Appendix 1
Supporting documentation

This page intentionally left blank

Appendix 1.1: Project resources and output

Software

The main software used in the production of this atlas were:

HealthWIZ – data analysis and mapping
Harvard Graphics – charting
Microsoft Excel for Windows – correlation analysis
Microsoft Word for Windows – word processing

Hardware

A variety of IBM compatible microcomputers were used in the production of the atlas. A HP Laser Jet 5000 Series printer was used for printing drafts of the text and maps.

Printing

The atlas was printed by Openbook Publishers, Adelaide. They were supplied with word processing documents containing the text, tables, graphs and the maps (the maps were pasted into frames in the document). The atlas was then electronically transferred to plates for offset printing, without the need for film or bromides.

Project output

Data in electronic and printed form

Separate atlases are available for each State and Territory and for Australia. For each atlas there is a companion volume comprising the data on which the maps are based: for Western Australia, it is Volume 6.1. Both of these can be purchased from Government Info Shops in the capital cities.

The text and maps can also be downloaded for reading and printing from the Public Health Information Development Unit World Wide Web site at www.publichealth.gov.au

In addition, the text, maps and data can be accessed electronically from a CD-ROM (for Windows). On the CD-ROM, the text is in documents in Microsoft Word format. The data are in spreadsheet files in Microsoft Excel format and include all of the data mapped in the atlas, in table format as presented in Volume 6.1. Some data are also available in the HealthWIZ database.

Additional analyses will be posted to the Public Health Information Development Unit web site from time to time.

HealthWIZ software

HealthWIZ is a comprehensive health statistics database product, with a small area focus, produced by the Commonwealth Department of Health and Aged Care. It is comprised of detailed, content-rich data collections from Australia's hospital systems, cause of death registries, Medicare and social security payment systems and population censuses, together with data from administrative systems such as aged care and child care.

The data are contained on a CD-ROM and are accompanied by high performance table-building software. The menu-driven interface allows for a range of statistical calculations (age-standardised rates, confidence intervals, indices, time series data) to be undertaken to choose the most appropriate for the dataset and the needs of the user. These calculations are built into the software. The HealthWIZ software is also accessible via the World Wide Web at www.prometheus.com.au

HealthWIZ Version 4.0 comes with an integrated high performance mapping module. All the datasets and variables in the database can be mapped without the need for specialist knowledge of mapping software. All necessary digitised boundaries are included for users to be able to copy the maps to their own documents for publication.

Selected data from the atlas will be available in HealthWIZ. This includes all of the deaths and income support payments data, as well most of the hospital data, although its inclusion is subject to approval from the States and Territories. Its inclusion in HealthWIZ will allow greater flexibility in mapping the variables in the atlas, as well as many more variables from the same and other topics. The Census data, as well as the remaining health status data (the disability and handicap predictions, Total Fertility Rate), cannot be incorporated at this stage because of restrictions imposed on its use by the Australian Bureau of Statistics.

This page intentionally left blank

Appendix 1.2: Geographic areas mapped

Introduction

The following notes are intended to amplify and explain points raised in Chapter 2, *Methods* as to the areas mapped in this atlas.

Areas

Background

The data variables in each chapter are mapped separately for **Perth** and for the whole State. The basic geographic area mapped for **Perth** and for the whole State is the Statistical Local Area (SLA): SLAs are described in Chapter 2. Maps have been produced in the HealthWIZ software using an approximation to Lambert's Conformal Conic Projection.

SLAs in Perth

The SLAs mapped for **Perth** and the Rest of State are shown in **Maps A1** and **A2** and listed in the accompanying tables. Copies of the boundaries to use as overlays with the maps in this volume are in a pocket inside the back cover.

Areas mapped in non-metropolitan areas

As noted, the data for non-metropolitan are mapped by SLA. SLAs which are predominantly urban centres (towns) have been separately identified and located on the maps as a circle. Many urban centres are not separate SLAs.

To increase the number and range of urban centres for which data could be published, an urban centre with a population of 7,500 or more was mapped separately where it comprised 75 per cent or more of the SLA in which it was located. This resulted in 14 of the 36 urban centres of this size in New South Wales being mapped (**Table A1**). In cases where the area of the SLA is larger than the area of the circle, the underlying SLA can be seen on the map: both are mapped in the same shade. Where the location of the circle in its correct geographic position would have hidden details of another SLA, the circle has been located off the map, with a line adjoining the circle and the correct geographic location. Similarly, areas on the map that are too small for variations in the shading to be seen have been enlarged and located off the map.

Table A1: Urban centres in Western Australia

Urban centre	Population		
	Urban centre	SLA	Urban centre as % of SLA
Mapped: urban centres comprising 75% or more of SLA			
Albany	14,590	14,590	100.0
Geraldton	19,812	19,816	100.0
Port Hedland	12,846	13,116	97.9
Mandurah	35,945	37,925	94.8
Kalgoorlie/Boulder	28,087	29,683	94.6
Bunbury	24,945	26,556	93.9
Not mapped: urban centres comprising less than 75% of SLA			
Esperance	8,647	11,837	73.1
Karratha	10,057	14,954	67.3
Busselton	10,642	17,490	60.8

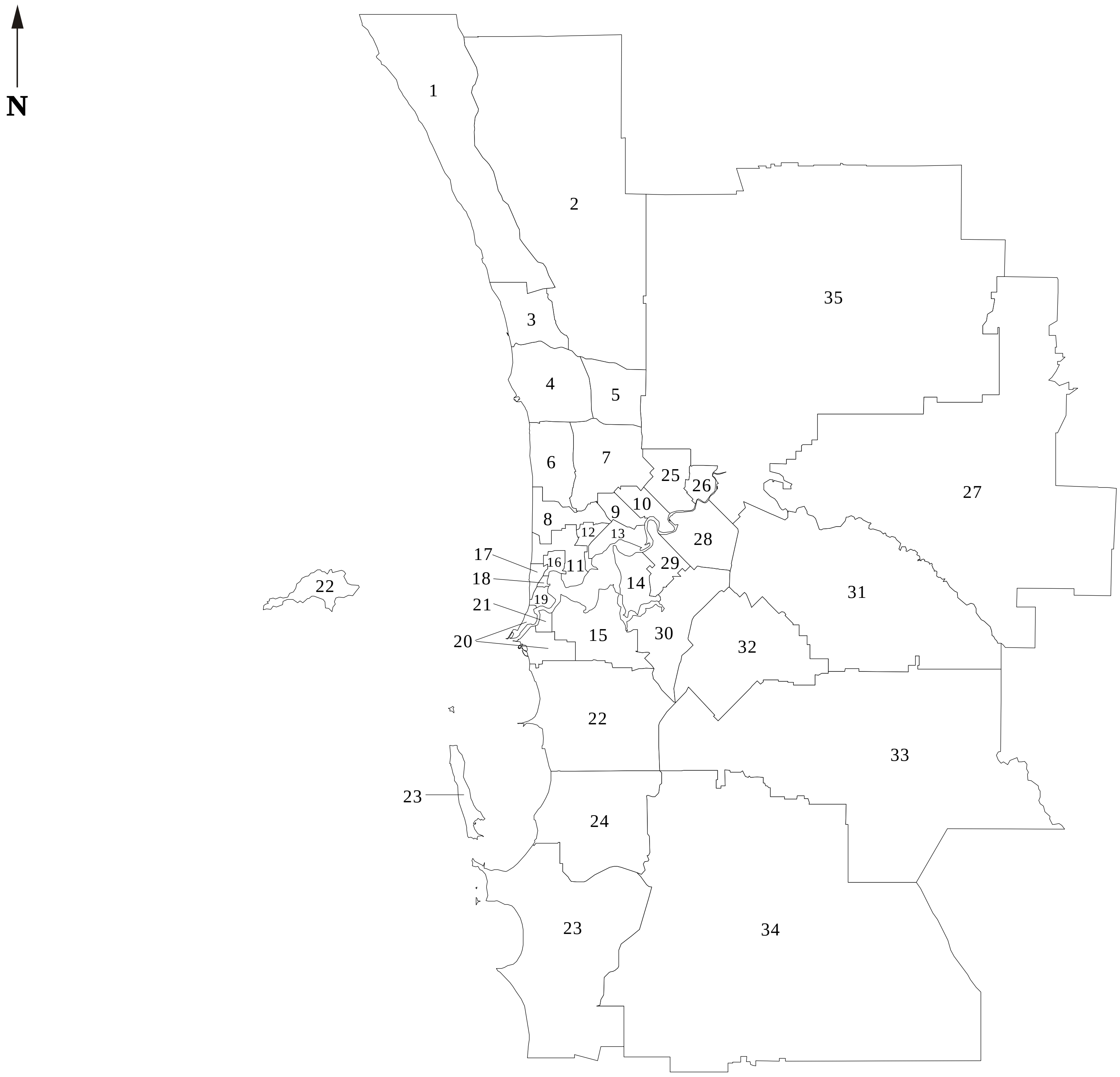
Source: Compiled from 1996 ABS Census data

Boundary changes

The boundaries of some SLAs have changed over the periods for which the data have been collected (periods varying from one year to four years). In some cases this requires that two or more areas be combined to enable the data to be mapped and compared, or for the correlation analysis to be undertaken. For example, boundary changes to the SLAs of Wanneroo (C) in 1996 meant that, for the 1996 Census data to be correlated with the data for deaths and hospital admissions, the SLAs of Wanneroo: Central-Coastal, North-East, North-West, South-East and South-West had to be combined. A list of the areas grouped and the name assigned to each is included in the beginning of the relevant chapter.

Map A1
Key to areas mapped for Perth[1]

(also included as a clear film overlay inside back cover flap)



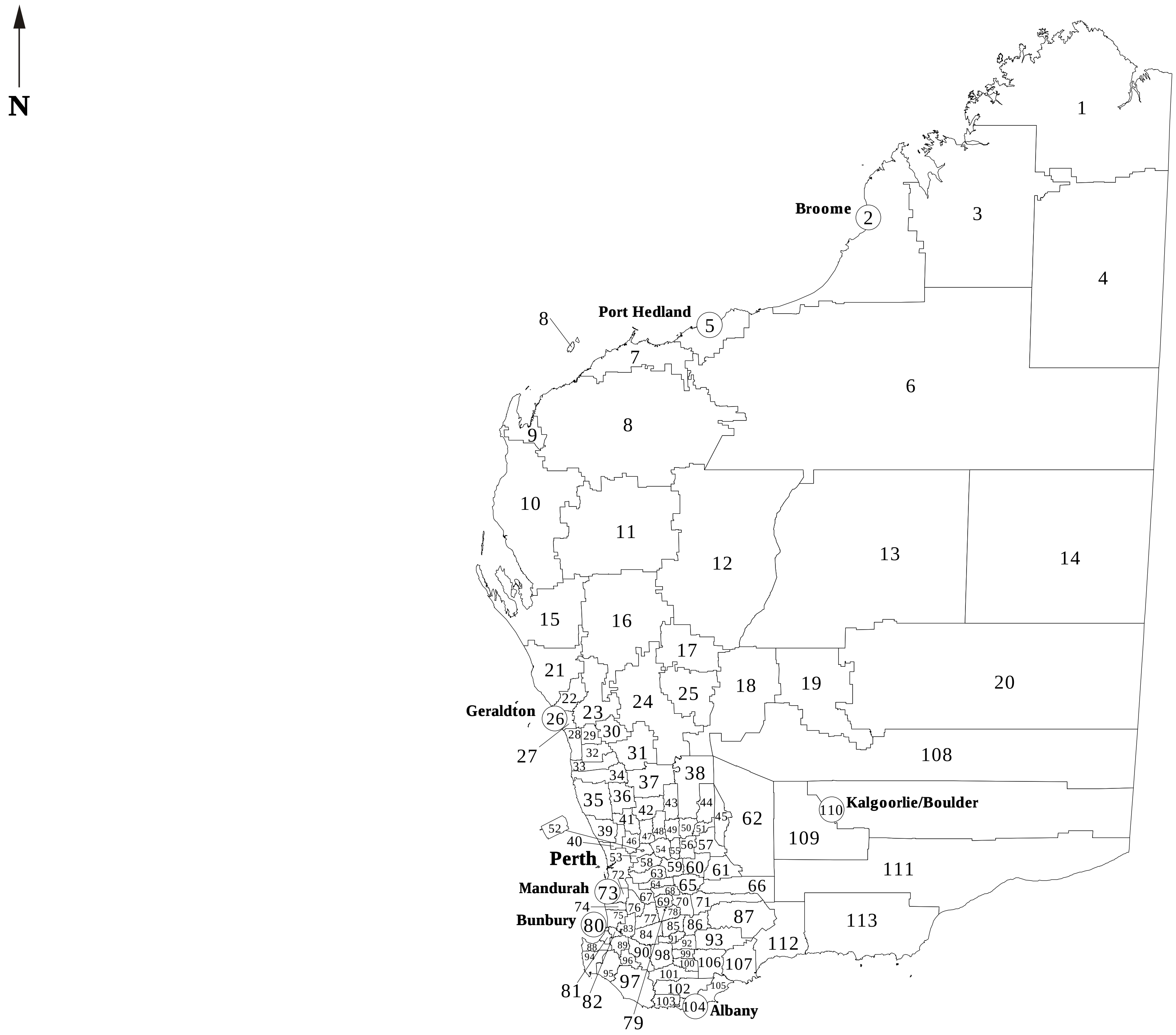
[1]*See footnotes to Table A2 for details of differences in boundaries for areas prior to 1996*

Table A2: Key to Statistical Local Areas in Perth, 1996

Statistical Local Area Name	Area number	SLA code	Statistical Local Area Name	Area number	SLA code
Armadale (C)	33	70	Peppermint Grove (S)	18	6930
Bassendean (T)	26	350	Perth (C)[1]	13	7080
Bayswater (C)	25	420	Rockingham (C)	23	7490
Belmont (C)	28	490	Serpentine-Jarrahdale (S)	34	7700
Cambridge (T)[1]	8	1310	South Perth (C)	14	7840
Canning (C)	30	1330	Stirling (C) - Central	7	7914
Claremont (T)	16	1750	Stirling (C) - Coastal	6	7915
Cockburn (C)	22	1820	Stirling (C) - South-eastern	10	7916
Cottesloe (T)	17	2170	Subiaco (C)	12	7980
East Fremantle (T)	21	3150	Swan (S)	35	8050
Fremantle (C)	20	3430	Victoria Park (T)[3]	29	8510
Gosnells (C)	32	3780	Vincent (T)[1]	9	8570
Kalamunda (S)	31	4200	Wanneroo (C) - Central Coastal[3]	3	8751
Kwinana (T)	24	4830	Wanneroo (C) - North-East[3]	2	8753
Melville (C)	15	5320	Wanneroo (C) - North-West[3]	1	8755
Mosman Park (T)	19	5740	Wanneroo (C) - South-East[3]	5	8757
Mundaring (S)	27	6090	Wanneroo (C) - South-West[3]	4	8758
Nedlands (C)	11	6580			

[1] For data sets prior to 1996 Cambridge (C), Perth (C) and Vincent (T) have been mapped as Cambridge/Perth/Vincent
[2] Victoria Park (T) is named Perth (C) South in Chapter 5 and Shepparton (T) in Chapter 6
[3] For data sets prior to 1996 Wanneroo (C) – Central Coastal, - North-East, - North-West, - South-East and – South-West have been mapped as Wanneroo (C)
Source: Compiled from project sources

Map A2
Key to areas mapped for Western Australia[1]
(also included as a clear film overlay inside back cover flap)



[1]See footnotes to Table A3 for details of differences in boundaries for areas prior to 1996

Details of map boundaries are in Appendix 1.2

Table A3: Key to Statistical Local Areas in non–metropolitan areas of Western Australia, 1996

SLA name	Area no.	SLA code	ARIA Index	SLA name	Area no.	SLA code	ARIA Index
Albany (T)	104	70	2	Laverton (S)	20	4970	5
Albany (S)	105	140	3	Leonora (S)	19	5040	5
Ashburton (S)	8	250	5	Mandurah (C)	73	5110	1
Augusta-Margaret River (S)	94	280	2	Manjimup (S)	97	5180	3
Beverley (S)	63	560	2	Meekatharra (S)	12	5250	5
Boddington (S)	76	630	2	Menzies (S)	108	5390	5
Boyup Brook (S)	90	770	2	Merredin (S)	57	5460	3
Bridgetown-Greenbushes (S)	96	840	2	Mingenew (S)	29	5530	3
Brookton (S)	64	910	2	Moora (S)	36	5600	3
Broome (S)	2	980	5	Morawa (S)	30	5670	3
Broomehill (S)	99	1050	3	Mount Magnet (S)	25	5810	5
Bruce Rock (S)	60	1120	3	Mount Marshall (S)	38	5880	4
Bunbury (C)	80	1190	1	Mukinbudin (S)	44	5950	4
Busselton (S)	88	1260	2	Mullewa (S)	23	6020	3
Capel (S)	81	1400	1	Murchison (S)	16	6160	5
Carnamah (S)	33	1470	3	Murray (S)	72	6230	1
Carnarvon (S)	10	1540	5	Nannup (S)	95	6300	2
Chapman Valley (S)	22	1610	3	Narembeen (S)	61	6370	4
Chittering (S)	40	1680	2	Narrogin (T)	79	6440	3
Collie (S)	83	1890	2	Narrogin (S)	78	6510	3
Coolgardie (S)	109	1960	3	Ngaanyatjarraku (S)[1]	14	6620	5
Coorow (S)	34	2030	3	Northam (T)	52	6650	1
Corrigin (S)	65	2100	3	Northam (S)	53	6720	1
Cranbrook (S)	101	2240	3	Northampton (S)	21	6790	4
Cuballing (S)	69	2310	3	Nungarin (S)	51	6860	4
Cue (S)	17	2380	5	Perenjori (S)	31	7000	4
Cunderdin (S)	54	2450	2	Pingelly (S)	68	7140	3
Dalwallinu (S)	37	2520	3	Plantagenet (S)	102	7210	3
Dandaragan (S)	35	2590	3	Port Hedland (T)	5	7280	5
Dardanup (S)	82	2660	1	Quairading (S)	59	7350	2
Denmark (S)	103	2730	3	Ravensthorpe (S)	112	7420	4
Derby-west Kimberley (S)	3	2800	5	Roebourne (S)	7	7560	5
Donnybrook-Balingup (S)	89	2870	2	Sandstone (S)	18	7630	5
Dowerin (S)	48	2940	2	Shark Bay (S)	15	7770	5
Dumbleyung (S)	86	3010	3	Tambellup (S)	100	8120	3
Dundas (S)	111	3080	5	Tammin (S)	55	8190	3
East Pilbara (S)	6	3220	5	Three Springs (S)	32	8260	3
Esperance (S)	113	3290	4	Toodyay (S)	46	8330	1
Exmouth (S)	9	3360	5	Trayning (S)	50	8400	3
Geraldton (C)	26	3500	2	Upper Gascoyne (S)	11	8470	5
Gingin (S)	39	3570	2	Victoria Plains (S)	41	8510	2
Gnowangerup (S)	106	3640	3	Wagin (S)	85	8610	3
Goomalling (S)	47	3710	2	Wandering (S)	67	8680	2
Greenough (S)	27	3850	2	Waroona (S)	74	8820	1
Halls Creek (S)	4	3920	5	West Arthur (S)	84	8890	2
Harvey (S)	75	3990	1	Westonia (S)	45	9030	4
Irwin (S)	28	4060	3	Wickepin (S)	70	9100	3
Jerramungup (S)	107	4130	4	Williams (S)	77	9170	2
Kalgoorlie/boulder (C)	110	4280	5	Wiluna (S)[1]	13	9250	5
Katanning (S)	92	4340	3	Wongan-Ballidu (S)	42	9310	3
Kellerberrin (S)	56	4410	3	Woodanilling (S)	91	9380	3
Kent (S)	93	4480	4	Wyalkatchem (S)	49	9450	3
Kojonup (S)	98	4550	3	Wyndham-East Kimberley (S)	1	9520	5
Kondinin (S)	66	4620	4	Yalgoo (S)	24	9590	4
Koorda (S)	43	4690	3	Yilgarn (S)	62	9660	4
Kulin (S)	71	4760	4	York (S)	58	9730	1
Lake Grace (S)	87	4900	4				

[1]For data sets prior to 1996 Ngaanyatjarraku (S) and Wiluna (S) have been mapped as Ngaanyatjarraku/Wiluna
Source: Compiled from project sources

This page intentionally left blank

Appendix 1.3: Analysis and presentation of data

Data ranges settings

The selection of data ranges for the maps in this atlas took into account a variety of factors. These factors were:

- the data ranges used for other maps, particularly closely related maps;
- the number of areas in each range; and
- the 'balance' of the visual impact of the map.

Indirect standardisation

In comparing populations, for example the mortality of two populations, crude rates (eg. the number of deaths per 1,000 persons) may be misleading. Mortality, for example, depends strongly on age and sex. If the two areas have different age structures this variation alone may explain a difference in crude rates. The technique of standardisation is used to prevent variations in population structure from distorting differentials in events.

Indirect standardisation, used in this analysis, calculates the number of events (eg. services by GPs) which would theoretically occur if the rates for each age/sex group in a given population (the standard – in this case the population of Western Australia) were applied to the population of interest. The result is termed the 'expected' number of events. If the actual number of events is then divided by this expected number and expressed as a percentage, we obtain the standardised ratio, a figure which is independent of population age and sex structure.

Thus the standardised ratio for a particular area will show the percentage by which it differs from the experience found in the whole population. Taking an example, the Standardised Death Ratio for deaths of males in the SLA of Stirling: South-Eastern was 166: that is, there were 66 per cent more deaths of male residents of Stirling: South-Eastern aged from 15 to 64 years than would have been the case had the Western Australian rates applied in Stirling: South-Eastern. In other words, the ratio was substantially above the State average.

The data for persons (ie. the total of females and males) have been standardised for both age and sex. That is, standardised ratios have been produced using separate details of the number of males and females in each age group. This eliminates distortion of the data which may occur where the illness or death experience of males and females is different (eg. as in the case for circulatory system disease among the population under 65 years of age). The ages used for all but the deaths data were generally each five year age group from 0 to 4 years to 80 to 84 years, and 85 years and over. For the deaths data, the ages were the five year age groups for the population aged from 15 to 64 years for all but accidents, poisonings and violence (where a separate analysis was undertaken for 15 to 24 year olds) and infant deaths. In the case of infant deaths (deaths of children under 12 months of age), the Infant Death Rate was calculated; the Infant Death Rate is the number infant deaths per 1,000 live births. Standardised ratios (SRs) were not calculated for areas where fewer than five events (deaths, admissions, etc.) were expected from the Australian rates, because of the doubtful reliability for such small numbers. All cases were, however, retained in the analysis for the calculation of capital city and State/Territory totals and ratios.

In some areas, however, high ratios are due to the relatively high proportion of Aboriginal and/or Torres Strait Islander people. This occurs because, in the methodology used, a standard population with a fixed age structure is introduced. The mortality or morbidity, etc., for a particular population (eg. people in an SLA) is then adjusted to allow for discrepancies in age structure between the standard and the particular population. When the particular population includes a sub group with a substantially different age structure and health experience (for example, mortality experience) the process is distorted. Indigenous people represent such a population. They have a substantially lower life expectancy than the total population, are a much younger population, have higher age-specific death rates at all ages and their average age at death is lower. However, since data relating to Indigenous people is not adequately identified in, for example, death or hospital statistics, they cannot be analysed as a discrete group.

The high SRs for some data for areas with a relatively large proportion of Indigenous people therefore reflect, in part, that the data have not been effectively standardised. This does not invalidate the data for these areas – on the contrary, it highlights the inequity evident in the health of Indigenous people, and the urgent need to address this inequity, as well as the need to identify Indigenous people more accurately in the statistics.

It should be noted that SRs derived for each area by this indirect method are comparable only by relation to the standard population (the State population) and not directly with each other.

For variables presented as SRs the text and tables include details of whether the ratios were statistically significant ie. that they differed significantly from the standard. Whether an SR for an area differs significantly from the standard depends not only on the size of the ratio but also on the population size of the area and the overall rate for the particular event (eg. a cause of death, use of a general medical practitioner), both of which contribute to the 'expected' number of cases in an area. The same SR value in two areas which differ greatly in population size may be significantly different from the standard in the area with the larger population, but not so in the area with the smaller population.

Data sources

Table A4 shows data sources in addition to those noted in the footnotes to the tables in the earlier chapters. Further details of the HealthWIZ software (referenced in the table) are on page 373.

Table A4: Data sources

Chapter	Data sources
Chapter 4	
Tables	
4.2 to 4.11	Data for **1989** from *A Social Health Atlas of Australia 1992*. Data for **1996** is at 30 June and was compiled in HealthWIZ from data supplied by the DFACS (for all variables), DVA (Service Pension (Age) and Service Pension (Permanently Incapacitated)) and ATSIC (Community Development Employment Program data, at 30 June 1998).
Maps	As for Tables, above
Chapter 5	
Tables	
5.4 to 5.7	Compiled in HealthWIZ from data supplied by the ABS.
5.8 to 5.9	Data for **1988** from *A Social Health Atlas of Australia 1992*. Data for **1993** was compiled in HealthWIZ from data supplied by the ABS.
5.11 to 5.33	Data for **1985** to **1989** from *A Social Health Atlas of Australia 1992*. Data for **1992** to **1995** was compiled in HealthWIZ from data supplied by the Registrars of Deaths.
5.34 and 5.35	Compiled in HealthWIZ from data supplied by the ABS.
Figures	
5.3 to 5.7, 5.10	See note for Tables, above
Maps	As for Tables, above
Chapter 6	
Tables	
6.3, 6.5	With the exception of data for Queensland, data was compiled in HealthWIZ from data supplied by the AIHW from the National Hospital Morbidity Database: this database comprises data supplied to the AIHW by the State and Territory health authorities. Data for SLAs in Queensland were not available from the AIHW database and were obtained directly from the Queensland Health Department. The data was supplemented with details of the postcode or SLA of patients admitted to hospital in a State/Territory other than the State/Territory of their usual residence: these details were obtained from the individual State/Territory health authorities.
6.4	Data for **1989** (1989/90 for New South Wales) is from *A Social Health Atlas of Australia 1992*. With the exception of the data for same day patients which was from *NSW Inpatient Statistics Data Book 1989-90* for NSW and for South Australia was supplied by the Department of Human Services. Data for **1995/96**: see notes re Table 6.3, above, other than for data for same day patients which was supplied by the NSW Health Department and the South Australian Department of Human Services.
6.6, 6.7, 6.12 to 6.15, 6.18 to 6.23, 6.28 to 6.39, 6.55 to 6.56	Data for **1989** is from *A Social Health Atlas of Australia 1992*. Data for **1995/96**: see notes re Table 6.3, above.
6.8 to 6.11, 6.16 and 6.17, 6.24 to 6.27, 6.42 to 6.54, 6.57 to 6.61	Data for **1995/96**: see notes re Table 6.3, above.
6.63 to 6.66	Data for **1989** from *A Social Health Atlas of Australia 1992*. Data for **1996** was compiled in HealthWIZ from Medicare statistics supplied by DHAC.
6.67 and 6.68	Data was compiled in HealthWIZ from immunisation rates supplied from the Australian Childhood Immunisation Register by the National Centre for Immunisation Research and Surveillance of Vaccine at the New Children's Hospital, Westmead, New South Wales.
Figures	
6.1 to 6.10	See note for Table 6.3, above
Maps	As for Tables, above
Chapter 7	
Tables	
7.3 and 7.4	Data for **1990/91** from *A Social Health Atlas of Australia 1992*. Data for **1996/97** was compiled in HealthWIZ from Medicare statistics supplied by DHAC.
7.5 to 7.8	Data for **1989** from *A Social Health Atlas of Australia 1992*. Data for **1995/96** (public acute hospitals) and **1997** (private hospitals) was compiled in HealthWIZ from data supplied by DHAC.
7.2 and 7.9 to 7.12	Data for **1992** from *A Social Health Atlas of Australia 1992*. Data for **1997** was compiled in HealthWIZ from data supplied by DHAC.
Maps	As for Tables, above

Note: Details of abbreviations used in the table are ABS, Australian Bureau of Statistics; ATSIC, Aboriginal and Torres Strait Islander Commission; DFACS, Department of Family and Community Services; DHAC, Department of Health and Aged Care; DVA, Department of Veterans' Affairs.

Appendix 1.4: Classification of deaths, admissions and procedures

Codes used

Causes of death are classified by the Australian Bureau of Statistics to the Ninth (1975) Revision of the World Health Organisation's International Classification of Diseases (ICD-9) which was adopted for world-wide use from 1979. The codes used for the variables mapped in Chapter 5 are listed in **Table A5**.

Diagnoses and procedures mapped in Chapter 6 are classified according to the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM October 1988 Revision). External causes are classified according to ICD-9-CM Supplementary Classification of External Causes of Injury and Poisoning ('E' codes) classification codes. The codes used for the variables mapped in Chapter 6 are listed in **Table A6** and **A7**.

Table A5: ICD–9 Codes for causes of death mapped in Chapter 5

Cause of death	ICD code
All cancers [malignant neoplasms]	140-208
Lung cancer	162
Circulatory system diseases	390-459
Respiratory system diseases	460-519
Accidents, poisonings and violence	E800-E999

Table A6: ICD–9 Codes for diagnoses/external causes mapped in Chapter 6

Diagnoses /External cause	ICD code
Infectious and parasitic diseases	001-139
Cancers [malignant neoplasms]	140-208
Lung	162
Female breast	174
Psychiatric conditions	290-319
Psychoses	290-299
Neurotic, personality and other disorders	300-316
Circulatory system diseases	390-459
Ischaemic heart disease	410-414
Respiratory system diseases	460-519
Bronchitis, emphysema, asthma	490-493
Accidents, poisonings and violence	E800-E999

Table A7: ICPM Codes for surgical procedures mapped in Chapter 6

Principal procedures	Codes
All procedures	010-169; 180-695; 704-789; 792-793; 795-796; 798-869
Tonsillectomy and/or adenoidectomy	28.2, 28.3
Myringotomy [limited to 0-9 year olds]	20.01
Hysterectomy [limited females aged 30 years and over]	68.3-68.7
Caesarean section [limited to females aged 15 to 44 years]	74.0, 74.1, 74.2, 74.4; 74.99
Hip replacement	81.51, 81.53
Lens insertion	13.1, 13.2, 13.3, 13.4, 13.5, 13.6, 13.7
Endoscopies	42.23, 42.24, 44.13, 44.14, 45.13, 45.14, 45.16, 45.23-45.25

This page left intentionally blank

Appendix 1.5: Synthetic estimation for small areas

Staff of the Adelaide office of the Australian Bureau of Statistics (ABS) produced the synthetic predictions discussed and mapped in Chapter 5 as a consultancy for the Public Health Information Development Unit. The following paper prepared by the ABS describes the techniques used in production of the estimates.

Introduction

Statistics for small geographic regions are generally available only through administrative sources or the population census. Although household surveys contain much data of value, they provide estimates at a broad geographic level, usually the State or Territory level or, for some of the more populous States, for large regions. Estimates are rarely available for small areas such as the Statistical Local Area (SLA) mapped in this atlas.

Estimates produced from sample surveys are subject to two types of error: non-sampling errors which arise from errors in collecting, recording and processing the data; and sampling errors which arise because a sample, rather than the entire population, is surveyed. The sampling error tends to increase as the sample size decreases. Thus estimates produced from small samples can be subject to such high sample errors as to make them too unreliable for most practical purposes. Since household surveys typically have a small sample from large regions, it is not possible to provide direct survey estimates of suitable reliability for small regions.

Through the use of synthetic estimation techniques it is possible to produce reliable region level statistics (Marker 1999). The method of synthetic estimation was applied in predicting, at the SLA level, two characteristics from the 1995 National Health Survey (NHS):

- the number of people who had a self-assessed poor or fair health status; and
- the Physical Component Summary score from the SF-36 component of the NHS (see page 109 for details of this measure).

Predictions are also provided in this atlas of the number of people with a handicap; these estimates were produced by the ABS using a similar technique as part of another project. This technical note concentrates on the prediction of the former characteristics.

Background

Synthetic estimation predicts a value for a small geographic region based on modelled survey data and known characteristics of the region. A synthetic prediction can be interpreted as the expected value, for the variable of interest, for a 'typical' area with those characteristics. The SLA was the regional level of interest for this project (in the Australian Capital Territory and, in some cases in Queensland and the Northern Territory, SLAs were grouped; details of these groupings are contained in the relevant State and Territory atlases).

The model used for predicting small region data is determined by analysing data at a higher geographic level, in this case Australia. The relationship observed at the higher level between the characteristic of interest and predictor variables is assumed to also hold at the lower level. The predictions are made by applying the model to the small region counts of the predictors. This modelling technique can be considered as a sophisticated pro-rating of Australian level characteristic of interest across the regions in accordance with the joint distributions across the regions of the predictors.

The process of producing the predictions consists of four parts:

- preparation of data;
- model fitting;
- synthetic prediction; and
- assessing the prediction.

Data

As noted above, the two characteristics predicted were self-assessed health status and the Physical Component Summary score, both from the 1995 NHS. Self-assessed health status is provided by respondents to the survey indicating their assessment of the health status, on a scale of 'Excellent', 'Very Good', 'Good', 'Fair' or 'Poor'. The variables of interest here were those of people reporting their health as being 'Fair' or 'Poor'. The Physical Component Summary score is calculated from responses to the SF-36 component of the NHS. It is derived from a subset of items that ask respondents to the NHS aged 18 years and over, about their general physical health and wellbeing. A higher score indicates a better state of physical health and wellbeing.

Predictor data must satisfy the following criteria. It must be

- well related to the characteristic of interest;
- available from the NHS;
- available for similar time periods, both date and length of period; and be
- available at a similar geographic level, both Australia and SLA.

Sources of predictor data utilised were:

- the 1995 NHS;
- the 1996 Census of Population and Housing;
- administrative data from the Department of Family and Community Services;
- hospital separations data; and
- unreferred attendances with general medical practitioners (GPs).

One of the most important data related tasks was to identify predictors from these potential sources which satisfy the above criteria. Data considered included variables such as:

- age;
- sex;
- employment status;
- currently a student;
- income;
- receiving a Disability Support Pension;
- receiving Sickness Allowance;
- receiving the Age Pension;
- Socio-Economic Indexes for Areas derived from the Census;

- whether an inpatient at a hospital; and
- whether consulted with a GP in the two weeks prior to interview.

Many of the available variables common with the NHS differed by definition, collection methodology, reference period and geography. In such instances, appropriate adjustments were made using information obtained by comparing counts, proportions and distributions of the common variables. For example, the income variable was available to the nearest dollar from the NHS, but was available from the Census by income range only. This required the NHS income data to be classified to similar ranges. A comparison of the counts and distributions of persons across the income ranges indicated that income data from the NHS and Census were closely aligned and for the purposes of prediction could be considered well aligned. Several variables also required conversion of their geography from postcode to SLA using the 1994 Australian Standard Geographical Classification (ABS 1994).

There was, however, a fair degree of commonality in the datasets, with the NHS taken over the 1995 year, the hospital inpatient data being for 1995-96, pensioner and beneficiary data being at 30 June 1996 and the Population Census at 4 August 1996.

Model fitting

Once data preparation was completed the relationship between the characteristic of interest and the predictor variables was modelled using data from the NHS at the Australian level. The self-assessed health status and Physical Component Summary score were modelled independently.

The model applied took the linear form:

$$Y = p_o + p_1X_1 + p_2X_2 + p_3X_3 + ... + p_jX_j$$

where
Y is the characteristic of interest

X_i are the predictor variables

p_i are the coefficients which are produced from the modelling process.

In the case of the variable for self-assessed health status, the Y takes the value 1 if the individual's status was fair or poor and 0 otherwise. For the Physical Component Summary score, Y ranges in value from around 45 to 55.

The X_i predictors take the value 1 if the individual has the predictor characteristic (eg. has visited a GP in last two weeks) or 0 otherwise.

The coefficients, p_i, were estimated using the linear regression technique. An original subset of data items from the NHS were compiled that satisfied the specified criteria. The NHS data file, with the subset of data items, was randomly split into two halves with a regression model fitted to both data sets. Data items that were not important in predicting the variable of interest in either, or both, of the two models were removed. This process continued until a final linear model was obtained whereby all variables were significant ($p<0.05$) in the estimation of the response variable (characteristic of interest). Fitting the model to the split data produces a more robust final model as it reduces the probability of including a variable with high variability.

The final form of the model was then fitted to the full data set to produce regression coefficients and diagnostics which were examined using Cook's D statistic (Cook 1979) to identify any individual respondent who had undue influence on the final parameter estimates. Any 'outliers' identified were removed from the data and the model refitted.

Below is a list of variables that were included in the final models.

Self-assessed health status:
- State/Territory of usual residence;
- age (in 10 year age groups);
- sex;
- employed;
- employed (aged 18 to 24 years);
- employed (aged 25 to 34 years);
- admitted to hospital for at least one night in the last two weeks;
- consulted a general medical practitioner in the last two weeks;
- receives Disability Support Pension;
- receives Disability Support Pension (aged 18 to 24 years);
- receives Sickness Allowance;
- receives Age Pension;
- SEIFA Index of Relative Socio-Economic Disadvantage.

Physical Component Summary score:
- State/Territory of usual residence;
- age (in 10 year age groups);
- income (gross personal annual income);
- studying (currently studying full or part-time at college, university, etc.);
- employed;
- admitted to hospital for at least one night in the last two weeks;
- consulted a general medical practitioner in the last two weeks;
- receives Disability Support Pension;
- receives Disability Support Pension (aged 18 to 24 years);
- receives Sickness Allowance;
- receives Age Pension;
- SEIFA Index of Relative Socio-Economic Disadvantage.

Synthetic prediction

The prediction for an SLA was derived from the linear combination, specified by the regression coefficients, of the counts of individuals within the SLA with the predictor characteristics.

Note that for the Physical Component Summary score the predicted value for the SLA was scaled to a person level score by dividing the prediction by the number of people aged 18 years and over. The final prediction can therefore be considered as a mean score for people living in the SLA.

The predictions of poor or fair health status give an indication of the number of persons aged 18 years and over who would assess their health as poor or fair.

The predictions were age-sex standardised to remove variations between SLAs solely related to variations in age and sex.

Assessing the predictions

The models were assessed in terms of how well they predicted for individuals, SLA and larger regions (Statistical Divisions and Sub-Divisions). This involved comparing predicted values against values determined directly from the NHS. For individuals, this was the reported value, while for SLA and larger regions it was the direct survey estimate. The comparisons were made by examining plots of the predictions against the NHS reported values and estimates. The plots were checked to ensure that there was a reasonable relationship between the predictions and NHS results.

The 95% confidence intervals were calculated for the direct survey estimates and compared to the predictions. If the majority of predictions fall within the confidence intervals then there is a high level of confidence that the predictions are reliable.

This page left intentionally blank

Appendix 1.6: Additional details of cluster analysis

Introduction

Some of the descriptions of the cluster analyses were more lengthy and technical than others. Where they were considered to be too detailed and/or technical, a shortened version is shown in Chapter 8 and the full version is shown below. Those included are the socioeconomic status clusters in **Perth**; the health service utilisation clusters in non-metropolitan Western Australia; and all of the analyses for towns.

Socioeconomic status clusters in Perth

The dendogram and agglomeration schedule suggested a three or four cluster solution. Both solutions were examined. The three cluster solution lacked definition between the Low and Medium clusters, which became the Low and Very Low clusters in the four cluster solution. Despite this lack of clarity the solution performed very well against the IRSD.

A factor analysis was run using Principal Components extraction and Varimax rotation. This produced a two factor solution.

The factor drivers of this solution (single parent families, unskilled and semi-skilled workers, unemployed people, female labour force participation, dwellings rented from the State housing authority and early school leavers) were entered into a cluster analysis. This produced a two cluster solution, which was disappointing. The three cluster solution from this analysis was also examined, and found to be of high quality, with fairly clean discrimination between factors, and very good performance against the IRSD.

The factor drivers of the first factor (single parent families, unemployed people, dwellings rented from the State housing authority, low income families and dwellings without a motor vehicle) were entered into a cluster analysis which produced an extremely clean three factor solution, with every input variable consistently ranked High, Medium or Low in the respective high, Medium and Low clusters. Unfortunately, this solution did not compare well to the IRSD, leading to confusion as to whether the clusters represented socioeconomic groups or something else.

A factor analysis was performed using Maximum Likelihood extraction and Oblimin rotation. This produced a two factor solution.

The factor drivers of the solution (unskilled and semi-skilled workers, early school leavers, the indigenous population, unemployed people, low income families and single parent families) were entered into a cluster analysis, which produced a two cluster solution. Again the three cluster solution was also examined and found to be of high quality.

The factor drivers of the first factor (unskilled and semi-skilled workers, early school leavers, the indigenous population and female labour force participation) were also entered into a cluster analysis, which again produced a two cluster solution. Again the three cluster solution was examined and found to be of high quality, but it was not as good as the solution based on the factor drivers of the Maximum Likelihood solution.

This string of analyses had uncovered four good quality solutions, three of which were really two factor solutions, with doubts about the other (cleanest solution) in terms of producing socioeconomic groupings.

The solution based on factor drivers for the Principal Components extraction, Varimax rotation factor solution was accepted. The main reasons for this were that it grouped the two diffuse clusters of the original four cluster solution using all input variables, which in itself was almost acceptable but for the fact that the two diffuse clusters did not amalgamate to form the three cluster solution in the same analysis. Also this solution performed as well as or better than any other solution against the IRSD (see **Table 8.4** and **Map 8.1**).

Although, as noted above, theoretically there is insufficient data to justify the model, the solution is so good it should be accepted (ie. the end justifies the means). This is supported by a comparison with the IRSD. This comparison showed that, of the 13 SLAs with the lowest IRSD scores in **Perth**, 11 were classified to the Low socioeconomic status group in this analysis; and that all of the SLAs with the highest scores for the IRSD were classified to the High socioeconomic status group.

Health service utilisation clusters of non-metropolitan SLAs

The variables for admissions of lung cancer, breast cancer, myringotomy, hysterectomy and hip replacement were excluded from the analysis because over 5 per cent of areas had no cases. Thus there were 24 variables to analyse 112 records. This was not quite enough data. Alternative strategies were tried in an attempt to produce a sensible solution.

A cluster analysis was tried to see if it gave a sensible solution despite the lack of data. This produced a three cluster solution of poor quality.

An exploratory factor analysis was run on the data using Maximum Likelihood extraction and oblique (oblimin) rotation. The analysis failed to converge at iteration 10.

An exploratory factor analysis was run on the data using Principal Component extraction and orthogonal (varimax) rotation. The analysis produced a seven factor solution. It should be noted that there was not quite enough data to sustain a factor analysis either.

Factor scales saved in the above analysis were used as input to a cluster analysis. This approach assumes the factor structure is accurate for the SLA data. This analysis resulted in a 3 cluster solution of dubious merit.

In an effort to produce a better solution, the drivers of the factor solution were selected for entry into a cluster analysis. The first four drivers of the first factor (total admissions, admissions of females, admissions to a public hospital and admissions for respiratory system diseases), the first three drivers of the second factor (same day admissions and admissions for same day procedures), the first two drivers of the third and fourth factors (admissions for circulatory system diseases and ischaemic heart disease and GP services to males and females), and the first drivers of the last three factors (admissions for psychosis, Caesarean section and lens insertion) were chosen.

This analysis again produced a three cluster solution of poor quality.

The drivers of the first factor (total admissions, admissions of males and females, admissions to a public hospital, and admissions for respiratory system diseases, infectious diseases and accidents, poisonings and violence) were entered into a cluster analysis. The solution contained three clusters but again was of poor quality.

Examination of previous solutions suggested that some variables are common to many solutions. The variables total admissions, admissions of males, admissions to a public hospital, and admissions for neurotic, personality or other mental disorder, infectious diseases and accidents, poisonings and violence were entered into a cluster analysis.

This analysis again produced a three cluster solution of poor quality. Although the low service use cluster was reasonably clearly defined, the high service use cluster was very diffuse (largely mixed with the medium cluster).

Running out of options to get a good solution, the factor drivers of the second factor extracted (same day admissions, admissions for same day procedures and admissions for endoscopy) were entered into a cluster analysis. This combination of variables produced a reasonably clean three factor solution. The SLAs in each cluster are listed in **Table 8.5** and shown in **Map 8.7**.

There was moderate agreement with the IRSD: of the lowest 41 SLAs for SEIFA, 14 (34.1 per cent) were in the High health service use cluster and, of the highest 10, none were in the Low health service use cluster.

Socioeconomic status clusters of towns

A cluster analysis was undertaken for the 55 towns (urban centres) across Australia that had populations of 7,500 or more at the 1996 Census and were identifiable in the non-Census datasets (see Appendix 1.2 for further details). These 55 records are theoretically sufficient to carry out a cluster analysis with nine input variables. A cluster analysis was performed on the available data, and the solution examined before attempting more complicated techniques to find a solution. This attempting provided a three cluster solution of fair to average quality. It did not discriminate particularly well between clusters, and the High socioeconomic cluster did not perform particularly well against the IRSD.

The 55 records also provided enough information for an exploratory factor analysis, since this analysis has the same data requirements as the previous model. A factor analysis was attempted using principal components extraction and varimax rotation, and a reasonable three factor solution was produced by this analysis, although it did not discriminate particularly well on the input variables between clusters.

The two main drivers of each factor were entered into a cluster analysis. The analysis excluded dwellings with no vehicles, single parent families and female labour force participation. This produced a three cluster solution which performed well against the IRSD, but again did not discriminate particularly well on the input variables between clusters.

The drivers of the first factor (low income families, unemployed people, female labour force participation and dwellings with no motor vehicle) were entered into a cluster analysis. This produced a four factor solution of poor quality.

A second exploratory factor analysis was tried using all nine input variables, but this time using maximum likelihood extraction, and oblimin (oblique, ie. not orthogonal) rotation. This analysis gave a three factor solution with the same factors (although in a different order, and the variables were in a different order of importance to the solution). The two main drivers of each factor were entered into a cluster analysis. The analysis excluded dwellings rented from the State/Territory housing authority, single parent families and female labour force participation. This analysis produced a four factor solution of good quality, although again the solution did not discriminate between clusters.

The drivers of the first factor of the oblique factor solution (dwellings rented from the State housing authority, indigenous people and single parent families) were entered into a cluster analysis. This analysis produced a three factor solution (with Broome ungrouped) which was of only fair quality.

The best solution was felt to be the four cluster solution produced from the first two factor drivers of each factor of the oblique factor solution (ie. based on low income families, unemployed people, early school leavers, unskilled and semi-skilled workers, indigenous people and single parent families). This analysis produced a solution of acceptable quality, which is reproduced in **Table 8.7**.

The ABS Index of Relative Socio-Economic Disadvantage (IRSD) was also available for the specified towns, but was withheld from the analysis and used as an independent check on the solution. It was found that, of the bottom 17 towns as classified by the IRSD, 16 (94.1 per cent) were classified to the Low socioeconomic group in this analysis. Further, of the top 20 towns under the IRSD, 15 (75.0 per cent) were classified to the High socioeconomic group.

Health status clusters of towns

There were 15 variables to analyse 55 records. This was not quite enough data. A cluster analysis of all the above variables was tried to see if it gave a sensible solution despite the lack of data. This produced a clear two cluster solution of good quality. The solution did not perform particularly well against the IRSD however, and a two cluster solution is not optimal.

Alternative strategies were tried in an attempt to produce a better solution. An exploratory factor analysis was run on the data using Principal Component extraction and orthogonal (varimax) rotation. The analysis produced a six factor solution. It should be noted that there was not enough data to sustain a factor analysis either.

The drivers of the factor solution were selected for entry into a cluster analysis. The first two drivers of the first two factors (deaths of 15 to 64 year old females, and deaths of 15 to 64 year olds from cancer, lung cancer and accidents, poisonings and violence) and the first drivers of the other four factors (people with a handicap, the Physical Component Summary score, infant deaths and the Total Fertility Rate) were chosen. They were entered into a cluster analysis, which produced a three cluster solution of good quality. Again the solution did not perform all that well against the IRSD.

The four drivers of the first factor (deaths of 15 to 64 year old females, deaths of 15 to 64 year olds from respiratory system diseases and accidents, poisonings and violence and years of potential life lost) were entered into a cluster analysis. This again produced a three factor solution which was very similar to the one produced based on the previous set of factor drivers (although slightly inferior to it).

The six factor scores saved from the above analysis were input into a cluster analysis. This produced a three cluster solution of good quality. The clusters were better spread than in other solutions, and the solution performed better against the IRSD than other solutions (**Table 8.7**).

The IRSD was again used as an independent check on the solution. It was found that, of the bottom 12 towns as classified by the IRSD, five (41.7 per cent) were classified to the Poor health status group in this analysis. Further, of the top 22 towns under the IRSD, 14 (63.6 per cent) were classified to the Good health status group.

Health service utilisation clusters of towns

There were 30 variables to analyse 55 records. This was not enough data. A cluster analysis of all the above variables was tried to see if it gave a sensible solution despite the lack of data. This produced a three cluster solution of reasonable quality.

Alternative strategies were tried in an attempt to produce a better solution. An exploratory factor analysis was run on the data using Principal Component extraction and orthogonal (varimax) rotation. The analysis produced an eight factor solution, but the varimax rotation failed to converge. Examination of the scree plot led to the conclusion that the factor analysis should only have six factors. This solution was forced, and the rotation then converged. It should be noted that there was not enough data to sustain a factor analysis either.

The drivers of the factor solution were selected for entry into a cluster analysis. The first two drivers of the first three factors (total admissions, same day admissions, admissions of females, same day admissions for a surgical procedure, and GP services for males and females) and the first drivers of the other three factors (admissions to a private hospital, and admissions for breast cancer and hip replacement) were chosen. They were entered into a cluster analysis, which produced a three cluster solution of reasonable quality (similar to the quality of the first solution examined).

The first nine drivers of the first factor (total admissions, admissions to a public hospital, admissions of males and females, and admissions for infectious diseases, respiratory system diseases and respiratory system diseases of children aged 0 to 4 years) were entered into a cluster analysis. The solution contained two clusters but was of a lower quality than the original solution.

The six factor scores saved from the above analysis were input into a cluster analysis. This produced a three cluster solution of good quality. The clusters were better spread than in other solutions, and the solution performed better against the IRSD than other solutions (**Table 8.7**).

A check with the IRSD showed that, of the bottom ten towns as classified by the IRSD, three (30.0 per cent) were classified to the High health service use group in this analysis. Further, of the top 26 towns under the IRSD, 13 (50.0 per cent) were classified to the Low health service use group.

Social health status clusters of towns

The cluster analysis technique has also been applied to a combination of the socioeconomic status and health status data sets. Data considered for inclusion were the variables in the final models for towns used to examine socioeconomic status and health status.

There were 24 variables to analyse 55 records. This was clearly not enough data. A cluster analysis of all the above variables was tried to see if it gave a sensible solution despite the lack of data. This produced a three cluster solution of fair to average quality. The solution did not perform at all well against the IRSD for the Low status group, and lacked definition between the Medium and Low status groups.

Alternative strategies were tried in an attempt to produce a better solution. An exploratory factor analysis was run on the data using Principal Component extraction and orthogonal (varimax) rotation. The analysis produced a six factor solution. It should be noted that there was not enough data to sustain a factor analysis either.

The drivers of the factor solution were selected for entry into a cluster analysis. The first three drivers of the first factor (deaths of 15 to 64 year old males, deaths of 15 to 64 year olds from accidents, poisonings and violence and years of potential life lost), the first two drivers of the second to fourth factors (single parent families, unskilled and semi-skilled workers, unemployed people, people with a handicap or disability and the Physical Component Summary score) and the first drivers of the last two factors (dwellings rented from the State housing authority and infant deaths) were chosen. They were entered into a cluster analysis, which produced a three cluster solution of only fair quality. Again the solution lacked discrimination between the middle and low status groups.

The eleven drivers of the first factor (the indigenous population, deaths of 15 to 64 year old males and females, deaths of 15 to 64 year olds from cancer, lung cancer, circulatory system diseases, respiratory system diseases and accidents, poisonings and violence, deaths of 15 to 24 year olds from accidents, poisonings and violence, years of potential life lost and Total Fertility Rate) were entered into a cluster analysis. This again produced a three factor solution which was of very similar quality to the original one based on all input variables (although slightly superior to it).

The six factor scores saved from the above analysis were input into a cluster analysis. This produced a four cluster solution of poor quality.

An exploratory factor analysis was run on the data using Maximum Likelihood extraction and oblique (oblimin) rotation. This produced a six factor solution.

The drivers of the factor solution were selected for entry into a cluster analysis. The first two drivers of the first four factors (dwellings rented from the State housing authority, people reporting fair or poor health, the Physical Component Summary score, people with a handicap or disability, deaths of 15 to 64 year old males and females and deaths of 15 to 64 year olds from cancer), and the first drivers of the last two factors (the indigenous population and single parent families) were chosen. They were entered into a cluster analysis, which produced a three cluster solution of only fair quality. Again the solution lacked discrimination between the Middle and Low status groups.

The eight drivers of the first factor (the indigenous population, deaths of 15 to 64 year old males, deaths of 15 to 64 year olds from cancer, lung cancer, circulatory system diseases, respiratory system diseases, accidents, poisonings and violence, deaths of 15 to 24 year olds from accidents, poisonings and violence, years of potential life lost and Total Fertility Rate) were entered into a cluster analysis. This again produced a three factor solution which was identical to the three cluster solution produced using the factor drivers of the first factor of the principal components extraction/varimax rotation factor analysis.

The six factor scores saved from the above analysis were input into a cluster analysis. This produced a three cluster solution of reasonable quality, with Charters Towers (C) not grouped. The clusters were better spread than in other solutions, and the solution performed better against the IRSD than other solutions. It is accepted since it was the best alternative found (**Table 8.7**).

Of the 17 lowest towns for the IRSD, nine (52.9 per cent) were classified to the Low social health status cluster; and of the top 14 towns for the IRSD, seven (50.0 per cent) were classified to the High social health status cluster.