How vulnerable is your community to a natural hazard? Using synthetic estimation to produce spatial estimates of vulnerability.

Susan Day
National Centre for Social and Economic Modelling (NATSEM)
University of Canberra. ACT 2601
Australia
Ph: +61 2 6201 2336
Fax: +61 2 6201 2751
Susan.Day@natsem.canberra.edu.au

Anita Dwyer
Geoscience Australia
GPO Box 378, Canberra. ACT 2601
Australia
Ph: +61 2 6249 9027
Fax: +61 2 6249 9986
Anita.Dwyer@ga.gov.au

How vulnerable is your community to a natural hazard? Using synthetic estimation to produce spatial estimates of vulnerability.

ABSTRACT

Geoscience Australia and NATSEM have recently collaborated to produce experimental estimates of the geographic distribution of vulnerability to a natural hazard. Geoscience Australia (GA) has devised a methodology to quantify community impacts of natural hazards. NATSEM has used its synthetic estimation techniques to produce estimates of vulnerability to flood in 224 Census Collection Districts in Perth. GA's methodology comprises four stages: indicator selection, a risk perception questionnaire, a decision tree analysis and a case study. This paper is about the fourth step, in which NATSEM was involved.

The Australian Bureau of Statistics (ABS) 1998/99 Household Expenditure Survey with MarketInfo 2001 small area weights derived from the ABS 1996 Census of Population and Housing, are used to create a detailed picture of small area populations for the year 2001 and estimate their vulnerability to natural hazards.

A major aim of this collaborative project was to assess each of the four stages and identify methodological improvements. An assessment of the estimates of vulnerability and the role of synthetic estimation must be done in the context of the preceding steps in the four-stage methodology. Due to the issues identified, we see the synthetic vulnerability outputs as powerful examples of the kind of small area data that can be produced using synthetic estimation rather than as authoritative final estimates of vulnerability in their own right.

1 INTRODUCTION

Geoscience Australia (GA) has devised a methodology to quantify community impacts of natural hazards. The first stages of the methodology are non-spatial – they identify variables, including socio-economic characteristics that are relevant to vulnerability, and define a relationship between those variables and perceptions of vulnerability. NATSEM, using its synthetic estimation techniques, has estimated those socio-economic variables, and hence vulnerability, in selected Census Collection Districts (CDs). The methodology could be used to deliver spatial estimates of vulnerability anywhere in Australia for CDs or larger aggregates of CDs.

To date, NATSEM's synthetic estimation and microsimulation techniques have largely been used to look at socio-economic characteristics of individuals and households in small areas. Recent work includes that on small area expenditure

patterns (using the MarketInfo model), telecommunications (Hellwig and Lloyd 2000), incomes and poverty (Harding et al 2000; Lloyd, Harding and Hellwig 2001; Lloyd, Harding and Greenwell 2001) and income support (King, McLellan and Lloyd 2002). This collaborative project with Geoscience Australia provided an opportunity to apply NATSEM's techniques in the physical sciences domain. This project, while experimental, has relevance to strategic and tactical emergency planning.

This paper describes the method used in applying synthetic estimation to quantifying vulnerability and reports on the results of the application. It is divided into four sections. Section 2 provides some contextual background on the Geoscience Australia Risk Research Group's modelling approach. Section 3 describes the methodology used to identify relevant variables in the Household Expenditure Survey (HES), create synthetic socio-economic profiles for small areas and synthetic estimates of vulnerability. In addition it presents the synthetic estimates and highlights a number of issues encountered in this first application of GA's four-stage methodology for quantifying vulnerability. Section 4 outlines some different ways of assessing the synthetic data.

2 BACKGROUND

2.1 Perth Cities Project

Geoscience Australia is currently performing a multi-disciplinary natural hazard risk assessment study in the Perth region. GA views risk as a combination of hazard, the exposure of elements (such as people and infrastructure) and their vulnerability. For the Perth Cities Project a new methodology devised by the Risk Research Group at GA is being employed for the first time.

The Perth Cities Project began in 2001 and is investigating natural hazard risks in 46 Local Government Areas around the Perth region, from as far north as Gingin, south to Busselton and east to Northam. These hazards include flood, earthquake, coastal erosion, storm surge and severe wind. GA's project partners include the Bureau of Meteorology, Fire and Emergency Services, Local Government Authorities and the University of Western Australia. Each of the four stages is seen as a pilot study to assess its usefulness in risk assessment. GA's four-stage methodology is described in detail elsewhere (Dwyer et al. 2003) and is summarised below. For this study, GA's definition of vulnerability is 'a person's ability to recover from a natural hazard impact', and recovery is defined as 'the time taken for a person to regain a lifestyle similar to the one they had prior to the event'.

Indicator selection

While actual vulnerability is likely to depend on a multitude of personal, family and community attributes (Granger and Hayne 2001), it is believed that a more limited set of quantifiable socio-economic characteristics can be used to assess relative vulnerability. In conjunction with two variables, injury and residence damage, GA has identified thirteen socio-economic variables for assessing important aspects of relative vulnerability (refer Table 1). These have been selected from a range of data sources, predominantly the ABS 1996 Census Basic Community Profile (BCP) and the ABS 1998/99 Household Expenditure Survey (HES).

Risk perception questionnaire

A risk perception questionnaire was developed based on a set of hypothetical individuals (one member of a household). The individuals were generated, primarily from 1996 Census data. Computer generated cases described each individual's socio-economic characteristics, as well as the degree to which they were injured or their home damaged, in terms of the fifteen selected variables. Between November 2002 and January 2003, questionnaire respondents were asked to rank each hypothetical person's ability to recover from each of four hazards (earthquake, flood, landslide and cyclone). A score of 1 signified little time to recover, while 10 indicated infinite time. Each questionnaire contained ten cases. Of the 460 questionnaires that were distributed, 152 completed questionnaires were returned; hence, 1520 hypothetical individuals were ranked. Box 1 shows an example of the questionnaire for a single hypothetical individual.

Throughout this paper, the word 'questionnaire' is used to refer to GA's questionnaire. The word 'survey' is used in general to refer to the kind of data that can be reweighted using synthetic estimation or to refer to the HES, which is the particular survey that is reweighted in this project.

Decision tree analysis of vulnerability

A decision tree analysis was used to establish a relationship between the fifteen selected variables and the vulnerability rankings assigned to flood victims by respondents to the questionnaire. The decision tree analysis resulted in two classes of vulnerability: high and low. The final decision tree classified the data into 22 nodes, 11 of which represent high vulnerability and 11 low. Injury and damage as well as five of the socio-economic variables appear in the decision rules that assign a person to the high vulnerability class. Listed according to the decision tree's relative order of importance, the socio-economic variables are: house insurance, income, tenure type, age and household type.

Table 1 Variables selected by GA for assessing vulnerability

Socio-economic characteristics & impact variables		
Age	English Skills	Income
Car ownership	Gender	Residence Type
Debt	Health Insurance	Tenure
Disability	House Insurance	Residence damage[1]
Employment	Household Type	Injury[1]

1. Impact variables.

Box 1 Hypothetical individual 1148

Person # 1148 is a female aged 38, is employed, earns $390 a week and is running into debt. They live as a couple with dependent children and own with no mortgage a flat in one or two storey building with house\contents insurance and has 1 car. This person does not have health insurance, does not have a disability and speaks English well.

Due to the hazard impact, this person requires basic medical treatment without hospitalisation and their residence is moderately damaged.

According to the recovery scale, indicate how long do you think it will take for this person to recover from the following impact?

A destructive earthquake (1-10)...7....
A major flood (1-10) ...5....
A major landslide (1-10) ...5....
A major cyclone (1-10) ...7....

1 5 6 10
Little time Sufficient time to interrupt life Infinite time

Source: GA questionnaire (computer generated cases for hypothetical individuals)

Case study

Synthetic estimation was used to predict the number of people in each CD in the high vulnerability class based on the rules determined by the decision tree for flood victims. These estimates are based on a number of scenarios defined by the injury and damage variables that describe the outcome of the flood event. NATSEM's application of synthetic estimation techniques in this case study builds on the three preceding steps undertaken by GA.

Study area

The CDs chosen for the case study are in the north and north east of the Perth metropolitan area and are in areas subject to hazards as estimated from GA's current hazard modelling. They include:

- All 153 CDs in Statistical Local Area (SLA) 58758 (Wanneroo (C) - South-West)
- All 44 CDs in SLA 58757 (Wanneroo (C) - South-East)
- 27 CDS in SLA 58050 (Swan (S))

2.2 Synthetic regional data

More detailed accounts of synthetic estimation, reweighting approaches and MarketInfo linkage variables occur elsewhere (Melhuish et al. 2002, King et al. 2002 and Williamson et al. 1998). Essentially, the techniques combine the rich information available in a Survey with Census data, to produce synthetic estimates for small areas of interest (Census districts or aggregates of them).

MarketInfo applies a fractional reweighting method to the ABS 1998/99 Household Expenditure Survey (HES) and the 1996 ABS Census of Population and Housing Basic Community Profile (BCP). MarketInfo 2001 is a model for the analysis of small area household expenditure, and is a joint commercial product of NATSEM and MDS Market Data Systems (refer www.natsem.canberra.edu.au /research/ marketinfo/marketinfo.html). The MarketInfo techniques have an established reputation in commercial applications – such as market identification – and have also increasingly been used in socio-economic applications (Lloyd, Given and Hellwig 2000; Lloyd, Harding and Hellwig 2001).

Synthetic CD populations can be generated using any selection of variables that are common to both the Census and survey datasets. These variables are referred to as linkage variables. There are 68 linkage variables used in MarketInfo derived from the household and personal characteristics that are common to both the HES and Census BCP. The characteristics are listed in Table 2.

The linkage variables used to derive the weights must be appropriate for the output variables of interest. The five socio-economic variables, which appear in the decision rules that assign a person to the high vulnerability class, are all either linkage variables (income, tenure type, age, household type) or considered closely associated with them (house insurance).

A set of weights (one for each of the HES survey records) has been generated by MarketInfo for each CD in the study area. The meaning of "a weight" (*w)* for a particular survey record for a particular CD is that that survey record represents *w* people or households in the CD. Hence if that survey respondent is assigned to the high vulnerability class, *w* synthetic people in the CD will be counted in that class.

Table 2 MarketInfo 2001 characteristics

Characteristics	
Age	Level of highest qualification
Total individual income	Age & income market segments[1]
Marital status	Housing type
Labour force status & gender[1]	Housing tenure
Country of birth	Household size
Occupation	Number of motor vehicles
Family type	Level of mortgage repayments
Student status	Level of rent payments

1. Cross-tabulated characteristics.
Source: MarketInfo 2001

3 Methodology and results

This third section of the paper describes the methodology used to:

- choose the socio-economic variables relevant to vulnerability
- estimate the socio-economic profiles of each CD
- estimate high vulnerability in each CD.

3.1 Socio-economic variables relevant to vulnerability

In the HES, there is a vast number of variables (with multiple classifications) providing detailed information on demographic, household, dwelling, income and expenditure characteristics. The data are captured in private dwellings and are available for a number of units of analysis including persons, families and households.

For this project, however, the selection and use of HES variables was constrained by the specification of variables and their classes in the GA questionnaire, in particular because of the unit of analysis, the individual, chosen for that questionnaire.

Matching the questionnaire's unit of analysis

The questionnaire's hypothetical cases describe an individual in terms of particular personal characteristics, but there is no information on the characteristics of other members of the household (apart from whether a partner or dependents exist). Hence the unit of analysis is persons, but not all persons – merely a random person in a residence or household. Questionnaire respondents assign a ranking to this person and the decision tree determines which of the particular characteristics contribute to that ranking. Hence the synthetic data must replicate these persons

and those characteristics in order for the decision rules to be appropriately applied to determine the small area geographic distribution of vulnerability.

In this application, it was decided to use the HES Household Reference person as a surrogate for the randomly generated hypothetical individual. The household reference person is chosen according to certain criteria (ABS 2002) for example partner or parent, highest earner or eldest person. The characteristics of household reference people are unlikely to match those of a randomly generated person. A future study could use a comprehensive unit of analysis, such as all persons or all households (rather than a random person in a household).

This choice of unit of analysis means that the synthetic results are for the number of households or household reference people in each CD as a surrogate for the number of hypothetical individuals in each CD.

Matching the questionnaire's variables and classes of interest

Often HES data has more detail than the questionnaire's classes of interest, so HES classes were merged in order to match the questionnaire classes. HES surrogates for English skills, and debt and savings were agreed as there were no exact matches. Table 3 shows the variables selected from the 1998/99 HES to match GA's characteristics of interest.

Other variables relevant to vulnerability

In addition to GA's selected household reference person characteristics, some other household level variables were identified. These were of two kinds:

- alternative measures of GA's selected variables, such as
  - equivalised household income (calculated by dividing gross household income by the square root of the number of usual residents in the household) as a potentially better measure of available income than the household reference person's income
  - 'Number Of Persons In The Household With A Disability/Health Condition' rather than whether the household reference person has a disability or restriction
- additional variables that might be relevant such as 'Principal Source of Household Income' , 'Ability Of Household To Raise Emergency Money', 'Hours Worked In All Jobs' or 'Car Insurance'.

These variables were identified as they illustrate the kind of data available in the HES that could be used in a future application of GA's four stage methodology.

Table 3 GA's characteristics of interest & HES equivalent

Characteristic	HES variable description & name
Age[1]	Age[2] (HDAGE)
Car	Number Of Registered Cars And Motor Cycles In The Household (DNCAR)
Debt & Savings[3]	Main Source Of Households Emergency Money (MSRCMNEY)
Disability	Severity Of Restriction Of The Person (HCAPP)
Employment	Labour Force Status And Status In Employment[2] (HDEMP)
English Skills [4]	Year Of Arrival In Australia[2] (HDRES)
Gender	Sex[2] (HDSEX)
Health Insurance	Hospital, medical and dental insurance (COMCOD10)
House Insurance	House Insurance[5]; Contents Insurance[5]; House & Contents Insurance[5]; (COMCOD10)
Household Type	Household Family Composition (DCOMP)
Income	Total Weekly Income (All Sources) (Gross)[2] (HDINC)
Residence Type	Dwelling Structure (DWSTR)
Tenure	Nature Of Housing Occupancy (DTENU)

1. Inexact match as HES classified; 2. Characteristic of the Household Reference Person; 3. Surrogate for quick access to money; 4. Surrogate for familiarity with surrounds; 5. Separable (selected dwelling);

Note: Shaded rows indicate the five socio-economic variables, which appear in the decision rules that assign a person to the high vulnerability class.

Source: ABS Household Expenditure Survey 1998/1999

3.2 Estimating socio-economic profiles of each CD

Estimates of the socio-economic characteristics of each CD are not used to derive the CD level vulnerability estimates. The socio-economic profiles were generated and are described here to provide contextual background for this study and the vulnerability predictions for each CD. Creating synthetic data for each of the CDs in the study area involved three steps:

- Extract relevant HES household, person and expenditure data: The majority of the variables are in the household level data. Disability status is extracted from the person level data and insurance variables are extracted from the expenditure data.
- Extract MarketInfo 2001 weights for each of the 224 CDs in the study area.
- Generate synthetic data: The extracted data in each HES record were classified and weighted in order to determine the number of people in each CD with each characteristic.

Table 4 Synthetic data for a CD in the study area

CD 1			
Age[1]		**Household Type**[1]	
15-39 years	43.9%	Couple without children	27.6%
40-44 years	16.9%	Couple with children	44.7%
45-49 years	10.3%	Single parent	10.3%
50-54 years	8.3%	Live alone	11.0%
55-59 years	6.1%	Group household	6.6%
60-64 years	5.5%	**Insurance**[1]	
65-69 years	3.7%	No house insurance	12.7%
70-74 years	3.1%	Contents insurance only	7.5%
75 years and over	2.4%	House insurance only	7.9%
Tenure[1]		House and contents insurance	71.9%
Own	34.9%		
Mortgage	50.2%	**Income**[2]	
Government renting	0.0%	Total gross weekly income of reference person	$624.05
Private renting	14.9%	Total equivalised household gross weekly income	$536.43

1. Percent of household reference people in the CD with each characteristic.
2. Income is in average dollars earned by each household in the CD.
Source: MarketInfo 2001

Synthetic socio-economic data

The resulting synthetic data for each CD has over 70 fields (i.e. all possible classes for all variables identified in the HES). Most fields in the synthetic data contain the number of households or reference people in that variable class in that CD. Table 4 shows the synthetic data (the subset of variables used in the decision rules) for one CD as percentages of the CD population.

The data for this CD show that there is a difference in income depending on the variable used. 'Total household reference person gross weekly income' is higher than 'total equivalised household gross weekly income' which may be a better measure of available income. Similarly, synthetic counts of household reference people with a disability are lower than counts for households comprising one or more disabled members (disability is defined as a schooling, employment, moderate, severe or profound restriction). In general, more low income or disabled people are identified using the household measures. It is therefore important to consider household measures in a future study.

3.3 Estimating high vulnerability in each CD

A decision tree methodology found that five socio-economic variables (plus two variables that describe the outcome of the hazard), explain the vulnerability rankings assigned to flood victims by respondents to the questionnaire. Synthetic estimation was used to predict the number of people in each CD in the high vulnerability class based on the rules determined by the decision tree for flood victims. This vulnerability information will be integrated with a flood hazard model, currently being developed at GA, in order to perform a risk assessment.

Personal injury and residence damage

The decision tree analysis of the questionnaire data revealed that personal injury and residence damage were the most important factors in determining an individual's vulnerability to flood. Injury and damage, which describe the outcome of the flood, were more important than any of the socio-economic characteristics. Consequently, no spatial assessment of vulnerability based on socio-economic characteristics alone will successfully capture vulnerability to a hazard. Rather, an assessment based on different injury and damage scenarios is necessary. Table 5 shows the different injury and damage scales.

Scenarios, nodes and socio-economic variables

There are twenty logically possible scenarios based on the four injury and five damage classes, however the decision rules produced by the decision tree have distinguished only seven actual scenarios. Table 6 shows the scenarios and their associated injury and damage scores, as well as the nodes that are populated in the scenario, and the socio-economic variables used in the rules.

In this case study, a particular scenario and its associated injury and damage scores are assumed to affect the whole of every CD in the study area whereas in an actual event, different buildings and people will be affected to differing degrees. Unlike the injury and damage values, the five socio-economic characteristics used in the decision tree rules (house insurance, income, tenure type, age and household type) vary across CDs. In most cases, injury, damage and one demographic characteristic (age) determine the node. At most, three socio-economic characteristics are required to distinguish between nodes, for example for node 9 in scenario 3 (injury is 1 or 2, damage is 5), age, home insurance and tenure are required.

Table 5 Injury and residence damage scales

Damage code	Residence damage description	Injury code	Injury description
1	not damaged	1	no injuries
2	slightly damaged	2	requires basic medical treatment, no hospitalisation
3	moderately damaged	3	requires hospitalisation and is expected to recover
4	extensively damaged	4	requires hospitalisation with life threatening injuries
5	completely destroyed		

Source: Geoscience Australia (modified from HAZUS Technical Manual, FEMA 1999)

Table 6 Scenarios, nodes and socio-economic variables

Scenarios based on injury and damage	Injury	Damage	Nodes[1]	Socio-economic variables[2]
Scenario 1 (inj2dam3)	1 or 2	1, 2 or 3	1-6	age, tenure, family type
Scenario 2 (inj2dam4)	1 or 2	4	7-10,14	age, insurance, tenure
Scenario 3 (inj2dam5)	1 or 2	5	7-9,11-14	age, insurance, tenure, income
Scenario 4 (inj3dam2)	3	1 or 2	15,19,21	age
Scenario 5 (inj3dam3)	3	3	15,20,21	age
Scenario 6 (inj3dam4)	3	4 or 5	16,17,18,20,21	age, family type, insurance
Scenario 7 (inj4dam0)	4	1, 2, 3, 4 or 5	22	none (based on injury alone)

1. The low vulnerability nodes are 1, 2, 3, 4, 7, 8, 10, 13, 15, 16 and 19.
The high vulnerability nodes are 5, 6, 9, 11, 12, 14, 17, 18, 20, 21 and 22.
2. One or more of these variables determine a particular node in a particular scenario.

Applying the decision rules

Creating synthetic vulnerability data for each of the CDs in the study area involved the following:

- Classify each of the HES households (6892 of them) according to the prevailing injury/damage scenario and each household reference person's characteristics, into one or other of the 22 nodes defined by the decision tree rules and then into the appropriate high or low vulnerability classes.
- Apply the MarketInfo small area weights to determine the number of household reference persons in the high vulnerability class in each CD.

Estimated vulnerability data

The resulting synthetic vulnerability data contains for each CD, the number and percent of household reference persons in the high vulnerability class for each of the seven identified scenarios. Table 7 provides descriptive statistics on those data.

For each level of injury, the average percent of household reference people in each CD that is affected, increases as residence damage increases. For example, for injury level 2: the mean percentage of people deemed highly vulnerable increases from 3.3% (when damage is 1, 2 or 3) to 32.4% (when damage is 5); the mean count of people deemed highly vulnerable increases from 8.2 (when damage is 1, 2 or 3) to 78.8 (when damage is 5). For injury level 3: the mean percentage of people deemed highly vulnerable increases from 15.5% (when damage is 1 or 2) to 76.4% (when damage is 4 or 5); the mean count of people deemed highly vulnerable increases from 38.3 (when damage is 1 or 2) to 187.1 (when damage is 4 or 5).

In addition, table 7 shows that in most scenarios for a particular injury level, the range of the data (that is, the difference between the minimum and maximum values) increases as the damage increases. For example, for injury level 2: the range in percentage of people deemed highly vulnerable increases from 15% (when damage is 1, 2 or 3) to 65% (when damage is 5); the range in count of people deemed highly vulnerable increases from 62 (when damage is 1, 2 or 3) to 424 (when damage is 5). For injury level 3: the range in percentage of people deemed highly vulnerable increases from 49% (when damage is 1 or 2) to 62% (when damage is 4), but then drops back down to 39% (when damage is 5); the range in count of people deemed highly vulnerable increases from 316 (when damage is 1 or 2) to 718 (when damage is 4), and then to 1052 (when damage is 5).

A map was produced for each scenario that shows the estimated percentages and counts of the highly vulnerable in each CD. A common legend was used for the percentage maps so they can be readily compared across scenarios. It is advisable to bear in mind that CDs with particularly high *percentages* of the vulnerable may have very small *counts* of the vulnerable because the population of the CD is small.

Figures 1 and 2 show the spatial estimates for Scenario 3, as counts and percentages respectively. In this scenario, the hypothetical individual does not require hospitalisation (they either have no injuries or require only basic medical treatment), but their residence is completely destroyed. The socio-economic variables that determine vulnerability are one or more of age, insurance, tenure and income. Figure 1 classifies counts in equal area classes, that is, the class intervals are set to ensure that a similar number of CDs occur in each class. Figure 2 classifies percentages in equal interval classes, that is, the class thresholds are every 20 percentage points.

Table 7 Descriptive statistics: Highly vulnerable household reference people in each CD

Variable	Number/percent	N	Mean	Std Dev	Minimum	Maximum
Num. HHs/ref. people	n/a	224	247.2	114.3	21	1403
Scenario 1 (inj2dam3)	Number	224	8.2	8.8	0	62
Scenario 1 (inj2dam3)	Percent	224	3.3	2.9	0	15
Scenario 2 (inj2dam4)	Number	224	31.3	24.3	1	202
Scenario 2 (inj2dam4)	Percent	224	12.8	7.6	1	46
Scenario 3 (inj2dam5)	Number	224	78.8	40.7	14	438
Scenario 3 (inj2dam5)	Percent	224	32.4	9.3	11	76
Scenario 4 (inj3dam2)	Number	224	38.3	28.9	4	320
Scenario 4 (inj3dam2)	Percent	224	15.5	7.1	2	51
Scenario 5 (inj3dam3)	Number	224	122.5	62.5	9	727
Scenario 5 (inj3dam3)	Percent	224	50.3	12.5	15	77
Scenario 6 (inj3dam4)	Number	224	187.1	85.7	18	1070
Scenario 6 (inj3dam4)	Percent	224	76.4	8.2	55	94
Scenario 7 (inj4dam0)	Number	224	247.2	114.3	21	1403
Scenario 7 (inj4dam0)	Percent	224	100.0	0.0	100	100

Source: MarketInfo 2001

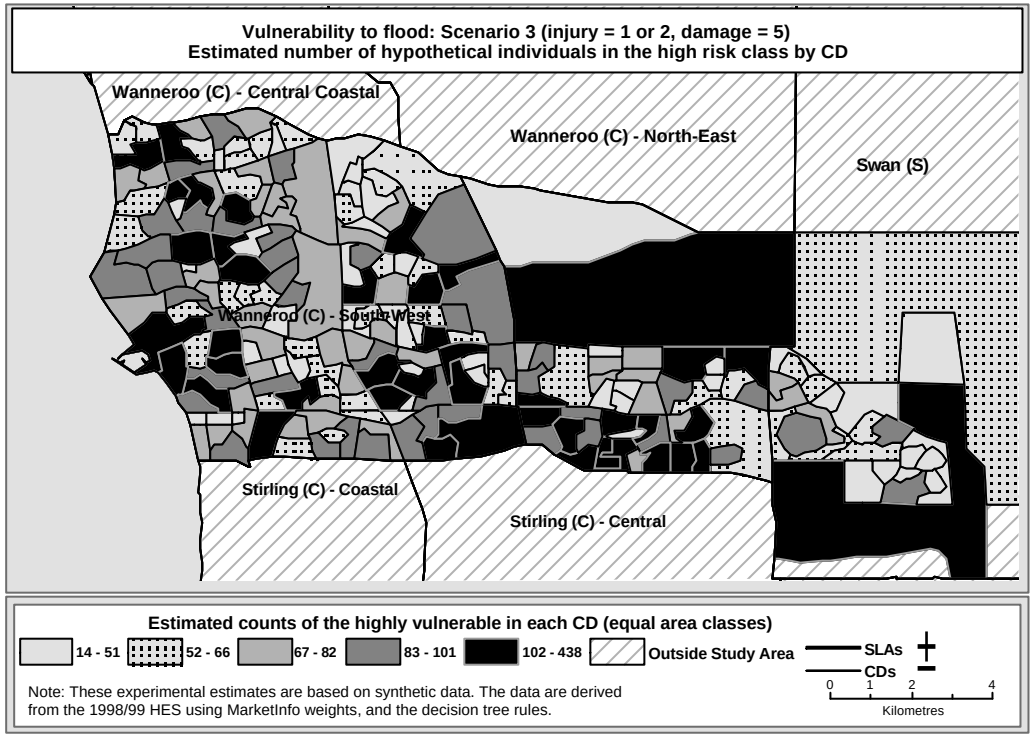

Figure 1. Experimental estimates (counts) for Scenario 3 (Source: MarketInfo 2001)

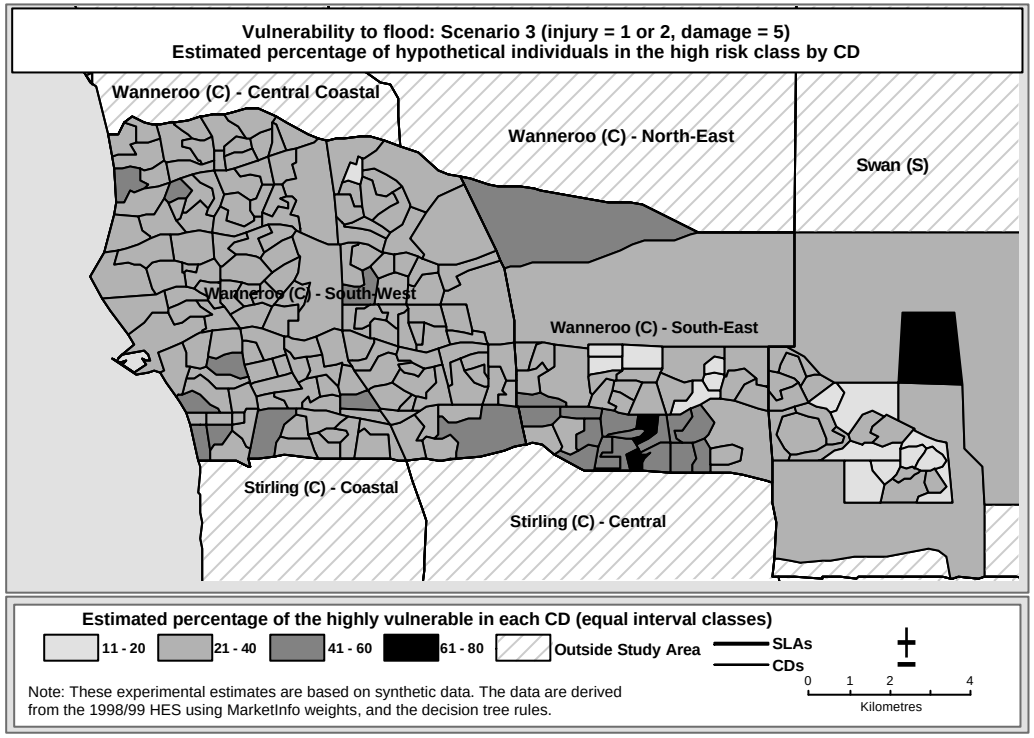

Figure 2. Experimental estimates (percentages) for Scenario 3 (Source: MarketInfo 2001)

The number of households in each of the 224 CDs in the study area ranges from 21 to 1403, with an average of 247 households in a CD. The number of hypothetical individuals in any one CD that is estimated to be vulnerable in scenario 3 ranges from 14 to 438 (refer Figure 1). In some cases this constitutes a substantial proportion of a CD (refer Figure 2).

The minimum percentage of hypothetical individuals (in a CD) that are considered vulnerable in scenario 3 is 11%. All CDs with fewer than 20% of hypothetical individuals estimated to be vulnerable are average in size or within two standard deviations of the mean number of households. The maximum percentage of hypothetical individuals in a CD that are considered vulnerable under this scenario is 76%. Two of the three CDs where more than 60% of hypothetical individuals are estimated to be vulnerable, are smaller than average (fewer than 190 households). Of those CDs where between 40% and 60% of residents are estimated to be vulnerable, a few are larger than average, so a sizeable number of people are affected.

4 Assessment

The synthetic estimation technique usually has the goal of creating new data that does not currently exist - in this project this new data is an estimate of vulnerability at the CD level. As the *true* vulnerability at CD level is unknown, the synthetic CD data can be assessed in the following ways:

1. against expert local knowledge
2. by comparison with available ABS data
3. in terms of the appropriateness of the method and its data inputs for prediction of vulnerability
4. by comparison with results from an alternative method.

These four possible avenues of assessment are discussed briefly below as possible areas of useful further work.

4.1 Assessment against expert local knowledge

Through GA's contacts in the Perth Cities Project there will be an opportunity to compare the mapped vulnerability outputs with what is known on the ground by local experts in the emergency planning and community services fields. At the time of writing, this validation step had not been completed.

4.2 Assessment by comparison with ABS data

It may be possible to purchase cross-tabulations of most of the decision rule variables from the ABS (where the variables are available from the Census) to assess similar cross-tabulations in the synthetic data. Comparison of the cross-tabulations could be at SLA or CD level. While it would be better to validate at CD level than SLA level, confidentiality issues may prevent ABS data being made available for CDs. In addition, the ABS cross-tabulations would be based on the 1996 or 2001 census, whereas the synthetic estimates are based on existing weights (derived from the 1996 census and uprated to 2001 prior to the release of the 2001 Census), so any differences due to the dates of the data and the geographies on which they are based might be difficult to reconcile.

4.3 Assessment of appropriateness of method and data

The data are crucial to the final result. If reformulated variables such as low household income and presence in the household of people with a disability are better measures for indicators and therefore better predictors of vulnerability, then such variables, had they been included in the questionnaire, might have been ranked differently by respondents and been assigned greater importance in the decision tree.

The following issues and options are raised, assuming that synthetic estimation is to be used to derive the small area estimates in the fourth stage of GA's methodology. This means that some kind of survey is to be reweighted, such as the HES. Following are the issues identified in the four-stage methodology that affect the reliability of the synthetic estimates.

Firstly, meaningful measures of important indicators for the appropriate population are critical:

- explore the survey that is to be reweighted for best available measures of variables as well as additional variables, so that they can be assessed for importance and included in the questionnaire
- ensure a comprehensive unit of analysis such as all persons or all households - if household is the chosen unit of analysis, collect sufficient data on the household and its members
- determine if and how non-private dwellings should be considered.

Secondly, in applying the relationship between socio-economic characteristics and vulnerability at the small area level

- there is an assumption that the profile of households in a small area captures important neighbourhood attributes of that small area and that it is valid to apply an a-spatial relationship in different small areas

- it is important to ensure that linkage variables are important drivers of, surrogates for, or closely related to the synthetically estimated output variables (for example, if disability is an important indicator, the linkage variables must be closely associated with disability for estimates to be valid)
- consider surveys other than the HES for regionalising at CD or SLA level (including other ABS surveys or a customised survey meeting GA and synthetic estimation requirements)

Synthetic estimation is a powerful means of estimating the small area distribution of variables of interest. The relationships between linkage variables and the variables to be estimated must be assessed, to determine the appropriateness of synthetic estimation. For example, home insurance is likely to be closely related to income and tenure, both of which are linkage variables, however the strength of such a relationship should be confirmed. This will provide a level of confidence about the small area estimates of important variables.

Synthetic estimation was used in this project to estimate the socio-economic profiles of each CD as well as to derive vulnerability to flood. It would be interesting to examine the synthetic estimates of selected variables, which were not used in the decision rules, to see if their relationship with the synthetic estimates of vulnerability accords with expectation. This comparison could be performed on particular CDs (of a reasonable size) that consistently appear in the high or low vulnerability class across scenarios. Such an assessment might provide insights into which variables should be considered in future studies.

4.4 Assessment by comparison with results from alternative methods

An alternative method or model could be used to produce vulnerability estimates that could be compared with the results produced using this four stage methodology. The synthetic socio-economic data or other CD level data could be used as inputs to an alternative method in order to produce vulnerability data at CD level. It is fair to expect that any differences in results from the different methods can be explained or may need to be explored further.

The earlier Cities Project methodology (Granger and Hayne (eds) 2001) could be used to produce CD level estimates of vulnerability, which could be compared with the synthetic estimates in order to investigate differences in overall levels of vulnerability. Alternatively a simple spatial overlay of variables considered to be important drivers, while subjective, may result in some insights.

5 CONCLUSION

The aims of the project were to create experimental estimates of the geographic distribution of vulnerability in 224 Census Collection Districts in Perth and to assess the role of synthetic estimation in a methodology aiming to quantify vulnerability to natural hazards. Due to the issues identified, we see the synthetic vulnerability outputs as powerful examples of the kind of small area data that can be produced using synthetic estimation rather than as authoritative estimates of vulnerability in their own right. Nevertheless, this first attempt by GA and NATSEM to quantify vulnerability to natural hazards is believed to be a promising pilot of a new methodology that allows spatial estimation of characteristics relevant to natural hazard response.

ACKNOWLEDGMENTS

The authors would like to acknowledge the earlier work at NATSEM of Dr Otto Hellwig (now of MDS Market Data Systems) in the development of the MarketInfo techniques. They would like to thank Christopher Zoppou and Ole Nielsen for their assistance in the decision tree analysis and Elizabeth Taylor for preparing the maps.

NOTE

NATSEM research findings are generally based on estimated characteristics of the population. Such estimates are usually derived from the application of microsimulation modelling techniques to microdata based on sample surveys. These estimates may be different from the actual characteristics of the population because of sampling and nonsampling errors in the microdata and because of the assumptions underlying the modelling techniques.
The microdata do not contain any information that enables identification of the individuals or families to which they refer.

REFERENCES

Australian Bureau of Statistics (2002) Catalogue No. 6544.0.30.001. *1998-99 Household Expenditure Survey Australia Confidentialised Unit Record File (CURF) Technical Paper.* Second edition (incl. Fiscal Incidence Study) August 2002.

Dwyer, A., Zoppou, C., Nielsen, O., Roberts, S. and Day, S.D. (2003) *Quantifying vulnerability: a methodology for identifying those at risk to natural hazards.* Draft version May 2003. Geoscience Australia, Canberra.

FEMA (Federal Emergency Management Authority) (1999) *HAZUS Technical Manual.* USA Government. Washington.

Granger, K. and Hayne, M. (eds) (2001) *Natural hazards and the risks they pose to South-East Queensland: comprehensive report.* Geoscience Australia in conjunction with the Bureau of Meteorology, Canberra.

Harding, A, Lloyd, R, Hellwig O and Bailey, G. (2000) *Building the Profile: Report of the Population Research Phase of the ACT Poverty Project, Poverty Task Group Paper No. 3*, ACT Government, Canberra, December.

Hellwig, O. and Lloyd, R. (2000) *Sociodemographic Barriers to Utilisation and Participation in Telecommunications Services and their Regional Distribution: a Quantitative Analysis*, Report commissioned by Telstra, Canberra.

King, A., McLellan, J., and Lloyd, R. (2002) *Regional Microsimulation for Improved Service Delivery in Australia: Centrelink's CuSP Model.* Paper Prepared for the 27th General Conference, International Association for Research in Income and Wealth Stockholm, Sweden, 18-24 August 2002.

Lloyd, R., Given, J. and Hellwig, O. (2000) The Digital Divide: Some Explanations, *Agenda*, 7(4), pp. 354-58.

Lloyd, R., Harding, A. and Greenwell, H. (2001) *Worlds Apart: Postcodes with the Highest and Lowest Poverty Rates in Today's Australia*, Paper prepared for the National Social Policy Conference 2001, Sydney, July.

Lloyd, R, Harding, A and Hellwig, O. (2001) Regional Divide? A Study of Incomes in Regional Australia, *Australasian Journal of Regional Studies*, 6(3), pp. 271-292.

Melhuish, T., Blake, M. and Day, S.D. (2002) An Evaluation of Synthetic Household Populations for Census Collection Districts Created Using Optimisation Techniques. *Australasian Journal of Regional Studies*, 8(3).

Williamson, P., Birkin, M. and Rees, P.H. (1998) The estimation of population microdata by using data from small area statistics and samples of anonymised records. *Environment and Planning*, vol. 30, pp. 785-816.