of-fit indexes by resulted biased standard errors that can invalidate the conclusions from the statistical tests. If non-normality is extreme (as can happen with the use of Likert items when a majority of respondents select the lowest choice such as 1 out of 5), the ML method will produce incorrect parameter estimates (Brown & Moore, ud; Bryant & Yarnold, 2004).

Citing Bentler, 1995, Brown & Moore (ud, p. 15) advise that in the case of non-normal but continuous indicators, is it better to use ML with robust standard errors and χ^2, which would provide the same parameter estimates as the standard ML, but the goodness-of-fit statistics such at the χ^2 and the standard errors would be corrected for non-normality. If one or more of the observed indicators are categorical or if non-normality is extreme, normal theory ML should be used where estimators such as a weighted and unweighted least squares methods should be considered (Bryant & Yarnold, 2004, p. 116, citing Muthen, 1993) and Brown & Moore (ud, p. 15).

5.4.4.5 Interpretation

Various output values are used in the interpretation of CFA, these include the standardised root mean square residuals, parameter estimates (factor loadings, error variances and covariances, and factor variances and covariances) as well as various measures and indexes of overall goodness of fit. For each estimated parameter in the model, CFA provides a probability level indicating the likelihood that the given parameter is different from zero. This can be used to decide which observed indicators can be eliminated from the model while retaining model reliability (Bryant & Yarnold, 2004, p. 111).

According to Brown & Moore (ud, p. 16) three main aspects of the results should be examined in considering the CFA model: a) overall goodness-of-fit; b) specific points of poor fit; and c) the interpretability, size and statistical significance of the parameter estimates.

Overall Goodness-Of-Fit Determining goodness-of-fit involves considering how well the model parameter estimates (factor loadings, factor correlations, error covariance) are able to reproduce the observed relationships.

> For example if the standardised loadings of a factor on two variables X_1 and X_2 were $\lambda_1 = 0.760$ and $\lambda_2 = 0.688$ and the standardised factor variance was given by $\phi_1 = 1$, the model-implied correlation of these indicators would be the product of their factor loading estimates and the factor variance. That is $Cov(X_1, X_2) = \lambda_1\phi_1\lambda_2 = (0.760)(1)(0.688) = 0.523$. Assuming the sample correlation of X_1 and $X_2$was 0.516, the model-implied estimate differs by only 0.007 standardised units Brown & Moore (ud, p. 17).

CFA determines an overall maximum likelihood χ^2 and an associated p-value, which describes the probability that the matrix of fitted residuals generated by the model is different from zero, hence when one considers the fit of an individual model a non-significant p-value suggests a model fit (Bryant & Yarnold, 2004, pp. 111-112).

In general factor analyses require large samples. Due to the nature of the χ^2 statistic and the definition of the hypothesis, small samples provide less power to estimate a model's true lack of fit and can therefore inflate the model's apparent goodness of fit Bryant & Yarnold (2004, p. 117). Citing Alwin & Jackson (1980), Bryant & Yarnold (2004, p. 113) warn that due the sensitivity of the χ^2 statistic to sample size, it may not be that useful to consider overall goodness of fit when large samples are used since "even reasonable models are likely to produce significant χ^2 values" and that the difference in χ^2 may be more informative. Citing Jöreskog (1978, p448), Bryant & Yarnold (2004, p. 113) give the following advice: If the χ^2 statistic is large compared to the degrees of freedom, the fit can be explored by considering the residuals, which may suggest an option of introducing more parameters, which could lead to a smaller χ^2 value. If however the drop is large compared to the difference in the degrees of freedom, the change may indicate an improved model. Should the drop in the χ^2 value be close to the difference in the numbers of degrees of freedom, the apparent improvement could be considered as "capitalising on chance", implying the added parameters may not be significant. A "useful heuristic" for comparing relative fit is to use a ratio of χ^2 divided by degrees of freedom and decreases approaching zero imply improvement of the fit.

In searching for an appropriate CFA model, one typically considers a range of alternative models; from very restrictive or *null* models with no latent factors and therefore no factor loadings and no factor variances or covariances, to unrestricted or *fully saturated* models in which all factor loadings and factor interrelationships are free to be estimated (Bryant & Yarnold, 2004, p. 115). When competing models are *nested*, i.e., when the model that is more restrictive can be obtained by imposing constraints on a more general, less restricted model, their chi-square test statistics can be directly compared by simply subtracting the chi-squared statistic for the more general model from that of the more restricted one, giving $\Delta\chi^2$. Similarly the difference in the degrees of freedom, Δdf, can be determined. Using the difference in χ^2 and the difference in the *df* in an ordinary chi-squared test for significance the models can be compared for best fit, with the model with the smaller chi-squared statistic considered a better fit (Bryant & Yarnold, 2004, pp. 110-112, citing Bentler and Bonner, 1980, and others).

Although χ^2 is effectively used in nested model comparisons it is seldom used as the only measure of fit since, as already described, it is very sensitive to sample size. It is however χ^2 is however utilised in the calculation of various other indexes, such as the standardised (or unstandarised) root mean square residual (SRMSR), root mean square-error of approximation (SMSEA), the Tucker-Lewis index (TLI) and the comparative fit index (CFI). Each of these should be considered since they provide different information about model fit. Considered together they provide a more conservative and reliable evaluation of goodness-of-fit (Brown & Moore, ud, pp. 17-18).

Citing Hu and Bentler (1999), Brown & Moore (ud, p. 18) suggest the following guidelines to acceptable model fit, although, citing Marsh, Hau & Wen (2004), they warn that some researchers consider these guidelines to be too conservative:

- SRMR values close to or below 0.08;
- RMSEA values close to or less than 0.06 ;
- CFI and TLI values close to or greater than 0.95

Specific Poor Fit The goodness-of-fit statistics and indices can only provide overall indications of model fit. Sometimes, especially in more complex models, while overall indicators may show acceptable model fit, some relationships are less than acceptable; or, alternatively overall goodness-of-fit is rejected, but the reasons for this rejection are not clear. Two statistics that are used to identify specific areas of misfit are *standardised residuals* and *modification indices (Brown & Moore, ud, p. 19).*

Standardising the residuals in order to consider them independent of their units of measurement and examining their distribution is one way of judging how well a CFA model fits the data. The difference between each of the predicted interrelationships and the actual observed interrelationships is referred to as a *fitted residual.* The goodness-of-fit of a particular model can be assessed by examining the overall size of the fitted residuals and the proximity of these residuals to zero (Bryant & Yarnold, 2004, pp. 110-111).

A standardised residual is analogous to standard scores in a sampling distribution and can be interpreted as z-scores. For example a standardised residual of 1.96 or more would indicate significant additional covariance between indicators that was not reproduced by the model's estimates.

The *modification index* (MI) represents the predicted decrease in the χ^2 value if a given parameter no longer constrained the model. A given parameter's MI can be evaluated by for example comparing it with 3.84, the critical value of a χ^2 distribution with one degree of freedom at $p < 0.05$. An MI value greater than this value would suggest overall fit can be significantly improved if the fixed or constrained parameter was freely estimated. However, since MI's are sensitive to sample size, such revisions should only be considered if they can be justified on empirical or conceptual grounds. Revising a model purely on the basis of large standardised residuals or MIs can result in further mis-specification and over-fittingBryant & Yarnold (2004); Brown & Moore (ud, pp. 111-114, citing Jöreskog and Sörbom (1989)).

Interpretability, Strength and Statistical Significance of Parameter Estimates The parameter estimates such as factor loadings and factor correlations should only be interpreted in the context of a fitting model solution since the parameter estimates of a poorly fitting model are likely to be biased. Assuming a good fit, the parameter estimates should first be evaluated to confirm that they make statistical as well as substantive sense. They should for example not take on values such as negative indicator variances, which could be indicative of model mis-specification, and they should be of a size and direction that is in line with the conceptual or empirical aspects of the model. For example very small estimates may indicate unnecessary parameters, while very large estimates may question the existence of distinct constructs (Brown & Moore, ud, p. 22).

Chapter 6

Framework for a Phd

6.1 Specific Context and Linkages with the MSc in Data Science

After establishing media repertoires as they pertain in 2016 in South Africa, questions still remain as to the trends and shifts in media repertoires as well as how the various groups use the different media types. Please refer to chapter 4 for a more detailed exposition of the contextual background where this research will be situated.

6.2 Aims and Objectives of the Research

To gain a better understanding of the trends in media repertoires over the past decade in South Africa.

To gain a deeper understanding of how these groupings, as defined by repertoires, use the various media types.

6.3 Methods

Replicating some of the factor analytic procedures developed and conducted in the masters thesis, again utilising AMPS data, but applied to at least two earlier periods with five-year intervals, resulting in media repertoires defined for 2006; 2001 and 2016. It may be possible to extend this work to include 2017, but utilising data from the AMPS successor surveys.

Applying various *multi-method* approaches using techniques involving a triangulating mixture of questionnaire-based surveys, focus groups and ethnological observation of participants' online practices, including methods described as *virtual shadowing* of users activities, as described in section 4.3 above.

Bibliography

Abdi, H. (u.d.). Factor Rotations in Factor Analyses. The University of Texas. `https://www.utdallas.edu/~herve/Abdi-rotations-pretty.pdf`, Accessed: 29 August 2016.

Afifi, A., May, S., & Clark, V. A. (2011). *Practical Multivariate Analysis*. CRC Press, 5th edition.

Anderson, C., Bell, E., & Shirky, C. (2012). *Post-Industrial Journalism: Adapting to the Present*. Technical report, Tow Center for Digital Journalism.

Anon (2007). *Toward Economic Sustainability of the Media in Developing Countries*. Technical report, Center for International Media Assistance.

Berndt, A. & Petzer, D., Eds. (2011). *Marketing Research*. Pearson Education South Africa.

Brown, T. A. & Moore, M. T. (u.d.). Confirmatory Factor Analysis. Research Gate. `https://www.researchgate.net/profile/Michael_Moore8/publication/251573889_Hoyle_CFA_Chapter_-_Final/links/0deec51f14d2070566000000.pdf`, Accessed: 2 September 2016.

Bryant, F. B. & Yarnold, P. R. (2004). *Reading and Understanding Multivariate Statistics*, chapter four, (pp. 99–136). American Psychological Association.

Carifio, J. & Perla, R. (2008). Resolving the 50-year debate around using and misusing likert scales. *Medical Education*, (42), 1150–1152.

Conrad, D. (2014). Deconstructing the community radio model: Applying practice to theory in east africa. *Journalism*, 15, 773–789.

Edgerly, S. (2015). Red Media, Blue Media, and Purple Media: News Repertoires in the Colorful Media Landscape. *Journal of Broadcasting & Electronic Media*, 59(1), 1–21.

Grimm, L. G. & Yarnold, P. R. (2004). *Reading and Understanding Multivariate Statistics*. American Psychological Association.

Johnson, R. A. & Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis*. Pearson Education International, 5th edition.

Kline, P. (1994). *An Easy Guide to Factor Analysis*. Routledge.

Morrison, D. F. (2005). *Multivariate Statistical Methods*. Thomson Learning Inc., 4th edition.

Peters, B. (2010). The future of journalism and challenges for media development. *Journalism Practice*, (4), 268–273.

Radloff, S. (2015). *Mathematical Statistics 3, General Linear Models: Course Notes*. Department of Statistics, Rhodes University.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of educational research*, 99(6), 323–338.

Schröder, K. (2015). News media old and new: Fluctuating audiences, news repertoires and locations of consumption. *Journalism Studies*, 16(1), 60–78.

Susman-Pena, T. (2012). *Making Media Development More Effective*. Special report, Center for International Media Assistance.

Turcotte, J., York, C., Irving, J., Scholl, R. M., & Pingree, R. J. (2015). News Recommendations from Social Media Opinion Leaders: Effects on Media Trust and Information Seeking. *Journal of Computer-Mediated Communication*, 20(5), 520–535.

Xu, J., Forman, C., Kim, J. B., & Van Ittersum, K. (2014). News Media Channels: Complements or Substitutes? Evidence from Mobile Phone Usage. *Journal of Marketing*, 78(4), 97–112.