1351.0.55.002

Australian Bureau of Statistics

Research Paper

Modelling Languages Other Than English Spoken in Australia Using Census Data

www.abs.gov.au

New Issue

Research Paper

Modelling Languages Other Than English Spoken in Australia Using Census Data

Lujuan Chen, Paul Romanis and Katie Palin

Analytical Services Branch

Australian Population Association
12th Biennial Conference

15–17 September 2004, Canberra, Australia

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) THU 31 MAR 2005

ABS Catalogue no. 1351.0.55.002
ISBN 0 642 47999 2

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper. For further information, please contact Lujuan Chen, Analytical Services Branch on Canberra (02) 6252 5917 or email <lujuan.chen@abs.gov.au>.

CONTENTS

MODELLING LANGUAGES OTHER THAN ENGLISH SPOKEN IN AUSTRALIA USING CENSUS DATA

Lujuan Chen, Paul Romanis and Katie Palin

Analytical Services Branch

EXECUTIVE SUMMARY

There has always been competition for space on the population census form. The Australian Bureau of Statistics (ABS) constantly needs to prioritise user requirements and find ways to make Census collection more cost effective and efficient, while presenting the least burden to respondents. Language information is an important part of the collection. However, the question on languages spoken by individuals at home takes up a good amount of space on the form and incurs significant coding costs.

Within this context, the ABS investigated whether the information, *Main Language Other than English Spoken at Home*, could be predicted every 10 years using other Census variables so that we could free up some space and funds to include other questions in every second Census. A mid range option was to ask whether a language other than English was spoken at home but not to ask what this was. In this way, we could free some amount of space and save the coding costs to meet increased user demands for questions like disability.

This paper presents the findings and methodologies engaged in this study. Although the ABS has since decided to include the language question in the 2006 Census, the methods and results are interesting and provide useful inputs to the formation of language questions in future Census of Population and Housing.

Objectives

The analysis aimed to investigate the option mentioned above. Specifically, its objectives are as follows:

- to construct and specify a regression model of the Census data item *Languages other than English spoken at home*, using other census responses or variables such as *Ancestry*, *Birthplace*, and *Religious Affiliation*;
- using the model, to estimate the number of people who speak a language other than English at home in each statistical local area (SLA) as if the detailed language information was not available;
- to assess the goodness of the fit and effectiveness of the model.

Methods and data

The analysis made use of the following techniques:

- Multivariate regression modelling;
- Distribution analysis; and
- Combination of the regression modelling and distribution analysis.

The analysis made use of New South Wales and Victoria data from the 1986, 1996 and 2001 Census of Population and Housing.

Major findings

The model constructed here predicts the number of speakers of the 30 most frequently spoken languages other than English on the basis of *Ancestry*, *Birthplace*, and *Religious Affiliation*. The findings include:

- Among the 30 languages included, 11 were underestimated and 19 were overestimated.
- Apart from one exception (Arabic), the percentage differences between the estimated numbers and the Census figures were below ten percent (25 out of 30 were below 5% and 4 out of 30 were between 5% and 10%).
- Arabic had the lowest percentage accuracy, which was 11%.

1. INTRODUCTION

The objective of the Census collection operation is to achieve a high quality census count that obtains maximum coverage of the population in a cost-effective manner.

It was suggested to reduce the coding for *Main Language Other than English Spoken at Home* topic in the 2006 Census and alternate censuses, if it can be shown that modelling language data from other census responses can provide detailed language data of an acceptable reliability.

Language use information are important for the implementation of national and state/ territory programs, and in particular, to ensure access to and equity of service delivery. They are also relevant to those interested in language retention issues.

This paper presents the results from multivariate modelling analysis of non-English languages. The research aimed to:

- construct a predictive model of languages spoken using other census responses such as *Ancestry*, *Birthplace*, and *Religious Affiliation*;
- estimate the number of people who speak a language other than English at home in each statistical local area (SLA), using the model above;
- evaluate the models and assess the feasibility of reducing the language question.

The paper is organized as follows:

- Section 2 provides a brief discussion of the data and the variables used in the modelling, as well as issues in the self-reported questions.
- Section 3 outlines the methodology used in the multivariate analysis, presents the results of the analysis, evaluates the models and the methodologies applied.
- Section 4 summarizes the analysis and concludes the discussions.

2. THE CENSUS DATA

This study made use of the 1996 and 2001 Censuses of Population and Housing data to model 'Other Language' for Victoria and New South Wales. The variables used in the model are *Ancestry*, *Birthplace*, and *Religious Affiliation*. To test the robustness of the models and goodness of the fit, we used the 1986 Census as apart from the 2001 Census, the 1986 Census was the latest which collected ancestry information.

Ancestry

The ancestry variable is an essential component in the modelling. The reliability of responses to the question plays a very important role in the estimation. In the 2001 Census, the question was "What is the person's ancestry?" Further instructions specifically allow respondents to provide multiple answers. The Census guide states 'When answering this question consider and mark the ancestries with which you most closely identify', then elaborates: 'Count your ancestry back as far as three generations, if known, for example, your parents, grandparents, or great grandparents.' These guidelines bring in both self-identification and descent criteria.

The interpretation of the question affects the way people answer this question. Some people may identify their preferred ancestry group. Others may report the countries in which their parents, or grandparents, or great grandparents were born before their arrival in Australia. While the question allowed for people to provide more than one ancestry only a few people provided this. A few others chose not to respond.

The investigation showed that self-identification based on *Ancestry*, and non-response to the question, both had an impact on the accuracy of language estimation to some extent. We take *Greek* as an example to highlight the issue.

Among 122,351 Greek speakers, 92% reported Greek ancestries, 2.6% did not state their ancestries, 1.4% reported South eastern Europe ancestries, 1.1% reported English ancestries, and 0.9% reported Australian ancestries. The percentages for other ancestries were each less than 0.8 and the total of them made 2.3%.

In order to predict based on the models developed, we needed to estimate the number of people speaking a language other than English at home as if the language information was not available. Therefore in the process of selecting the population to include in the model the following people had to be excluded:

- people who did not state their ancestries and not report their birthplaces;
- people who reported Australian ancestry and Australian birthplace or birthplace not stated;
- similarly, people who reported English ancestry and their birthplaces were, for example, Australia, not stated, or England.

To avoid underestimation and eliminate the impact of non-responses and self-identification, in the estimation process we select people whose ancestries were either Greek or South Eastern European, or whose birth places were either Greece or Cyprus.

Religious affiliation

The religious affiliation question also involves elements of self perception. Answering this question is optional. Our study showed that the religion variable was not as highly correlated as ancestry with languages. Therefore it was not utilised in every language we investigated. We used this variable to further identify people's characteristics when overestimation occurred. The effect of inconsistent answers to the religious question on the estimation was not expected as much as that of ancestry.

A language spoken at home

On the 2001 Census, the question was "Does the person speak a language other than English at home?". Some people interpreted the language as the one they can speak other than the one which is actually spoken at home, as we found out that some languages were spoken by only one person in some statistical local areas.

The impact of this misinterpretation can not be quantified in the study. The effect on the modelling and estimation remains unknown.

3. METHODOLOGY AND RESULTS

3.1 Multivariate regression analysis

Method

Multivariate regression analysis was the technique used in modelling. Broadly speaking, the technique describes and evaluates the relationship between a given variable and other variables.

In this analysis, the number of people speaking a language other than English at home in each SLA is the variable to be explained (dependent variable). The explanatory variables are the person's ancestry, birthplace and/or religion (independent variables).

Preliminary exploration of variables shows that these three explanatory variables are highly correlated and linearly related to the number of people speaking other languages. Thus we chose a linear regression model. We hypothesize that Y, the number of people speaking a language other than English in each SLA, can be expressed by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

where β_0, β_1, β_2 and β_3 are the intercept and slopes due to X_1, X_2 and X_3 respectively. X_1 is the ancestry variable, X_2 is the birthplace variable and X_3 is the religious affiliation variable. ε is the unknown error. In some cases we use more than one ancestry/ birthplace variable. The religious affiliation variable is used only in cases where the population needs to be further defined or when overestimation occurs.

The regression analysis is carried out in two phases:

In phase 1, the detailed language information is used. We count the number of people speaking a specific language other than English at home in each SLA and use ancestry, birthplace, and/or religion as explanatory variables to estimate the model.

In phase 2, we select population based on reported ancestry, birthplace, and/or religion as if the information about the language was not available. We apply the generated regression model to estimate the number of people speaking the language in each statistical local area. The estimated figures using established models are given in column 3 (Estimated figures using the models) in table 3.1.

Results

Using 2001 Census data for Victoria, we applied the multivariate regression analysis to model 25 languages. The summary results are shown in table 3.1. The languages are arranged in ascending order of the percentage differences between Census figures

and estimated figures using the regression models. Greek has the smallest percentage difference, and Arabic has the largest.

3.1 Regression analysis applied to SLAs in Victoria

Language	*Victoria 2001 Census figures*	*Estimated figures using the models*	*% difference between Census and estimation using the models*
1. Greek	122,351	122,442	0.07
2. German	20,253	20,272	0.09
3. Tagalog	18,010	18,039	0.16
4. Korean	3,186	3,191	0.16
5. Macedonian	32,632	32,566	0.20
6. Maltese	21,488	21,442	–0.21
7. Polish	19,576	19,630	0.28
8. Italian	149,185	148,647	–0.36
9. Croatian	25,555	25,455	–0.39
10. Persian	5,875	5,914	0.66
11. Vietnamese	63,816	63,221	–0.93
12. Spanish	22,874	22,660	–0.94
13. Portuguese	3,895	3,947	1.34
14. Russian	13,911	14,121	1.51
15. Khmer	8,546	8,678	1.54
16. Turkish	28,594	29,050	1.59
17. Indonesian	9,138	8,977	–1.76
18. Samoan	4,062	4,134	1.77
19. Japanese	5,153	4,920	–4.25
20. Sinhalese	11,641	12,200	4.80
21. Netherlandic	10,621	11,295	6.35
22. Serbian	16,036	14,857	–7.35
23. French	11,093	10,145	–8.05
24. Hungarian	8,913	9,671	8.50
25. Arabic	47,182	42,011	–10.96

In terms of the percentage differences between the Census figures and estimated numbers using the models (assuming that we had no knowledge of the languages) (see column 4), we observe that:

- For 20 out of the 25 languages investigated differences are below 5%, 4 are between 5% to 10%, and only 1 exceeds 10%;
- The language with lowest accuracy is Arabic (11% difference);

- The methods overestimate 15 languages including: Greek, German, Tagalog, Korean, Macedonian, Polish, Persian, Portuguese, Russian, Khmer, Turkish, Samoan, Sinhalese, Netherlandic, Hungarian, and underestimate 10: Maltese, Italian, Croatian, Vietnamese, Spanish, Indonesian, Japanese, Serbian, French, Arabic.

3.2 Distribution analysis

Method

Distribution analysis applies to the 'other language' distribution in each Statistical Local Area (SLA) from 1996 Census to the current (2001) Census. This method is based on the assumption that the language distribution in each SLA does not change very much in five years time. If the distribution changed dramatically, the estimation will not be reliable.

We use this technique to estimate the number of people speaking Aboriginal languages, as ancestry and birthplace are not valid explanatory variables for these particular languages in the modelling process.

Results

Using the 1996 Census data, we first calculated the proportion of the people who spoke an Indigenous language to those who spoke a language other than English at home in each SLA. We then selected the population who reported that they spoke a language other English at home in the 2001 Census. Finally, we applied the proportion obtained from 1996 Census to the selected population to predict the number of people speaking one of the Indigenous languages at home in 2001. The results are shown in table 3.2.

3.2 Distribution analysis applied to SLA s for indigenous languages in Victoria

Language	*2001 Census figures*	*Estimated figures using 1996 Census distribution*	*% difference Census and estimation using 1996 distribution*
Indigenous Languages	318	330	3.64

The estimated number for 2001 is 330, 3.64% more compared to Census figure 318, overstating the total number of people who speak an Indigenous languages.

However, using this approach and Census data only, we are not able to distinguish the specific Indigenous languages spoken at home among the Indigenous language speakers in Australia.

3.3 Combined regression and distribution analysis

Method

This approach used 2001 Census data to construct a regression model at the aggregated language levels, then applied the 1996 distribution to the estimated figures at the aggregated levels to predict the number of people speaking a language at the language base unit level. This technique was applied when we were not able to accurately distinguish the characteristics of several language speakers.

Languages such as Cantonese and Mandarin are typical examples of this. We could not distinguish between Mandarin and Cantonese speakers, as both language speakers are the descendants of Chinese. To deal with it, we first aggregated Chinese_nfd [1], Cantonese, Hakka, Hokkien, Mandarin, Teochew, Wu, and Chinese_nec [2] as one language, and then conducted modelling at the aggregated level using detailed information of language, ancestry and birthplace collected through 2001 Census. Following that, we computed the distribution of Chinese languages using the 1996 Census. Finally, we allocated Mandarin, Cantonese, and other Chinese language speakers to each SLA based on the distribution.

The language information was used as the dependent variable and ancestries and birthplaces reported from 2001 Census are used as independent variables in regression. Once the model is estimated we select Chinese language speakers based on their ancestries and birthplaces and apply the established model to estimate the number of Chinese language speakers as if we did not know the language details.

The allocation of this estimated number of the aggregated language to individual languages is undertaken in two steps. First using the proportion of each Chinese language collected from the 1996 Census, we distributed this number to estimate the numbers of various Chinese language speakers in Victoria for 2001. We then allocate these state totals to SLAs according the 1996 Chinese language distributions in SLAs to obtain the numbers of various Chinese language speakers in each SLA. The results are given in table 3.3.

Results

As can be seen, the estimated number (107,902) at the aggregated level understated the Census number (113,129), each individual Chinese related language was underestimated.

The estimation of Hindi and Tamil was undertaken in similar fashion. The Southern Asian languages, including Southern Asian_nfd. Malayyalam, Tamil, Telugu, Gujarati,

1 not further defined

2 not elsewhere classified

Hind, Konkani, and Southern Asian_nec, were aggregated in the modelling process. The results were also shown in table 3.3.

3.3 Results from combined regression and distribution analysis

Language	*Census figures*	*Estimated figures*	*% difference between Census and estimation*
Chinese Language	113,129	107,902	–4.62
Chinese_nfd	4,445	4,244	–4.52
Cantonese	60,583	58,100	–4.10
Hakka	4,492	3,956	–11.93
Hokkien	2,738	2,615	–4.49
Mandarin	38,880	37,075	–4.64
Teochew	1,503	1,439	–4.26
Wu	321	317	–1.25
Chinese_nec	167	163	–2.40
Southern Asian	46,787	48,036	2.67
Hindi	10,723	11,023	2.80
Tamil	7,968	8,202	2.94
Other [3]	28,096	28,802	2.51

3.4 Comparison with the 2001 Census

In general, the approaches used in this study have resulted in very high accuracy. The percentage differences between Census and estimated figures at the State level for Victoria are ranged from 0.07% (the lowest) to 10.96% (the highest).

As shown in tables 3.1, 3.2, and 3.3, of the 30 languages studied, Greek achieves the highest accuracy as measured by percentage difference (0.07) between the estimated figure (122,442) using models and the reported number (122,351) in the 2001 Census. Then it is followed by German. The percentage difference is 0.09 and the estimated number is 20,272 compared to Census figure 20,253. Tagalog comes third, the estimate is 18,039 compared to 18,010, the Census figure.

The languages that have high modelling accuracy are generally spoken by a group of homogenous people in terms of their ancestries, birthplaces and/or religions. Korean is a typical example of this. Around 94% of Korean speakers reported Korean ancestries. About 87% were born in Korea and 10% were born in Australia. Australian Maltese are another very homogenous group: 67% of them were born in Malta, 88% identified their ancestries were Maltese, and Western Catholic represented 96% of

3 Including: Southern Asian_nfd, Malayyalam, Telugu, Gujarati, Konkani, Southern Asian_nec.

their religious affiliation. As the religious affiliation is a very important factor for Australian Maltese it is built into the modelling process.

In contrast, the ancestries, birthplaces, and religions of Arabic speakers are very diversely distributed. 51% of Arabic speakers reported Lebanese ancestries, 12% reported Egyptian ancestries, 6% reported Arabic_nfd ancestries, and 4% reported Iraqi ancestries. The other reported ancestry figures are all less than 4%.

The diversity of birthplaces of Arabic speakers is another major contributor to the low modelling accuracy: 40% of them were born in Australia, 27% were born in Lebanon, 11% were born in Egypt, and 6% were born in Iraq. The other reported birthplaces are all less than 5%.

The variety of their religious affiliations include Islam religion (49%), Western Catholic (16%), Coptic Catholic (9%), Greek Orthodox (7%), Maronite Catholic (4%), and other which are all less than 4%.

To make the model reflect the characteristics of Arabic speakers as much as possible, we took account of four birthplaces (Australia, Lebanon, Egypt and Iraq), four ancestries (Lebanese, Egyptian, Arabic_nfd, and Iraqi), and the five above mentioned religions in model building. We experimented with different combinations of these factors used as the dependent variables in model building process as well as in selection of population for estimation once the model was established (see phase 2 in section 3.1). Among dozen of tested models, the best model managed to achieve 10.96% difference between Census figure (47,182) and the estimation (42,011).

Overall, the techniques applied in the study predict the number of people speaking a language other than English at home reasonably well. For the languages spoken by a group of people who are very diverse with respective to their ancestries, birthplaces, and religious affiliation, other factors such as education levels, year of migration, ages and etc., perhaps should be introduced in model estimations. Approaches like multilevel regression may also be explored.

3.5 Robustness of models

The methodology applied to the Victoria data can be applied to other states or territories. This will test the robustness of the methods.

Our initial investigation on New South Wales Census data is done by:

- selecting populations according to their ancestries, birthplaces, and in some cases religions in the same fashion as the selection process for Victoria;
- using models built on Victoria Census data for estimations;
- using the New South Wales language data to re-estimate the model coefficients and applying the models for estimations again;

- comparing the results from steps 2 and 3.

The outcomes of the investigation demonstrates that:

- the selection process based on persons' ancestries, birthplaces, and/or religions works well for New South Wales;
- the explanatory variables are still valid in the models;
- the linear regression models are still appropriate;
- the models built on Victorian data give reasonably good estimations for some languages, and need modifications for others.

We have used 1986 Census data to test the robustness of the modelling approaches as, apart from 2001 Census, it is the most recent Census in which ancestry question was asked. The test is carried out in the following steps:

- applying the models as they were built on 2001 Census directly to 1986 Census data to obtain an estimation;
- keeping the model unchanged but recalculating the model coefficients using 1986 language information;
- applying the models built on 1986 Census to get another estimation.

Our investigations demonstrate that the regression methods are statistically sound and quite robust with regard to the model selection and the explanatory variables included in the models.

However, the distribution approach and the combination of modelling and distribution techniques require language information from previous Census. This information of "a language other than English spoken at home" was not collected in 1981 Census. We have to limit our test to regression modelling only.

4. CONCLUSION

We have investigated the feasibility of predicting the number of people speaking a language other than English at home in each SLA given available information on ancestry, birthplace and religious affiliation. We have built up models based on 2001 Census data and use the generated models for estimation.

We find that ancestries, birthplaces and religious affiliation can be used as reasonably good predictors of language. The models established in the process can be used to predict the number of people speaking non-English languages at home in each SLA. The estimations using the models as if the language information was not available are reasonably accurate among the most languages examined. There are a few exceptions where more work is needed.

Australian Indigenous languages are modelled differently. We use Indigenous language distribution patterns from the 1996 Census to estimate the number of people speaking one of the Indigenous languages in each SLA. This approach only works well if the Indigenous language distributions do not change very much over a five year period.

For languages such as Cantonese, Mandarin, Tamil, and Hindi we combine modelling and distribution analysis techniques in the estimation.

We have achieved reasonably accurate predictions in most languages investigated at SLA level. Our investigation suggests the feasibility of reducing the language question output categories to only 'English' and 'Other' for future Censuses. If the detailed language information is not available, the estimations can provide some useful alternative information.

However, there are some drawbacks that need to be considered carefully. There is considered to be too much risk in using models over a period which is more than five years as the characteristics of people speaking a specific non-English language may change over time. Also we cannot say with confidence that the relationship established at one Census will hold over a long time period. Also, the modelling is not conducted at the personal level but at SLA level. The estimated number of people speaking a specific non-English language using models at SLA level cannot be linked back to each individual.

REFERENCES

Australian Bureau of Statistics (2003) *2001 Census: Ancestry – Detailed Paper*, Census Paper No. 03/01b, Canberra.

Australian Bureau of Statistics (2000) *Australian Standard Classification of Cultural and Ethnic Groups*, ABS cat. no. 1249.0, Canberra.

Australian Bureau of Statistics (1996) *Australian Standard Classification of Religious Groups*, ABS cat. no. 1266.0, Canberra.

Australian Bureau of Statistics (1997) *Australian Standard Classification of Languages*, ABS cat. no. 1267.0, ABS, Canberra.

Australian Bureau of Statistics (2003) *Australian Standard Classification of Countries*, ABS cat. no. 1269.0, Canberra.

Australian Bureau of Statistics (1999) *Standards for Statistics on Cultural and Language Diversity*, ABS cat. no. 1289.0, Canberra.

Australian Bureau of Statistics (2001) *2001 Census Dictionary*, ABS cat. no. 2901.0, Canberra

Chiswick, B.R. and P.W. Miller (2001) "A Model of Destination Language Acquisition: Application to Male Immigrants in Canada", *Demography*, Vol. 38, No. 3, August 2001, pp. 391–409.

De Vries, J. (1985) "Some Methodological Aspects of Self-Reporting Questions on Language and Ethnicity", *Journal of Multilingual and Multicultural Development*, Vol. 6, No. 5, pp. 347–368.

Haug, W. (2001) "Ethnic, Religious and Language Groups: Towards a Set of Rules for Data Collection and Statistical Analysis", *International Statistical Review*, Vol. 69, No. 2, pp. 303–311.

Stevens, G. (1999) "A Century of U.S. Censuses and the Language Characteristics of Immigrants", *Demography*, Vol. 36, No. 3, August 1999, pp. 387–397.

FOR MORE INFORMATION...

INTERNET	**www.abs.gov.au** the ABS web site is the best place to start for access to summary data from our latest publications, information about the ABS, advice about upcoming releases, our catalogue, and Australia Now.
LIBRARY	A range of ABS publications is available from public and tertiary libraries Australia-wide. Contact your nearest library to determine whether it has the ABS statistics you require, or visit our web site for a list of libraries.
CPI INFOLINE	For current and historical Consumer Price Index data, call 1902 981 074 (call cost 77c per minute).
DIAL-A-STATISTIC	This service now provides only current Consumer Price Index statistics call 1900 986 400 (call cost 77c per minute).

INFORMATION SERVICE

Data which have been published and can be provided within five minutes are free of charge. Our information consultants can also help you to access the full range of ABS information—ABS user-pays services can be tailored to your needs, time frame and budget. Publications may be purchased. Specialists are on hand to help you with analytical or methodological advice.

PHONE	**1300 135 070**
EMAIL	**client.services@abs.gov.au**
FAX	1300 135 211
POST	Client Services, ABS, GPO Box 796, Sydney 2001

WHY NOT SUBSCRIBE?

ABS subscription services provide regular, convenient and prompt deliveries of ABS publications and products as they are released. Email delivery of monthly and quarterly publications is available.

PHONE	1300 366 323
EMAIL	subscriptions@abs.gov.au
FAX	03 9615 7848
POST	Subscription Services, ABS, GPO Box 2796Y, Melbourne 3001

2000001488607
ISBN 0 642 47999 2

RRP $10.00