Assigning a Socio-Economic Status Value to Student Records: A Useful Tool for Planning, Reporting and Institutional Research

JULIE DELANEY[1], PLOY TANGTULYANGKUL[2] AND ROBERT MCCORMACK[3]

[1] *Institutional Research Unit, The University of Western Australia.*

[2] *Statistics Office, The University of Western Australia.*

[3] *Planning Services, The University of Western Australia.*

Submitted to the *Journal of Institutional Research* January 20 2013, accepted for publication March 28 2013.

Abstract

In an educational context, the accurate determination of each student's socioeconomic status (SES) is important for planning, reporting and general institutional research. This article describes a project undertaken to develop the means to derive a proxy measure of students' SES, based on home address location and Australian Bureau of Statistics (ABS) data, using free Geographic Information System (GIS) software.

Keywords: geocoding; socio-economic status (SES); student load planning; geographic information system (GIS); institutional research

This article describes a project undertaken in the Institutional Research Unit and Statistics Office of Planning Services at The University of Western Australia (UWA) to develop the means to map student permanent address information (i.e., the student's street address, suburb, postcode and country of permanent residence) to a proxy for SES.

The project comprised two main phases as follows:

- Emulation of the Department of Industry, Innovation, Climate Change, Science, Research and Tertiary Education (DIICCSRTE) classification of student permanent addresses into SES categories; and
- Preparation of an automated system to undertake such processing of the data in a timely, reliable, repeatable and cost-effective manner.

This article was first presented at the Annual Conference of the Australasian Association for Institutional Research, *Evolution of IR practice and use of IR*, Terrigal, New South Wales, November 12–14, 2013.

Correspondence to: Robert McCormack, Planning Services, The University of Western Australia. E-mail: robert.mccormack@uwa.edu.au

Context

Classification of student permanent residential addresses into SES categories is important for a range of internal and external purposes. These purposes include the offering of places to support and encourage equal access to university for people from the full range of SES backgrounds, and funding arrangements. Papers by Edwards et al. (2005), Edwards and Marks (2008), James (2002), Machin (2006) and Stevenson et al. (2001) highlight the need to facilitate and maintain such schemes.

The University of Western Australia's Operational Priorities Plan (OPP) 2009–2013 states that the university is seeking to 'enhance further the quality of its student body with a deep commitment to equity access and diversity' and has articulated as its first operational objective that it will aim 'to recruit and graduate a diverse student cohort of the highest quality' (UWA, 2009, p. 12). The OPP also articulates a range of strategies to advance this objective. The development of the means to map student addresses to SES will allow the university to better measure its progress in achieving a more diverse student body.

The Australian Government (2009) has articulated an ambition for a 'fairer Australia' in which 'all Australians will benefit from widespread equitable access to a diverse tertiary education sector that allows each individual to develop and reach their potential' (p. 7). It has proceeded to develop 'Mission-Based Compacts' with universities, which outline the government's policies, plans and priorities for the higher education sector and allow each university to articulate their plans and priorities and how they relate to the government's overall direction. The Mission-Based Compacts also specify a range of performance indicators and associated targets, including a performance indicator on the proportion of domestic undergraduates who are from a low-SES background. As explained below, this indicator is now based on a combination of a proxy-SES derived from each student's permanent residential address and the proportion of students in receipt of student support provided by the Australian Government.

In its information paper: *An Introduction to Socio-Economic Indexes for Areas* (Australian Bureau of Statistics, 2006), the ABS notes the concept of relative socioeconomic disadvantage is neither simple nor well-defined. It nonetheless defines relative socioeconomic advantage and disadvantage in terms of people's access to material and social resources, and their ability to participate in society. It first derived a measure of SES from data collected in the 1971 *Census of Population and Housing*. It combines data from the census to produce four composite Socio-Economic Indexes for Areas (SEIFA):

- Index of Relative Socio-Economic Disadvantage: low income earners, relatively lower educational attainment and high unemployment
- Index of Relative Socio-Economic Advantage and Disadvantage: economic and social resources including measures of both relative advantage and disadvantage
- Index of Economic Resources: rent paid, income by family type, mortgage payments, and rental properties
- Index of Education and Occupation: general level of education and occupation-related skills of people within an area.

Prior to a review undertaken in 2009, the Department of Education, Employment and Workplace Relations (DEEWR; now the Department of Industry, Innovation, Climate Change, Science, Research and Tertiary Education [DIICCSRTE]) mapped the postcode of

student permanent home addresses to SES values (of high-SES, medium-SES and low-SES) using the SEIFA Index of Education and Occupation (IEO). In particular, low-SES postcodes were defined as those with the bottom 25% of the population on the IEO aged between 15 to 64 years. Following the review, DEEWR decided to refine its method for classifying students into SES groups from one based on the postcode of the permanent home address to one based on mapping the permanent home address to SES via Census Collection District (CD). It combined the resultant proportion of domestic undergraduate students from low-SES backgrounds with the proportion of students on Centrelink benefits (i.e., student financial support provided by the Australian Government) in a ratio of two to one to establish the principal measure of participation and social inclusion used in Mission-Based Compacts with Australian universities.

The impetus for the project was that while the Australian Government had initially indicated that it was likely to provide institutions with appropriate software that would mirror their CD mapping, the software had not been provided at the time the project was initiated. The project needed to address complications stemming from the nature of the raw data, the ability and reliability of the tools to process the data, definitions of the variables/categories and accuracy of the end product.

In undertaking this project, it also became apparent that there may be multiple opportunities to use student data to improve SES classifications, as suggested in numerous other articles such as Ainley and Long (1995), Graetz (1995) and Marks (2011). However, this article focuses on the work undertaken to deliver a workable method of mapping student addresses to a proxy for SES.

Phase 1: Emulation of Previous Classifications

The initial phase of the project attempted to replicate the DIICCSRTE classification of student permanent addresses into SES categories using data from the 2008, 2009 and 2010 Student Data Collection files. In particular, the project used the details of undergraduate Commonwealth-supported students submitted on the Commonwealth Assisted Students—Help Due files. However, only one record was used for each student. Furthermore, cross-institutional students and students who changed into postgraduate courses were excluded from the analysis. The project also used a DIICCSRTE-supplied data file that contained student address data and corresponding SES classifications assigned to students for these years. The DIICCSRTE also provided some advice on how it had mapped student addresses to SES.

Phase 1—Method Used

The method that was developed maps each permanent address to spatial coordinate data through georeferencing, and then uses a Geographic Information System (GIS) to map the spatial coordinates to a CD, before mapping the resultant CD to an average SES value associated with that location. The map of CDs and postcode areas to the relevant SES index was undertaken by the ABS and released by DIICCSRTE. Examples of the use of GIS in similar situations can be found in Delaney and van Niel (2007).

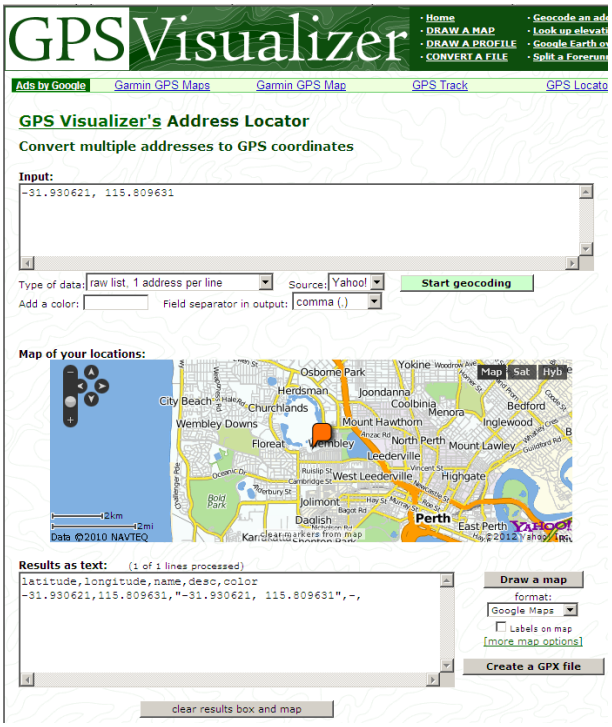
The method developed in Phase 1 is displayed in Figure 1.

1. Obtain student address data from records

For example:
StudID = xxxxxxxx, 1 Moondine Drive, WEMBLEY, WA, Australia, 6014, DIICCSRTE SES = 2

2. Georeference student data using an online, free tool, to generate spatial coordinates

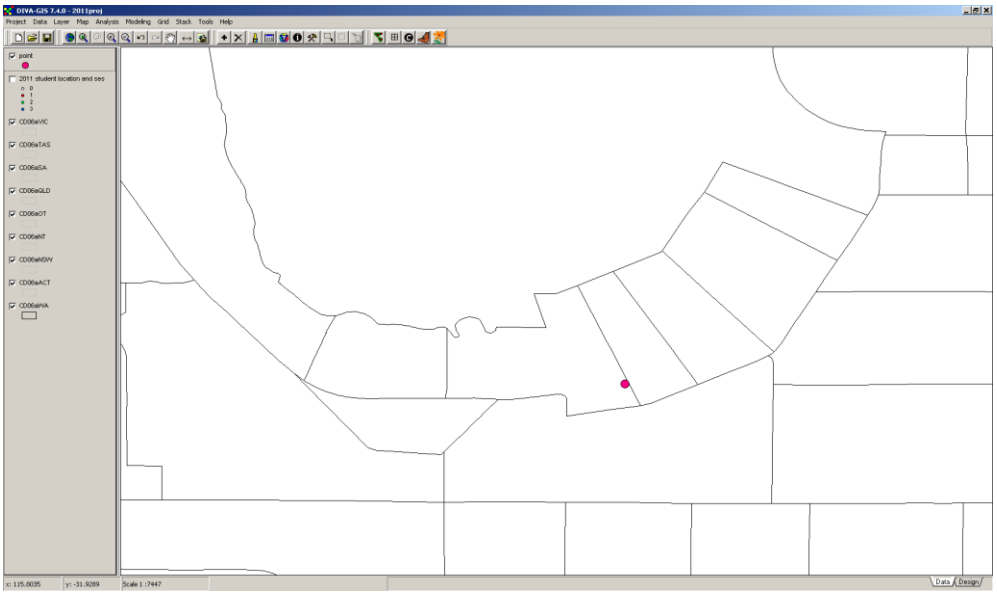
http://www.gpsvisualizer.com/geocoder/



For example: -31.930621, 115.809631

3. Read the spatial coordinate data into a GIS software package, such, as DIVA-GIS

4. Overlay the CD boundaries in the GIS. Use the GIS software to link polygon data (CD) to point data (student residential location—spatial coordinate data). Keep the unique student id number to identify each student



5. Export the linked data back into a spreadsheet.

For example: StudID = xxxxxxxx, CD_CODE06 = 5101214

6. Use a look-up table of CDs with associated SES values (supplied by DIICCSRTE) to obtain the SES value for each student record.

For example: StudID = xxxxxxxx, CD_CODE06 = 5101214, UWA-SES = 2

7. Use the common unique item, StudID, to link all data in the one spreadsheet.

For example, StudID = xxxxxxxx, 1 Moondine Drive, WEMBLEY, WA, Australia, 6014, DIICCSRTE -SES = 2, -31.930621, 115.809631, CD_CODE06 = 5101214, UWA-SES = 2

8. Compare the assigned SES classification (UWA-SES) with the DIICCSRTE assigned SES classification (DIICCSRTE -SES)

Figure 1. The method for determining and linking SES classification to a student's record—emulating previous classifications.

Phase 1—Results

Tables 1, 2 and 3 summarise the results of the attempts to emulate the SES categories assigned to student records by DIICCSRTE.

Table 1

Percentage of UWA Students by SES Groups, UWA versus DIICCSRTE Method, 2008

	DIICCSRTE Low-SES	DIICCSRTE Medium-SES	DIICCSRTE High-SES	DIICCSRTE Unclassified	DIICCSRTE Total
UWA Low-SES	4.27	0.89	0.42	0.04	5.62
UWA Med-SES	0.75	32.03	3.10	0.07	36.00
UWA High-SES	0.22	2.79	54.18	0.08	57.27
UWA Unclassified	0.06	0.26	0.79	0.00	1.11
UWA Total	5.33	35.97	58.49	0.20	100.00

Note. Columns report percentage of students. *N* = 12,015

Table 2

Percentage of UWA Students by SES Groups, UWA versus DIICCSRTE Method, 2009

	DIICCSRTE Low-SES	DIICCSRTE Medium-SES	DIICCSRTE High-SES	DIICCSRTE Unclassified	DIICCSRTE Total
UWA Low-SES	4.52	0.96	0.35	0.05	5.86
UWA Med-SES	0.62	32.97	2.54	0.08	36.24
UWA High-SES	0.14	2.20	57.28	0.10	56.72
UWA Unclassified	0.08	0.29	0.82	0.00	1.19
UWA Total	5.39	36.41	49.97	0.23	100.00

Note. Columns report percentage of students. *N* = 12,980

Table 3

Percentage of UWA Students by SES Groups, UWA versus DIICCSRTE Method, 2010

	DIICCSRTE Low-SES	DIICCSRTE Medium-SES	DIICCSRTE High-SES	DIICCSRTE Unclassified	DIICCSRTE Total
UWA Low-SES	4.35	0.84	0.20	0.06	5.44
UWA Med-SES	0.70	33.64	2.41	0.08	36.83
UWA High-SES	0.09	2.12	53.22	0.10	55.53
UWA Unclassified	0.10	0.58	1.48	0.03	2.19
UWA Total	5.24	37.17	57.32	0.27	100.00

Note. Columns report percentage of students. *N* =13,762

As can be seen, the SES obtained under both methods agreed for 90.5% of the students enrolled in 2008, 94.8% of the students enrolled in 2009 and 91.2% of the students from 2010. After removing the students that DIICCSRTE did not classify, the percentage of records that agreed increased marginally to 90.7% for 2008, 95.0% for 2009 and 91.5% for 2010.

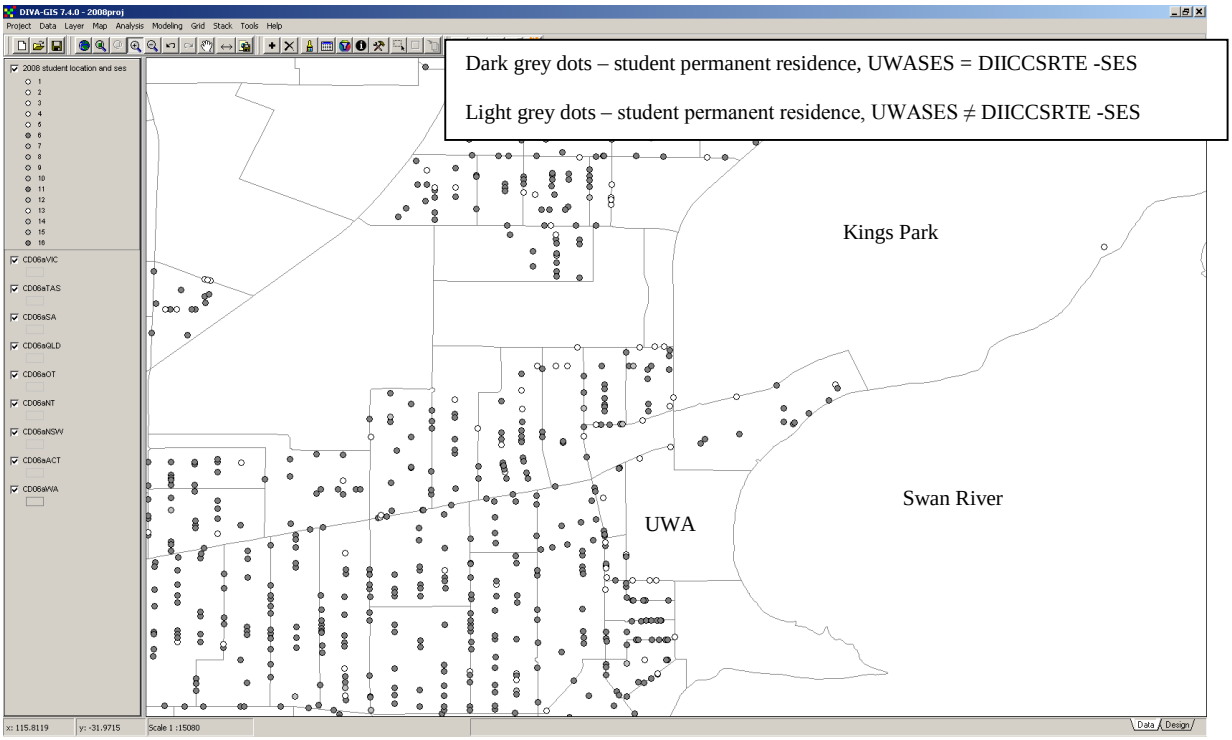
While both methods produced similar results, it was not possible to fully replicate the method used by DIICCSRTE to assign SES classifications to student permanent address data.

Some of the differences may be due to, for example, the rules used and the way the data were treated in preprocessing as well as the software used. Nonetheless discrepancies were further examined with a view to improving the initial method as explained below.

Phase 1—Discussion

As can be seen, the initial method developed by UWA produced relatively high numbers of records that could not be assigned an SES value. This appeared to be because DIICCSRTE used the postcode for some records when the software used by DIICCSRTE was unable to georeference the address that was provided, whereas the initial UWA method left these records unclassified. The rationale for using the postcode is that it may give some information about the SES of households in an area, especially for those postcodes that are fairly homogeneous.

Further differences became apparent when inspecting the data spatially. Figure 2 shows a screen snapshot from the GIS software DIVA-GIS. The dark grey dots show the locations of students who were mapped to the same SES by both the UWA method and the DIICCSRTE method, whereas the light grey dots locate students who were classified differently by UWA and DIICCSRTE and the clear circles show unclassified cases. It appeared that many of the addresses that produced different SES values under the two methods tended to be located very close to the CD boundaries.



Source: CD boundaries from ABS (2006), Screenshot from DIVA-GIS software.

Figure 2. Differences between DIICCSRTE-SES and UWA-SES, 2008 data.

Examination of addresses near CD boundaries showed that a significant number were mapped to the wrong CD because the polygonal boundaries of CDs and student latitudes/longitudes were captured at different scales and levels of accuracy.

The GPS Visualiser interface (the free online utility that uses georeferencing from Yahoo!™ Map Web Services) also places the 'pin' for residential location at the front of the property boundary, and sometimes in the street, increasing the likelihood of spatial data errors when overlaying the data with CD boundaries. This can be avoided by using other georeferencing tools, such as Google™ Maps, which place the 'pin' for a residential address inside the land parcel. Alternatively, another GIS tool, such as the Geographic Resources Analysis Support System (GRASS, 2012), can be used to 'buffer' the border zones to identify potentially problematic data.

In a similar vein, these spatial issues lead to some permanent home addresses being incorrectly mapped over the CD boundary to adjacent space that is unclassified. This occurred occasionally when a student's permanent home residence bordered an unclassified portion of land, such as park land (e.g., Kings Park) or a natural feature without census data (e.g., the Swan River).

The analysis also identified that there were significant numbers of student address records with deficiencies in the details provided that included typing errors, incorrect data in fields, incorrect postcodes and use of PO Boxes. In addition to leaving addresses unable to be geocoded, these errors may also lead to records being assigned an SES that is inconsistent with the student's permanent home address. Table 4 provides examples of the type of data limitations discovered, using 'dummy' data.

Table 4

Examples of Input Data Challenges

StudID	Street Address 1	Street Address 2	Suburb	State	Country	Postcode
xxxxxxx x	Unit 35	1 Moondine Drive	WEMBLEY	WA	AUSTRALIA	6014
xxxxxxx x	Unit 35	1 Moondine Drive		WA		6014
xxxxxxx x	Unit 35	1 Moondine Drive	WEMBLEY	WA	AUSTRALIA	6998
xxxxxxx x	Unit 35	1 Moondine Drive	WEMBLEY	WA	SWEDEN	6014
xxxxxxx x	Unit 35	1 Modine Drive	WEMBELY	WA	AUSTRALIA	
xxxxxxx x	PO Box 17		WEMBLEY	WA	AUSTRALIA	6998
xxxxxxx x	U. 35/1	Moondine Drive	WEMBLEY	WA	AUSTRALIA	
xxxxxxx x	UWA	St George's College	Nedlands			
xxxxxxx x	35	Wembley	Australia			

Phase 2—An Automated System

The next phase of the project drew on the lessons learned from the first phase of the project to develop an automated system to map student addresses to a proxy for SES.

Phase 2—Method

The method that was established uses SAS® (a Business Analytics and Business Intelligence Software currently used in Planning Services, UWA) and Java™ to process the data. It comprises five main steps (described in Figure 3).

- Obtain student permanent address data from records
- Pre-process (clean) address data
- Remove overseas records
- Remove any characters that might confuse the georeferencing (/, ', ., Unit X, Apartment X)
- Add missing country name/state when they can be generated or determined
- Check for confused term/permanent addresses
- Remove Lots, PO Boxes, Location, RMB from addresses
- Replace postcodes which are for PO Boxes with a valid postcode using a lookup table
- Georeference student permanent address data to obtain spatial coordinates
- Use SAS script to call 'middleware'—a Java library to retrieve and process the information from the Google Map API georeferencing service or similar
- Map student spatial coordinate data to the CD information
- Use SAS script to retrieve boundaries information from the GIS library called GeoTools
- This requires the CD files to be loaded via the middleware. Query the information with latitude and longitude to get the boundary name or code
- Determine SES
- The CD boundary is converted to SES using a SAS format, which has a KeyValuePair data structure generated from the spreadsheet of SEIFA data provided by DIICCSRTE

Figure 3. Automating the allocation of SES to a student's record.

A key part of this phase was to develop the means to identify and handle anomalous student address records, including developing the means to clean deficient records and working out how to handle records that could not be geocoded after data cleaning. As can be seen from Figure 3, data cleaning involving several processes is undertaken in step 2.

In addition to the steps shown in Figure 3, records that could not be geocoded after data cleaning were coded on the basis of the postcode associated with the permanent home addresses, when valid postcodes were available.

Phase 2—Results

Reprocessing the results for domestic undergraduate students enrolled in 2010, as expected, reduced the number of records that could not be mapped to a CD from 302 records to 264 records. As can be seen from Table 5, only 1.92% of records could not be assigned an SES value using the automated method. However, this is still more than the 0.27% of records that were left unclassified by the DIICCSRTE method.

Furthermore, the automated method did not improve the percentage of records that agreed with the SES value allocated under the DIICCSRTE method. The refined automated method agreed with the DIICCSRTE method for 90.9% of all records, as compared with 91.2% under the initial method. Even after removing records that could not be coded by DIICCSRTE, the automated method agrees for 91.0% of cases as compared with 91.5% under the initial method.

Table 5

Percentage of UWA Students by SES Groups, UWA Automated Method versus DIICCSRTE Method, 2010

	DIICCSRTE Low-SES	DIICCSRTE Medium-SES	DIICCSRTE High-SES	DIICCSRTE Unclassified	DIICCSRTE Total
UWA Low-SES	3.94	0.57	0.20	0.00	4.72
UWA Med-SES	0.49	32.79	2.00	0.04	35.33
UWA High-SES	0.69	3.31	54.00	0.03	58.04
UWA Unclassified	0.12	0.49	1.11	0.20	1.92
UWA Total	5.24	37.17	57.32	0.27	100.00

Note. Columns report percentage of students. N = 13,762

Phase 2—Discussion

While the automated method allows an SES classification to be assigned to student records with minimal cost and effort, it allocates a different SES value to a significant number of records and produces more unclassified records than the method employed by DIICCSRTE. Further work is being undertaken to investigate these differences and consider refinements to the method to suit changing needs and data sources. One possible enhancement would be to validate student address records against an authoritative address database.

Further General Discussion

The project achieved its aim of delivering an automated system to produce proxy SES values derived from student address data in a timely, reliable, repeatable and cost-effective manner. The divergence between the results produced by the system from those used by DIICCSRTE requires further investigation, with a view to delivering the means for Australian universities to map the addresses of their students to proxy SES values on a consistent, cost-efficient basis.

The project also highlighted the need for student address data to be of a suitably high quality. It became clear that the methods examined and developed in the project rely heavily on the quality and formatting of the address data. In this regard, the quality of address data is compromised at many universities, including The University of Western Australia, because students enter and maintain their contact and home address details directly onto student systems. The quality of these data may be improved if data validation procedures are developed and applied when address details are being entered. For example, requiring an actual address and rejecting a PO Box in the online form used by the students would significantly lessen the work needed for cleaning the data. However, the need to balance standardised data entry (e.g., a postcode is a 4-digit number) and validating postcode ranges with the ability of a system to accept all possible cases (e.g., some overseas postcodes are longer than 4 digits) should be acknowledged.

A further issue with using the student-entered data is the definition of permanent address. Some students change their permanent address to move closer to their university and in doing so their attributed SES changes. This can lead to an understating of the proportion of

students from low-SES areas, although it is acknowledged that most school leavers from metropolitan schools will remain in their parental home.

As indicated above, this project was undertaken primarily to develop the means to map student address details to a proxy for SES, which in turn can be used gain a better understanding of the SES mix of the university's student population and to investigate the educational performance in relation to SES. Beyond these uses, the project has delivered the means to generate spatial location information from student addresses. This spatial information can be used, or exploited, for many diverse purposes such as identifying for marketing purposes, localities where enrolments appear unusually low, investigating physical access to campus along public transport routes, organising car-pooling groups or locating off-campus events accessible to students. Literature such as Cresswell and Underwood (2004) and Blakers et al. (2003) also provide examples of opportunities to use similar data in a spatial context.

While this project covered how to map student address details to a proxy for SES, as noted above, DIICCSRTE also uses the proportion of students on Centrelink benefits in assessing each university's performance in relation to targets specified in relation to the proportion of domestic undergraduates who are from a low-SES background. It is not possible for each university to verify the proportion of its students on Centrelink benefits because relevant Centrelink data for individual students are not released to universities and are unlikely to be released to institutions in the future due to requirements to protect individual privacy.

Finally, it should be noted that more relevant and complex measures of SES have been proposed, for example those in Graetz (1995) and Marks (2011), and these are available for use in exercises such as this project.

Conclusion

The need to classify students into SES for the purpose of planning and reporting has led to the development of a tool that may be of interest to many in the same context. While the initial focus was on attempting to replicate classifications used by DIICCSRTE, it has led to the development of an automated system that maps address details to spatial coordinates, which may be used for various other purposes beyond understanding the SES mix of the student body.

Acknowledgments

The authors would like to acknowledge and thank their supervisors Greg Marie and Luke Minchin for their support and encouragement in this project and in preparing and presenting this report. Additionally, Ploy Tangtulyangkul would like to thank Nicola Powell, Strategic Information Analysis Manager at Monash University, and Jody Garnett from GeoTools for technical assistance.

References

Ainley, J., & Long, M. (1995). Measuring student socioeconomic status. In J. Ainley, B. Graetz, M. Long, & M. Batten (Eds.), *Socioeconomic status and school education* (pp. 53–76). Canberra: AGPS.

Australian Bureau of Statistics. (2006). *An introduction to socio-economic indexes for areas*. Retrieved from http://www.abs.gov.au/ausstats/abs@.nsf/mf/2039.0

Australian Government. (2009). *Transforming Australia's Higher Education System.* Retrieved from http://www.innovation.gov.au/HigherEducation/Documents/TransformingAusHigherED.pdf

Blakers, R., Bill, A., MacLachlan, M., & Karmel, T. (2003). *Mobility: Why do university students move?* (Occasional Paper Series 03-A). Canberra: Department of Education, Science and Training.

Cresswell, J., & Underwood, C. (2004). *Location, location, location: Implications of geographic situation on Australian student performance in PISA 2000*. Camberwell, Victoria: Australian Council for Educational Research.

Delaney, J., & van Niel, K. (2007). *Geographical information systems: An introduction* (2nd ed.). Melbourne, Australia: Oxford University Press.

Department of Education, Employment and Workplace Relations. (2009). *Measuring the socio-economic status of higher education students*. Retrieved from http://www.innovation.gov.au/HigherEducation/Documents/LowSES_Discussionpaper.pdf

Edwards, D., Birrell, B., & Smith, T.F. (2005), *Unequal access to university: Revisiting entry to tertiary education in Victoria.* Melbourne, Australia: Centre for Population and Urban Research.

Edwards, D., & Marks, G. (2008). Preliminary report on university participation access and entry in Victoria. In *Victorian Government Submission to Review of Australian Higher Education* (Appendix 4). Melbourne, Australia: Government of Victoria.

Graetz, B. (1995). Socioeconomic status in education research and policy. In J. Ainley, B. Graetz, M. Long, & M. Batten (Eds.), *Socioeconomic status and school education* (pp. 23–51). Canberra: AGPS.

Geographic Resources Analysis Support System. (2012). [Online GIS software]. Retrieved from http://grass.osgeo.org/

James, R. (2002, April). *Socioeconomic background and higher education participation: An analysis of school students' aspirations and expectations*. (Report from Evaluations and Investigations Programme). Canberra: Department of Education, Science and Training.

Machin, S. (2006). *Social disadvantage and educational experiences* (OECD Social, Employment and Migration Working Paper No. 32). Paris, France: Organization for Economic Co-operation and Development.

Marks, G. (2011). Issues in the Conceptualization and measurement of socioeconomic background: Do different measures generate different conclusions? *Social Indicators Research*, *104*, 225–251.

Stevenson, S., Evans, C., MacLachlan, M., Karmel, T., & Blakers, R. (2001). *Access: Effects of campus proximity and socioeconomic status on university participation rates in regions* (Occasional Paper Series 01/C). Department of Education, Science and Training: Canberra.

The University of Western Australia. (2009). *Operational Priorities Plan 2009–2013*. Retrieved from http://www.uwa.edu.au/university/strategy