DigitalCommons @WayneState

Journal of Modern Applied Statistical Methods

Volume 8 | Issue 2 Article 29

11-1-2009

Generating and Comparing Aggregate Variables for Use Across Datasets in Multilevel Analysis

James Chowhan
McMaster University, Ontario, Canada, chowhan@mcmaster.ca

Laura Duncan
McMaster University, Ontario, Canada, duncanlj@mcmaster.ca

Recommended Citation

Chowhan, James and Duncan, Laura (2009) "Generating and Comparing Aggregate Variables for Use Across Datasets in Multilevel Analysis," *Journal of Modern Applied Statistical Methods*: Vol. 8: Iss. 2, Article 29.
Available at: http://digitalcommons.wayne.edu/jmasm/vol8/iss2/29

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 613-625

1538 – 9472/09/$95.00

Generating and Comparing Aggregate Variables for Use Across Datasets in Multilevel Analysis

James Chowhan Laura Duncan
McMaster University,
Ontario, Canada

This article examines the creation of contextual aggregate variables from one dataset for use with another dataset in multilevel analysis. The process of generating aggregate variables and methods of assessing the validity of the constructed aggregates are presented, together with the difficulties that this approach presents.

Key words: Aggregate variables, contextual variables, multilevel analysis.

Introduction

Contextual effects influence individual outcomes and behaviors. The importance of including community level variables has been gaining ground in the social sciences. Despite their popularity and the presence of theory corroborating the existence of contextual effects, proper measurement and selection of contextual variables continues to challenge researchers. Furthermore, researchers often face the additional difficulty presented by surveys that are not designed to contain contextual information at the geographic area of interest. Even when data is available at the appropriate geographic level, a deficiency of individuals in each area may prohibit the calculation of reliable estimates in multilevel models and thus make it difficult to successfully model contextual effects. A suitable approach to address these difficulties might be to construct aggregate variables in one dataset that has sufficient sample size in the area of interest for use with other datasets.

Conducting multilevel analyses requires contextual information at the level of interest, for example, family, household, neighborhood, province or country. Datasets are selected by researchers based on their ability to provide answers to research questions and the presence of key variables of interest at the level of interest. In many cases, datasets do not contain the contextual information at the required level for a multilevel analysis. In such cases, researchers could turn to another dataset to construct the desired measure and match this information, using geographical or other identifiers, to their original dataset.

Despite the apparent simplicity of this approach, the issue of checking aggregate variables must be addressed. If possible, the aggregate variables should be checked in some way to assess their validity (i.e., do they measure what they are supposed to measure?). One possible way of checking aggregate variables for validity is presented here, together with problems encountered during the process. These are presented as a means of highlighting some of the hidden complexities of creating aggregate variables that researchers should take into consideration when using this approach.

Methodology

The Survey of Labor and Income Dynamics (SLID) is a longitudinal survey on labor market activity and income. The survey follows individuals with yearly questionnaires

James Chowhan is a Ph.D. Student in the DeGroote School of Business. Email him at: chowhan@mcmaster.ca. Laura Duncan is a Research Coordinator in Psychiatry and Behavioural Neurosciences in the Offord Centre for Child Studies. Email her at: duncanlj@mcmaster.ca.

administered for six consecutive years, with a new wave starting every three years since the survey's 1993 initiation. The SLID contains variables that may be used to construct numerous interesting and relevant Economic Region (ER) level variables; thus, researchers could use the following procedures to construct ER level variables of their own choosing. For this example ten aggregate variables on employment and education were constructed; these variables were selected for their potential value to researchers for use in conjunction with other datasets. Table 1 contains the variable names, definitions and the original SLID variables from which they were constructed.

Creation of an Analytic SLID Dataset

First, the variables that were used to create the aggregates (shown in Table 1), were extracted from the SLID together with the appropriate cross-sectional weights and individual and geographical identifiers for each survey year using the SLIDret program. SLID identifies ERs by two separate variables: erres25 and xerres25. The explanation for the presence of two identifiers instead of one is that Statistics Canada amended their ER identification codes in 1999, thus, the SLID contains two sets of ER identification codes. One code refers to the 1991 Census boundaries for all survey years of the SLID (xerres25), and the other refers to the 1991 Census boundaries up to 1999 and to the amended 1999 Census boundaries in subsequent years (erres25). Researchers must decide upon the most appropriate variable to use in any particular research scenario. This will often be determined by the geographical code used in the dataset in which the constructed aggregate variables will be used.

Table 1: Defining Variables of Interest

Variable Name	Definition	SLID Variables
non-employee	Proportion of total labor force self employed	clwrkr1
non-employee_f	Proportion of female labor force self employed	clwrkr1
pct_mgt	Proportion of occupations perceived to be managerial	manag1
pct_mgt_f	Proportion of female occupations perceived to be managerial	manag1
less_hs	Proportion of individuals with less than a high school education	hlev2g18
hs	Proportion of individuals with at least a high school education	hlev2g18
non_univ_ps	Proportion of individuals with a non-university post-secondary certificate	hlev2g18
uni_ps	Proportion of individuals with a university post-secondary certificate	hlev2g18
ps	Proportion of individuals with a post-secondary certificate	hlev2g18
ps(_f)	Proportion of females with a post-secondary certificate	hlev2g18

This article compares the constructed aggregate variables and both the 1996 and 2001 Census profile data. Because the 1996 Census profile data uses the pre-1999 Census boundaries, the xerres25 variable was used to calculate the 1996 SLID ER level estimates. Similarly, because the 2001 Census profile data uses the post-1999 Census boundaries, the erres25 variable was used for the 2001 SLID ER level estimates.

Construction of Aggregate SLID Variables

After creating a SLID dataset, the ER aggregate variables can be constructed. Ten aggregate SLID variables were constructed, seven for the entire population and three for the female population only. The approach was to create a count of individuals in each ER possessing the characteristics of interest and to use this count to construct weighted proportions aggregated at the ER level that could then be exported for use with other datasets.

For each characteristic of interest individuals with that characteristic are dummy coded as 1. This results in dummy variables for individuals aged 15 to 69 who are self employed, individuals aged 15 to 69 whose occupations are perceived as managerial, individuals aged 16 and over who have less than a high school education, individuals aged 16 and over who have at least a high school education, individuals aged 16 and over who have a non-university post-secondary certificate, individuals aged 16 and over who have a university post-secondary certificate, and individuals aged 16 and over who have a post-secondary certificate (university or non-university). There was also a dummy variable for gender so dummy variables could be created for females, for females aged 15 to 69 who are self employed, females aged 15 to 69 whose occupations are perceived as managerial, and females aged 16 and over who have a post-secondary certificate (university or non-university).

Aggregating SLID to the ER Level

After creating SLID dummy variables; the final step was to aggregate these variables. In all cases these aggregates will be proportions for each ER created by aggregating up to the ER level. Because the SLID produces an annual cross-section of individuals it is also necessary to aggregate to the ER level by survey year in order to obtain an accurate estimate of area level characteristics for each year. Taking the mean of a dummy variable is one way to calculate the proportion of individuals with a certain characteristic. Hence, proportions for each ER are calculated by collapsing the dummy variables to their mean for each ER level and for each survey year. These proportions are weighted using the cross-sectional weight. The resulting aggregate variables represent proportions of individuals in ERs with the characteristics of interest outlined.

Once created, aggregates are ready for use; however, it is highly recommend that a check be carried out to assess their validity as aggregate measures. This is accomplished in the following example by comparing the provincial and national population totals followed by the basic gender and age characteristics of the samples. The logic being that, if the population totals are similar and sample characteristics are similar across these demographics, there is some reason to assume that they will be similar in other ways. It is not guaranteed that this is actually the case, however.

As an additional check, similar education and employment aggregates constructed using the Census profile data from 1996 and 2001 were compared as well. (This will not be an option readily available to researchers if one of the main reasons for going to another dataset is that the variables of interest are not available in the Census profile data.) These comparisons are recommended because they will alert a researcher to oddities about the variables or dataset used and to inconsistencies that may require investigation.

To assess the validity of the aggregate SLID variables constructed, a comparison was made to the 1996 and 2001 Census. The Census is, by design, the most accurate and representative approximation of true population parameters. In order for the SLID aggregates to be useful they should reflect true population parameters. It may be argued that using the Census to verify how closely the SLID data and constructed aggregates reflect the true population is the most suitable method of comparison available. As the SLID weighting is calibrated on Census population totals, it is expected that estimates will match well. The following is a step-by-step guide to comparing aggregate variables.

Choose a Method of Comparison

Two methods of comparison were used in this example. The first involved simply calculating and comparing provincial and national population totals for both SLID datasets for 1996 and 2001. If no similarity existed at this level it would not be sensible to continue with the comparison and the validity of the aggregate variables would be questionable. The second method of comparison used confidence intervals as a means of statistically assessing how close the estimates match. This requires similarly defined variables to be created using Census profile data so that aggregates are created from the SLID and the Census profile data at the ER level. Confidence intervals (assuming a Normal distribution) can be created around the SLID estimates and observations made as to whether the population estimates from the Census fall within these confidence intervals for each ER. The confidence level chosen for this example is 95% but researchers can choose any level they think is suitable. A high number of matches show the SLID estimates are a good match to true population parameters.

Choose and Generate Demographic Variables and Confidence Intervals

For the provincial and national population totals, weighted sums were calculated in STATA broken out by province. At the ER level, two characteristics were chosen for comparison: gender and age. Twenty-one age and gender breakouts by ER were calculated using the SLID data for 1996 and 2001. In addition to the proportion of females, age breakouts for the whole population and for females only are generated using different age intervals. Using STATA, 95% confidence intervals were created for each SLID estimate.

Recreate Aggregate Variables Using Census Profile Data

Because the Census profile data does not contain ER identification codes it is first necessary to merge the Census data with the Postal Code Conversion File (PCCF), matching the data by Enumeration Area (EA) for 1996 and Dissemination Area (DA) for 2001. Enumeration areas (EA) in the 1996 Census and Dissemination areas (DA) in the 2001 Census are smaller geographical areas making up various larger Statistics Canada geographical areas, including ERs. The 1999 change in Census boundaries lead to a name and definition change from EA to DA (for more information on using the PCCF and the change from EA to DA see Gonthier, et al., 2006).

For 1996 the EA code is an eight-digit code constructed from provincial, federal and EA identifiers. The provincial code composes the first two digits; the federal code the following three and the EA code the final three. To construct the eight-digit EA code from its composite parts the provincial code is multiplied by 1,000,000 and the federal code is multiplied by 1,000, and these numbers are added to the EA code. The Census data is then merged with the 1996 PCCF file using this eight-digit EA identifier.

For 2001 the DA code is an eight-digit code constructed from provincial, census division and DA identifiers. The provincial code composes the first two digits, the census division code the following two and the DA code the final four. The eight-digit DA code for 2001 is created the same way as the EA code for 1996. Merging results in each record being assigned an ER identification code. Once again, to ensure the production of accurate estimates, data is aggregated to the ER level by first creating a sum of all individuals within ERs with the characteristics of interest. This ensures accurately weighted estimates reflecting the numbers of individuals in ERs. After these sums are created for each ER proportions are then calculated that correspond to the ten aggregate variables created in the SLID. Table 2 shows the variable names and the 1996 and 2001 Census variables from which they were constructed.

Compare SLID and Census Profile Data Estimates

With the SLID education and employment aggregates, population totals, demographic variables, confidence intervals for these estimates for 1996 and 2001 and similar

Table 2: Concordance between 1996 and 2001 Census Variables

Variable Name	Definition	1996 Census Variable Range Used	2001 Census Variable Range Used
non_employee	Proportion of total labor force self employed	v1211-v1222	v949-v960
non_employee_f	Proportion of female labor force self employed	v1235-v1246	v973-v984
pct_mgt	Proportion of occupations perceived to be managerial	v1031-v1090	v985-v1044
pct_mgt_f	Proportion of female occupations perceived to be managerial	v1151-v1210	v1105-v1164
less_hs	Proportion of individuals with less than a high school education	v1338-v1351	v1382-v1395
hs	Proportion of individuals with at least a high school education	v1338-v1351	v1382-v1395
non_univ_ps	Proportion of individuals with a non-university post-secondary certificate	v1338-v1351	v1382-v1395
uni_ps	Proportion of individuals with a university post-secondary certificate	v1338-v1351	v1382-v1395
ps	Proportion of individuals with a post-secondary certificate	v1352-v1375	v1358-v1381
ps_f	Proportion of females with a post-secondary certificate	v1352-v1375	v1358-v1381

variables recreated using Census profile data from 1996 and 2001, the comparison was carried out. First, weighted provincial and national population totals were compared by year and by province; results are shown in Tables 3a and 3b.

It is important to note that some variation in the totals is to be expected due to rounding error in the Census. In both tables it was expected that column 1 and 2 add up to column 3. In 1996, there was a difference of 685 and in 2001 there is a difference of 235. These differences are likely due to rounding error. It would also be expected that column 4 and column 6 would be similar and that column 5 would be less than both of these. In 1996 the total population is SLID is 271,963 more than the Census total population and in 2001, the total population in the SLID is 1,828,145 below the Census total population: no obvious reason exists to explain this. Even with the minor disparity, population totals in the SLID are close enough to the Census to conclude that the data matches reasonably well.

Second, basic demographics were compared by year and by ER in order to determine the number of Census profile estimates that would fall within the 95% confidence intervals generated around the SLID estimates. Each Census profile estimate falling within the confidence interval was called a match. Table 4 shows the percentage of matches across 66 ERs in 1996 and 73 ERs in 2001.

The proportion of females in the population variable matched perfectly and the age breakouts had a high, but not perfect, percentage of matches. The only variables with suspiciously low numbers of matches were the percentage of individuals aged 15 to 19 and the percentage of females aged 15 to 19. The age breakouts for individuals and females aged 15 to 25 and 20 to 24 showed much better matching. This suggests that the discrepancy is occurring at the lower end of the age spectrum in the 15 to 19 age range.

Based on observations of similarities in the population totals, gender and age characteristics across the SLID and Census profile data samples, it may be suggested that the SLID and the Census profile data will also be similar across other characteristics, in this case education and employment. To test this, the constructed aggregates were checked for validity in a similar manner. Again, 95% confidence intervals (assuming a Normal distribution) were created around the SLID estimates and observations were made as to whether the population estimates from the Census fell within these confidence intervals for each ER. Table 5 shows the percentage of matches across 66 ERs in 1996 and 73 in 2001.

Given the excellent age and gender match of the data, the low number of matches for the constructed aggregate variables is surprising. Without a clear explanation as to why the variables do not match, the constructed aggregates cannot be trusted as representative and should not be used. However, if explanations can be found for the low matching then the aggregates may be of some use. An investigation of the data and variable definitions was carried out to identify possible causes for the low number of matches.

Investigation of the data and examination of the documentation highlighted several limitations with the variables chosen for use in both the Census profile data and the SLID. These limitations are very likely the cause of the low number of matches across the aggregate variables. First, the internal consistency of the constructed estimates was investigated. In particular, confirmation was required that the total populations being used on the SLID and in the Census Profile data as the denominator in the proportions calculations were in fact the sum of their composite parts. In both the SLID and the Census Profile data, age and education populations were verified. A check was made of the proportions of individuals aged under 25, 25 to 49, 50 to 74 and 75; these proportions should total 1 as this range of ages encompasses all possible ages in the population. The same check was carried out for the female proportions and for the proportions of individuals aged under 25, 25 to 49, 50 to 64 and

Table 3a: Provincial and National Totals for SLID and Census Profile Data, 1996

1996	Census					SLID
Province	1. Male Subtotal	2. Female Subtotal	3. Total Population	4. Total Population 15+	5. Total Labor Force 15+	6. Total Population 15+
10	271,740	278,575	550,420	435,985	245,165	423,747
11	65,990	68,450	134,440	103,580	70,695	100,100
12	441,490	466,175	907,635	718,015	438,010	669,414
13	362,490	374,665	737,255	583,550	363,055	556,031
24	3,318,800	3,462,665	6,781,570	5,382,325	3,357,080	5,394,101
35	4,794,345	5,011,300	9,805,685	7,669,850	5,084,190	7,848,826
46	492,640	509,980	1,002,730	769,900	511,145	782,124
47	450,690	461,550	912,085	689,015	463,360	687,939
48	1,225,800	1,227,510	2,453,330	1,864,640	1,348,880	1,895,376
59	1,706,985	1,751,340	3,458,715	2,743,105	1,819,185	2,874,269
Total	13,130,970	13,612,210	26,743,865	20,959,965	13,700,765	21,231,928

Table 3b: Provincial and National Totals for SLID and Census Profile Data, 2001

2001	Census					SLID
Province	1. Male Subtotal	2. Female Subtotal	3. Total Population	4. Total Population 15+	5. Total Labor Force 15+	6. Total Population 15+
10	249,805	260,815	510,545	422,170	240,600	404,336
11	65,450	69,145	134,530	107,940	73,570	98,323
12	437,335	466,330	903,505	739,060	450,075	681,910
13	355,380	371,485	726,990	597,500	370,920	548,849
24	3,521,985	3,689,680	7,212,255	5,923,010	3,734,615	5,270,975
35	5,458,005	5,701,920	11,159,880	8,972,500	5,950,800	8,426,920
46	547,455	567,110	1,114,400	881,395	582,590	796,246
47	478,785	494,380	973,075	766,390	509,670	691,965
48	1,470,895	1,473,690	2,944,620	2,334,465	1,678,965	2,193,306
59	1,908,975	1,978,245	3,887,305	3,183,715	2,046,190	2,994,323
Total	14,494,070	15,072,800	29,567,105	23,928,145	15,637,995	22,100,000

Table 4: Comparison of SLID and Census Profile Aggregate Estimates for Gender and Age Variables

Variable Name	Variable Definition	1996	2001
		% of Matches	% of Matches
female	% population that is female	100	97
pct_15to25	% population aged 15 to 25	71	70
pct_25to49	% population aged 25 to 49	79	78
pct_50to74	% population aged 50 to 74	70	84
pct_75over	% population aged 75 & over	62	78
pct_50to64	% population aged 50 to 64	68	79
pct_65over	% population aged 65 & over	70	74
pct_15to19	% population aged 15 to 19	44	52
pct_20to24	% population aged 20 to 24	80	86
pct_40to44	% population aged 40 to 44	90	88
pct_75to79	% population aged 75 to 79	74	89
pct_15to25_f	% female population aged 15 to 25	68	74
pct_25to49_f	% female population aged 25 to 49	85	88
pct_50to74_f	% female population aged 50 to 74	76	86
pct_75over_f	% female population aged 75 & over	73	88
pct_50to64_f	% female population aged 50 to 64	79	86
pct_65over_f	% female population aged 65 & over	77	82
pct_15to19_f	% female population aged 15 to 19	70	66
pct_20to24_f	% female population aged 20 to 24	80	92
pct_40to44_f	% female population aged 40 to 44	85	92
pct_75to79_f	% female population aged 75 to 79	83	92

Table 5: Comparison of SLID and Census Profile Aggregate Estimates for Employment and Education Variables

Variable Name	Variable Description	% of Matches	
		1996	2001
non_employee	Proportion of total labor force self employed	44	55
non_employee_f	Proportion of female labor force self employed	61	67
pct_mgt	Proportion of occupations perceived to be managerial	20	8
pct_mgt_f	Proportion of female occupations perceived to be managerial	30	22
less_hs	Proportion of individuals with less than a high school education	33	51
hs	Proportion of individuals with at least a high school education	2	0
non_univ_ps	Proportion of individuals with a non-university post-secondary certificate	42	79
univ_ps	Proportion of individuals with a university post-secondary certificate	35	60
ps	Proportion of individuals with a post-secondary certificate	74	59
ps_f	Proportion of females with a post-secondary certificate	88	85

65 and over. It was found that the 15 to 19 age category produced low numbers of matches. Verification was made that the difference between the proportion of individuals aged under 25 and the proportion of individuals aged 15 to 19 added to the proportion of individuals aged 20 to 24 equals 0. The same verification was made for the female proportions. The results were either extremely close or exactly 0 or 1 (See Appendix 1 for details).

Additional checks were carried out for the education variables in the Census Profile data for both 1996 and 2001. The difference between the proportion of individuals with postsecondary certificates and the proportion of individuals with university certificates added to the proportion of individuals with non-university certificates was checked with the expectation that, if accurate, they should equal 0. Results were either 0 or less than 0.0001 above or below 0. The proportion of individuals with less than a high school education were added to individuals with at least a high school education with the expectation that they would equal 1. This was not the case: most totals ranged from 0.4 to 0.6. Referring to the documentation and exploring the data illuminated the reason. The Total population 15 years and over by highest level of schooling is a poorly defined population. The Census Profile data contains numerous population totals broken out by different characteristics. For example, the education variables include ‘Total population 15 years and over by highest level of schooling’, the marital status variables include ‘Total population 15 years and over by marital status’ and the labor force variables include ‘Total population 15 years and over by labor force activity’. It was expected that summing together the number of individuals aged 15 years and over using the age breakouts in the Census Profile data would include the same population as these ‘Total population 15 years and over by…’ variables. This was checked for the ‘Total population 15 years and over by highest level of schooling' variable. The difference between these two totals was more than can be explained by rounding error in the Census Profile data. Checking this variable against other variables that call themselves 'Total population 15 years and over by…' a sizeable and apparently unexplainable difference was found. Another drawback with the 2001 education aggregate variables is that in the 2001 Census Profile data education data is supplied for individuals aged 20 and over (Statistics Canada, 1999). By contrast, SLID education data was available for individuals aged 15 and over. This, added to the other problems described, provides the reason why the education aggregate variables do not match well.

Having identified an explanation for the low matching across education variables, similar explanations were sought for the employment variables. Three main limitations in both the Census Profiles and SLID documentation regarding ambiguous definitions of populations and variables were found that could explain the low number of matches across the employment variables. First, there was some ambiguity over the definition of the labor force. SLID defines the labor force as persons aged 16 to 69 who were employed during the survey reference period. Therefore, the employment variables used in the construction of the aggregate variables only refers to these individuals. The Census Profile data on the other hand defines the labor force as employed individuals aged 15 and over. Although this may cause some disparity, it is unlikely to be the only cause of the low number of matches. Further investigation revealed a more severe limitation regarding the classification of individual labor force status (Statistics Canada, 1997; Statistics Canada, 1999).

One of the strengths of the SLID as a longitudinal survey is that it asks for information on every job an individual has held during the reference year, rather than focusing on the job at the time of the survey. Regarding class of worker (paid worker, employee, self-employed, etc.) individuals can, therefore, hold several statuses, in addition, they are asked to report their status for each month so that they have 12 statuses over the year. By contrast, the Census Profile data only reports the class of worker for individuals at the time the Census is carried out. For the construction of the non_employee aggregate variable, it was necessary that individuals only fall into one class of worker category. However, the SLID uses the main job concept to categorize individuals into the class of worker variable clwrkr1. Main job is typically the job with the longest duration, greatest number of hours worked over the year, and most usual hours worked in a given month (Statistics Canada, 1997; Statistics Canada, 2007). Thus, the difference in the reference periods of the samples and the SLID's focus on main job is a possible explanation for lower matching rates.

The Census Profile data has its own ambiguities around the class of worker variable. In the Census Profile data on class of worker there is a category defined as Class of worker-Not Applicable (Statistics Canada, 1999). The documentation does not explain who this group consists of or what characteristics of individuals in this category make class of worker not applicable to them. In an effort to take this into account, the aggregate variable of non-employee was constructed using an all classes of worker variable as the denominator (a variable that does not include the class of worker-not applicable individuals). This was used in place of the variable total labor force 15 years and over by class of worker, which did include those individuals. Although this avoids using unclear population definitions as a denominator, it does not help explain where that category of individuals should be included most accurately in the class of worker categorization. These problems may explain why the non_employee variables are not matching well.

Finally, there was a difference in the definition of the Census Profile data and SLID managerial occupation variables that may render them incomparable. In the Census Profile data, individuals are asked to explain the type of job they have and their main responsibilities, and from this information they are coded into occupation classifications. This classification includes a section on management occupations that was used to produce the proportion of individuals with occupations perceived to be managerial variable (Statistics Canada, 1999). In the SLID, on the other hand, individuals are asked if they perceive their job to be managerial

(Statistics Canada, 1997). The self-identification involved here suggests that this variable is likely to be largely inconsistent: what defines a job as managerial is not clearly defined. Individuals may identify themselves as having a managerial job when in fact they do not. This could explain why the pct_mgt variable does not show a high number of matches.

Conclusion

The investigations and results are outlined here as a precaution to researchers wishing to create aggregate variables or use the SLID aggregate variables created in this study. The construction and comparison of aggregate variables should not be undertaken without caution. Despite their limitation, it is hoped that the constructed SLID aggregates could be of some use to researchers. The following points serve as a set of cautions to those wishing to use this approach based on difficulties that might be encountered in aggregate variable construction and comparison.

Internal Consistency

When constructing aggregate variables it is important that the variables are internally coherent. For example, imagine creating two aggregate education variables at the EA level; one is the proportion of individuals with less than a high school education, the other is the proportion of individuals with at least a high school education. If a researcher added the two proportions together across all EAs all totals should equal 1, if it does not, further investigation would be required to uncover reasons why.

Target Population

When creating and comparing aggregate variables it is important to know the target population of the variable. Some variables apply to individuals over a certain age, some apply to individuals who only answered positively to survey questions, and some apply to all respondents. This becomes more important if researchers wish to check the validity of their constructed aggregates by comparing the sample characteristics with the Census profile data characteristics. If constructing proportions of individuals with certain characteristics, it is important that the denominator be the same in both variables. The Census profile data documentation clearly defines its target population for each variable but it can be unclear how individuals were included. For example, in the 1996 Census profile data the ‘Total population 15 years and over by highest level of schooling’ was not the same as ‘All individuals aged 15 years and over’. In some cases, the target populations for the same variable in the 1996 and 2001 Census profile data were different. For example, in the 1996 Census profile data, the education data was available for individuals aged 15 and over; and in the 2001 Census profile data, it was supplied for individuals aged 20 and over (Statistics Canada, 1999).

Definitions

It is important to understand how variables are defined in order to construct useful aggregate variables that are as accurate as possible. What a researcher may consider to be a standard classification may in fact be different across different datasets. In the example used in this article, it was found that the definition of labor force was not clear. In SLID, labor force is defined as persons aged 16 to 69 who were employed during the survey reference period. In the Census profile data, the labor force is defined as employed individuals aged 15 and over. This difference made comparing the Census profile and SLID aggregate employment variables inappropriate. Variable definitions may also have unexplained ambiguities that must be taken into account. For example, in the Census profile data there were ambiguities with the class of worker variable, which made comparison difficult.

Survey Design

The way in which surveys are designed can make constructing and comparing aggregate variables problematic. One of the strengths of the SLID as a longitudinal survey is that it asks for information on every job an individual has held during the reference year rather than focusing on the job at the time the survey is carried out. The result is that many records may exist for one individual. By contrast, the Census profile data holds one record per individual. Researchers must be clear on the survey design and number of records per individual. If several records exist for each individual, some rationale must be used to select the most suitable record.

Classification

It is important to understand how variables have been coded into categories and whether individuals self-identify for certain classifications. This has implications for category definitions and how comparable they are across datasets. In the example provided, a difference was identified between the Census profile data and SLID in how managerial occupations are defined. The difference between coding by self-identification and coding by an external classifier is an important one that could lead to inconsistent definitions.

This article outlined the importance of aggregate level variables for use in multilevel analysis and introduced the idea of generating aggregate level variables in one dataset for use across other datasets. Generating and comparing aggregate variables was described using an example generating employment and education aggregate variables in the SLID for 1996 and 2001 cross-sectional samples at the ER Level and comparing them to similar estimates constructed using the 1996 and 2001 Census profile data. The difficulties encountered resulted in a set of cautions for researchers wishing to use this approach. As a whole, this article may serve as a guide to researchers in the generation and comparison of these or similar aggregate variables and also emphasizes the precautions that must be taken when using this approach.

References

Gonthier, D., Hotton, T., Cook, C., & Wilkins, R. (2006). Merging area-level census data with survey data in statistics Canada research data centres. *The Research Data Centres Information and Technical Bulletin*, *3*(*1*), 21-39.

Statistics Canada. 1997. Survey of labour and income dynamics: Microdata user's guide. *Dynamics of Labour and Income.* Catalogue No. 75M0001GPE. Ministry of Industry, Ottawa.

Statistics Canada. 1999. *1996 Census Dictionary, Final Edition Reference.* Catalogue No. 92-351-UIE. Ministry of Industry, Ottawa.

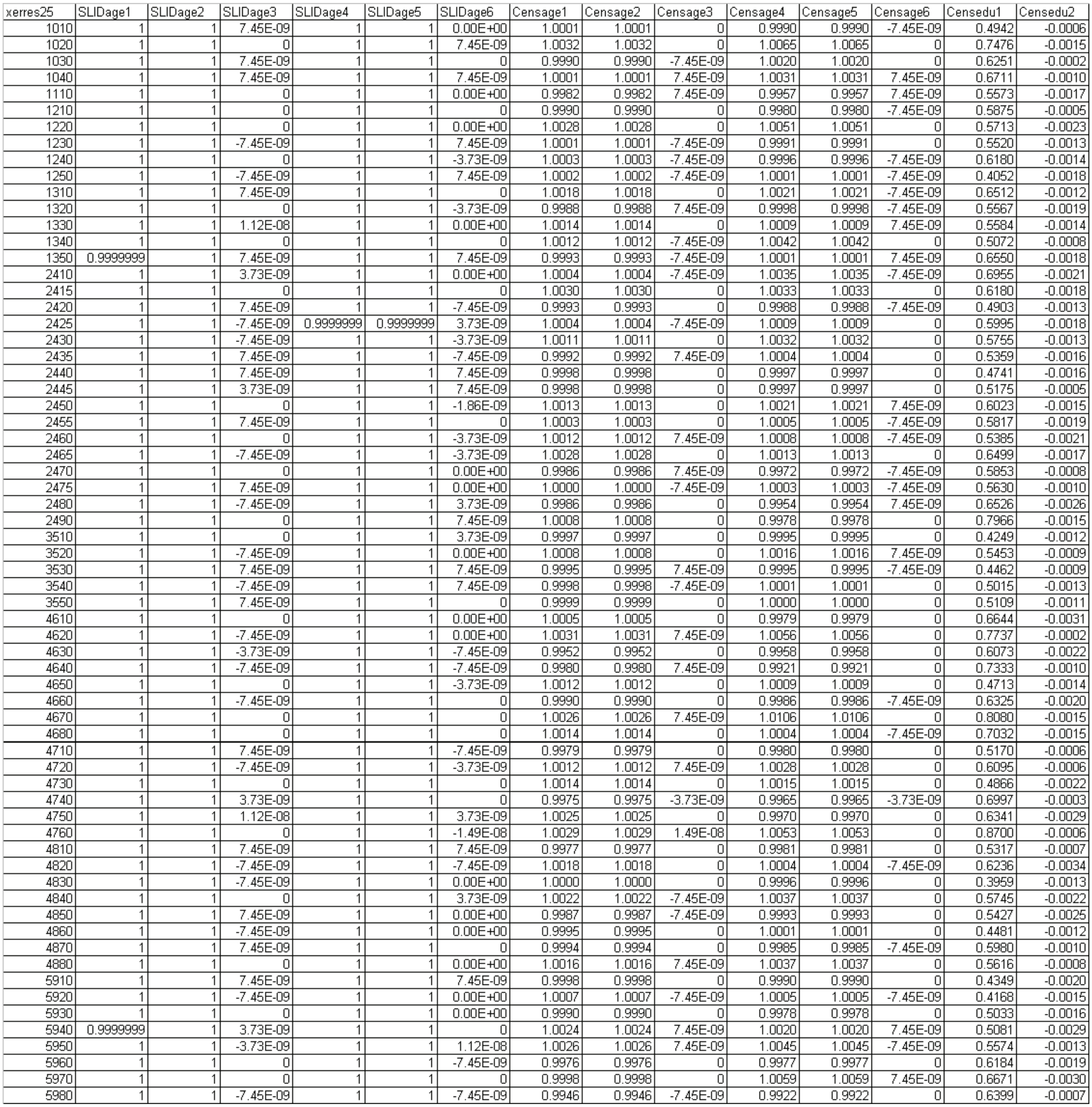
Statistics Canada. 2007. Guide to the Labour Force Survey. *Dynamics of Labour and Income.* Labour Statistics Division. Catalogue No. 71-543-GIE. Ministry of Industry, Ottawa.

Appendix

Tables A1 and A2 show the age and education verifications by ER for both the Census and the SLID for 1996 and 2001. The variable definitions are as follows:

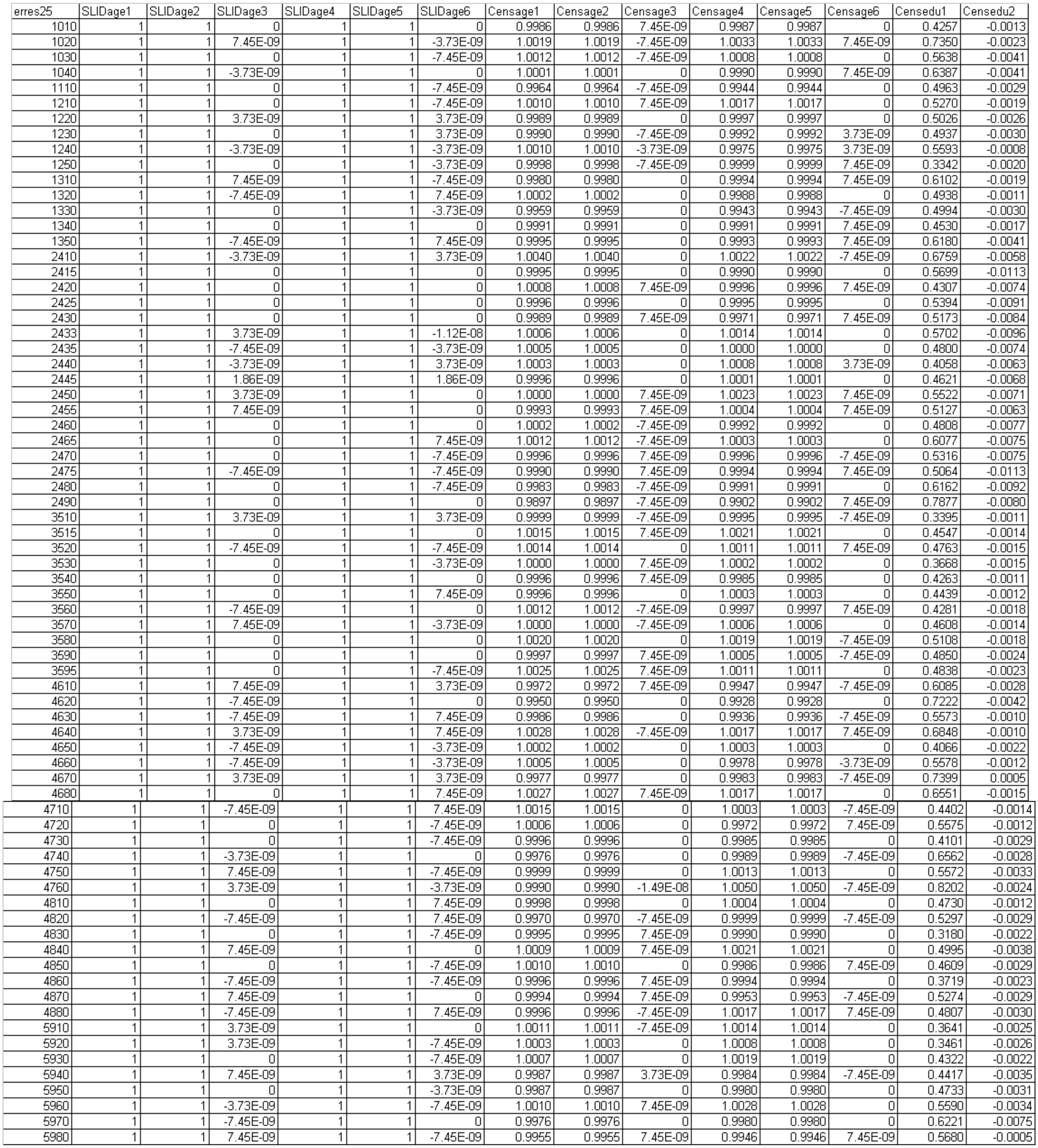
SLIDage1:	'pct_15to25' + 'pct_25to49' + 'pct_50to74' + 'pct_75over'
SLIDage2:	'pct_15to25' + 'pct_25to49' + 'pct_50to64' + 'pct_65over'
SLIDage3:	Difference between 'pct_15to25' and ('pct_15to19' + 'pct_20to24')
SLIDage4:	SLIDage1 for females
SLIDage5:	SLIDage2 for females
SLIDage6:	SLIDage3 for females
Censage1:	'pct_15to25' + 'pct_25to49' + 'pct_50to74' + 'pct_75over'
Censage2:	'pct_15to25' + 'pct_25to49' + 'pct_50to64' + 'pct_65over'
Censage3:	Difference between 'pct_15to25' and ('pct_15to19' + 'pct_20to24')
Censage4:	Censusage1 for females
Censage5:	Censusage2 for females
Censage6:	Censusage3 for females
Censedu1:	'Less than high school' + 'at least high school'
Censedu2:	Difference between 'postsecondary certificate' and 'university certificate' + 'non-university certificate'

Table A1: 1996 Age and Education Verifications by ER for the Census and SLID



xerres25	SLIDage1	SLIDage2	SLIDage3	SLIDage4	SLIDage5	SLIDage6	Censage1	Censage2	Censage3	Censage4	Censage5	Censage6	Censedu1	Censedu2
1010	1	1	7.45E-09	1	1	0.00E+00	1.0001	1.0001	0	0.9990	0.9990	-7.45E-09	0.4942	-0.0006
1020	1	1	0	1	1	7.45E-09	1.0032	1.0032	0	1.0065	1.0065	0	0.7476	-0.0015
1030	1	1	7.45E-09	1	1	0	0.9990	0.9990	-7.45E-09	1.0020	1.0020	0	0.6251	-0.0002
1040	1	1	7.45E-09	1	1	7.45E-09	1.0001	1.0001	7.45E-09	1.0031	1.0031	7.45E-09	0.6711	-0.0010
1110	1	1	0	1	1	0.00E+00	0.9982	0.9982	7.45E-09	0.9957	0.9957	7.45E-09	0.5573	-0.0017
1210	1	1	0	1	1	0	0.9990	0.9990	0	0.9980	0.9980	-7.45E-09	0.5875	-0.0005
1220	1	1	0	1	1	0.00E+00	1.0028	1.0028	0	1.0051	1.0051	0	0.5713	-0.0023
1230	1	1	-7.45E-09	1	1	7.45E-09	1.0001	1.0001	-7.45E-09	0.9991	0.9991	0	0.5520	-0.0013
1240	1	1	0	1	1	-3.73E-09	1.0003	1.0003	-7.45E-09	0.9996	0.9996	-7.45E-09	0.6180	-0.0014
1250	1	1	-7.45E-09	1	1	7.45E-09	1.0002	1.0002	-7.45E-09	1.0001	1.0001	-7.45E-09	0.4052	-0.0018
1310	1	1	7.45E-09	1	1	0	1.0018	1.0018	0	1.0021	1.0021	-7.45E-09	0.6512	-0.0012
1320	1	1	0	1	1	-3.73E-09	0.9988	0.9988	7.45E-09	0.9998	0.9998	-7.45E-09	0.5567	-0.0019
1330	1	1	1.12E-08	1	1	0.00E+00	1.0014	1.0014	0	1.0009	1.0009	7.45E-09	0.5584	-0.0014
1340	1	1	0	1	1	0	1.0012	1.0012	-7.45E-09	1.0042	1.0042	0	0.5072	-0.0008
1350	0.9999999	1	7.45E-09	1	1	7.45E-09	0.9993	0.9993	-7.45E-09	1.0001	1.0001	7.45E-09	0.6550	-0.0018
2410	1	1	3.73E-09	1	1	0.00E+00	1.0004	1.0004	-7.45E-09	1.0035	1.0035	-7.45E-09	0.6955	-0.0021
2415	1	1	0	1	1	0	1.0030	1.0030	0	1.0033	1.0033	0	0.6180	-0.0018
2420	1	1	7.45E-09	1	1	-7.45E-09	0.9993	0.9993	0	0.9988	0.9988	-7.45E-09	0.4903	-0.0013
2425	1	1	-7.45E-09	0.9999999	0.9999999	3.73E-09	1.0004	1.0004	-7.45E-09	1.0009	1.0009	0	0.5995	-0.0018
2430	1	1	-7.45E-09	1	1	-3.73E-09	1.0011	1.0011	0	1.0032	1.0032	0	0.5755	-0.0013
2435	1	1	7.45E-09	1	1	-7.45E-09	0.9992	0.9992	7.45E-09	1.0004	1.0004	0	0.5359	-0.0016
2440	1	1	7.45E-09	1	1	7.45E-09	0.9998	0.9998	0	0.9997	0.9997	0	0.4741	-0.0016
2445	1	1	3.73E-09	1	1	7.45E-09	0.9998	0.9998	0	0.9997	0.9997	0	0.5175	-0.0005
2450	1	1	0	1	1	-1.86E-09	1.0013	1.0013	0	1.0021	1.0021	7.45E-09	0.6023	-0.0015
2455	1	1	7.45E-09	1	1	0	1.0003	1.0003	0	1.0005	1.0005	-7.45E-09	0.5817	-0.0019
2460	1	1	0	1	1	-3.73E-09	1.0012	1.0012	7.45E-09	1.0008	1.0008	-7.45E-09	0.5385	-0.0021
2465	1	1	-7.45E-09	1	1	-3.73E-09	1.0028	1.0028	0	1.0013	1.0013	0	0.6499	-0.0017
2470	1	1	0	1	1	0.00E+00	0.9986	0.9986	7.45E-09	0.9972	0.9972	-7.45E-09	0.5853	-0.0008
2475	1	1	7.45E-09	1	1	0.00E+00	1.0000	1.0000	-7.45E-09	1.0003	1.0003	-7.45E-09	0.5630	-0.0010
2480	1	1	-7.45E-09	1	1	3.73E-09	0.9986	0.9986	0	0.9954	0.9954	7.45E-09	0.6526	-0.0026
2490	1	1	0	1	1	7.45E-09	1.0008	1.0008	0	0.9978	0.9978	0	0.7966	-0.0015
3510	1	1	0	1	1	3.73E-09	0.9997	0.9997	0	0.9995	0.9995	0	0.4249	-0.0012
3520	1	1	-7.45E-09	1	1	0.00E+00	1.0008	1.0008	0	1.0016	1.0016	7.45E-09	0.5453	-0.0009
3530	1	1	7.45E-09	1	1	7.45E-09	0.9995	0.9995	7.45E-09	0.9995	0.9995	-7.45E-09	0.4462	-0.0009
3540	1	1	-7.45E-09	1	1	7.45E-09	0.9998	0.9998	-7.45E-09	1.0001	1.0001	0	0.5015	-0.0013
3550	1	1	7.45E-09	1	1	0	0.9999	0.9999	0	1.0000	1.0000	0	0.5109	-0.0011
4610	1	1	0	1	1	0.00E+00	1.0005	1.0005	0	0.9979	0.9979	0	0.6644	-0.0031
4620	1	1	-7.45E-09	1	1	0.00E+00	1.0031	1.0031	7.45E-09	1.0056	1.0056	0	0.7737	-0.0002
4630	1	1	-3.73E-09	1	1	-7.45E-09	0.9952	0.9952	0	0.9958	0.9958	0	0.6073	-0.0022
4640	1	1	-7.45E-09	1	1	-7.45E-09	0.9980	0.9980	7.45E-09	0.9921	0.9921	0	0.7333	-0.0010
4650	1	1	0	1	1	-3.73E-09	1.0012	1.0012	0	1.0009	1.0009	0	0.4713	-0.0014
4660	1	1	-7.45E-09	1	1	0	0.9990	0.9990	0	0.9986	0.9986	-7.45E-09	0.6325	-0.0020
4670	1	1	0	1	1	0	1.0026	1.0026	7.45E-09	1.0106	1.0106	0	0.8080	-0.0015
4680	1	1	0	1	1	0	1.0014	1.0014	0	1.0004	1.0004	-7.45E-09	0.7032	-0.0015
4710	1	1	7.45E-09	1	1	-7.45E-09	0.9979	0.9979	0	0.9980	0.9980	0	0.5170	-0.0006
4720	1	1	-7.45E-09	1	1	-3.73E-09	1.0012	1.0012	7.45E-09	1.0028	1.0028	0	0.6095	-0.0006
4730	1	1	0	1	1	0	1.0014	1.0014	0	1.0015	1.0015	0	0.4866	-0.0022
4740	1	1	3.73E-09	1	1	0	0.9975	0.9975	-3.73E-09	0.9965	0.9965	-3.73E-09	0.6997	-0.0003
4750	1	1	1.12E-08	1	1	3.73E-09	1.0025	1.0025	0	0.9970	0.9970	0	0.6341	-0.0029
4760	1	1	0	1	1	-1.49E-08	1.0029	1.0029	1.49E-08	1.0053	1.0053	0	0.8700	-0.0006
4810	1	1	7.45E-09	1	1	7.45E-09	0.9977	0.9977	0	0.9981	0.9981	0	0.5317	-0.0007
4820	1	1	-7.45E-09	1	1	-7.45E-09	1.0018	1.0018	0	1.0004	1.0004	-7.45E-09	0.6236	-0.0034
4830	1	1	-7.45E-09	1	1	0.00E+00	1.0000	1.0000	0	0.9996	0.9996	0	0.3959	-0.0013
4840	1	1	0	1	1	3.73E-09	1.0022	1.0022	-7.45E-09	1.0037	1.0037	0	0.5745	-0.0022
4850	1	1	7.45E-09	1	1	0.00E+00	0.9987	0.9987	-7.45E-09	0.9993	0.9993	0	0.5427	-0.0025
4860	1	1	-7.45E-09	1	1	0.00E+00	0.9995	0.9995	0	1.0001	1.0001	0	0.4481	-0.0012
4870	1	1	7.45E-09	1	1	0	0.9994	0.9994	0	0.9985	0.9985	-7.45E-09	0.5980	-0.0010
4880	1	1	0	1	1	0.00E+00	1.0016	1.0016	7.45E-09	1.0037	1.0037	0	0.5616	-0.0008
5910	1	1	7.45E-09	1	1	7.45E-09	0.9998	0.9998	0	0.9990	0.9990	0	0.4349	-0.0020
5920	1	1	-7.45E-09	1	1	0.00E+00	1.0007	1.0007	-7.45E-09	1.0005	1.0005	-7.45E-09	0.4168	-0.0015
5930	1	1	0	1	1	0.00E+00	0.9990	0.9990	0	0.9978	0.9978	0	0.5033	-0.0016
5940	0.9999999	1	3.73E-09	1	1	0	1.0024	1.0024	7.45E-09	1.0020	1.0020	7.45E-09	0.5081	-0.0029
5950	1	1	-3.73E-09	1	1	1.12E-08	1.0026	1.0026	7.45E-09	1.0045	1.0045	-7.45E-09	0.5574	-0.0013
5960	1	1	0	1	1	-7.45E-09	0.9976	0.9976	0	0.9977	0.9977	0	0.6184	-0.0019
5970	1	1	0	1	1	0	0.9998	0.9998	0	1.0059	1.0059	7.45E-09	0.6671	-0.0030
5980	1	1	-7.45E-09	1	1	-7.45E-09	0.9946	0.9946	-7.45E-09	0.9922	0.9922	0	0.6399	-0.0007

Table A2: 2001 Age and Education Verifications by ER for the Census and SLID



erres25	SLIDage1	SLIDage2	SLIDage3	SLIDage4	SLIDage5	SLIDage6	Censage1	Censage2	Censage3	Censage4	Censage5	Censage6	Censedu1	Censedu2
1010	1	1	0	1	1	0	0.9986	0.9986	7.45E-09	0.9987	0.9987	0	0.4257	-0.0013
1020	1	1	7.45E-09	1	1	-3.73E-09	1.0019	1.0019	-7.45E-09	1.0033	1.0033	7.45E-09	0.7350	-0.0023
1030	1	1	0	1	1	-7.45E-09	1.0012	1.0012	-7.45E-09	1.0008	1.0008	0	0.5638	-0.0041
1040	1	1	-3.73E-09	1	1	0	1.0001	1.0001	0	0.9990	0.9990	7.45E-09	0.6387	-0.0041
1110	1	1	0	1	1	-7.45E-09	0.9964	0.9964	-7.45E-09	0.9944	0.9944	0	0.4963	-0.0029
1210	1	1	0	1	1	-7.45E-09	1.0010	1.0010	7.45E-09	1.0017	1.0017	0	0.5270	-0.0019
1220	1	1	3.73E-09	1	1	3.73E-09	0.9989	0.9989	0	0.9997	0.9997	0	0.5026	-0.0026
1230	1	1	0	1	1	3.73E-09	0.9990	0.9990	-7.45E-09	0.9992	0.9992	3.73E-09	0.4937	-0.0030
1240	1	1	-3.73E-09	1	1	-3.73E-09	1.0010	1.0010	-3.73E-09	0.9975	0.9975	3.73E-09	0.5593	-0.0008
1250	1	1	0	1	1	-3.73E-09	0.9998	0.9998	-7.45E-09	0.9999	0.9999	7.45E-09	0.3342	-0.0020
1310	1	1	7.45E-09	1	1	-7.45E-09	0.9980	0.9980	0	0.9994	0.9994	7.45E-09	0.6102	-0.0019
1320	1	1	-7.45E-09	1	1	7.45E-09	1.0002	1.0002	0	0.9988	0.9988	0	0.4938	-0.0011
1330	1	1	0	1	1	-3.73E-09	0.9959	0.9959	0	0.9943	0.9943	-7.45E-09	0.4994	-0.0030
1340	1	1	0	1	1	0	0.9991	0.9991	0	0.9991	0.9991	7.45E-09	0.4530	-0.0017
1350	1	1	-7.45E-09	1	1	7.45E-09	0.9995	0.9995	0	0.9993	0.9993	7.45E-09	0.6180	-0.0041
2410	1	1	-3.73E-09	1	1	3.73E-09	1.0040	1.0040	0	1.0022	1.0022	-7.45E-09	0.6759	-0.0058
2415	1	1	0	1	1	0	0.9995	0.9995	0	0.9990	0.9990	0	0.5699	-0.0113
2420	1	1	0	1	1	0	1.0008	1.0008	7.45E-09	0.9996	0.9996	7.45E-09	0.4307	-0.0074
2425	1	1	0	1	1	0	0.9996	0.9996	0	0.9995	0.9995	0	0.5394	-0.0091
2430	1	1	0	1	1	0	0.9989	0.9989	7.45E-09	0.9971	0.9971	7.45E-09	0.5173	-0.0084
2433	1	1	3.73E-09	1	1	-1.12E-08	1.0006	1.0006	0	1.0014	1.0014	0	0.5702	-0.0096
2435	1	1	-7.45E-09	1	1	-3.73E-09	1.0005	1.0005	0	1.0000	1.0000	0	0.4800	-0.0074
2440	1	1	-3.73E-09	1	1	3.73E-09	1.0003	1.0003	0	1.0008	1.0008	3.73E-09	0.4058	-0.0063
2445	1	1	1.86E-09	1	1	1.86E-09	0.9996	0.9996	0	1.0001	1.0001	0	0.4621	-0.0068
2450	1	1	3.73E-09	1	1	0	1.0000	1.0000	7.45E-09	1.0023	1.0023	7.45E-09	0.5522	-0.0071
2455	1	1	7.45E-09	1	1	0	0.9993	0.9993	7.45E-09	1.0004	1.0004	7.45E-09	0.5127	-0.0063
2460	1	1	0	1	1	0	1.0002	1.0002	-7.45E-09	0.9992	0.9992	0	0.4808	-0.0077
2465	1	1	0	1	1	7.45E-09	1.0012	1.0012	-7.45E-09	1.0003	1.0003	0	0.6077	-0.0075
2470	1	1	0	1	1	-7.45E-09	0.9996	0.9996	7.45E-09	0.9996	0.9996	-7.45E-09	0.5316	-0.0075
2475	1	1	-7.45E-09	1	1	-7.45E-09	0.9990	0.9990	7.45E-09	0.9994	0.9994	7.45E-09	0.5064	-0.0113
2480	1	1	0	1	1	-7.45E-09	0.9983	0.9983	-7.45E-09	0.9991	0.9991	0	0.6162	-0.0092
2490	1	1	0	1	1	0	0.9897	0.9897	-7.45E-09	0.9902	0.9902	7.45E-09	0.7877	-0.0080
3510	1	1	3.73E-09	1	1	3.73E-09	0.9999	0.9999	-7.45E-09	0.9995	0.9995	-7.45E-09	0.3395	-0.0011
3515	1	1	0	1	1	0	1.0015	1.0015	7.45E-09	1.0021	1.0021	0	0.4547	-0.0014
3520	1	1	-7.45E-09	1	1	-7.45E-09	1.0014	1.0014	0	1.0011	1.0011	7.45E-09	0.4763	-0.0015
3530	1	1	0	1	1	-3.73E-09	1.0000	1.0000	7.45E-09	1.0002	1.0002	0	0.3668	-0.0015
3540	1	1	0	1	1	0	0.9996	0.9996	7.45E-09	0.9985	0.9985	0	0.4263	-0.0011
3550	1	1	0	1	1	7.45E-09	0.9996	0.9996	0	1.0003	1.0003	0	0.4439	-0.0012
3560	1	1	-7.45E-09	1	1	0	1.0012	1.0012	-7.45E-09	0.9997	0.9997	7.45E-09	0.4281	-0.0018
3570	1	1	7.45E-09	1	1	-3.73E-09	1.0000	1.0000	-7.45E-09	1.0006	1.0006	0	0.4608	-0.0014
3580	1	1	0	1	1	0	1.0020	1.0020	0	1.0019	1.0019	-7.45E-09	0.5108	-0.0018
3590	1	1	0	1	1	0	0.9997	0.9997	7.45E-09	1.0005	1.0005	-7.45E-09	0.4850	-0.0024
3595	1	1	0	1	1	-7.45E-09	1.0025	1.0025	7.45E-09	1.0011	1.0011	0	0.4838	-0.0023
4610	1	1	7.45E-09	1	1	3.73E-09	0.9972	0.9972	7.45E-09	0.9947	0.9947	-7.45E-09	0.6085	-0.0028
4620	1	1	-7.45E-09	1	1	0	0.9950	0.9950	0	0.9928	0.9928	0	0.7222	-0.0042
4630	1	1	-7.45E-09	1	1	7.45E-09	0.9986	0.9986	0	0.9936	0.9936	-7.45E-09	0.5573	-0.0010
4640	1	1	3.73E-09	1	1	7.45E-09	1.0028	1.0028	-7.45E-09	1.0017	1.0017	7.45E-09	0.6848	-0.0010
4650	1	1	-7.45E-09	1	1	-3.73E-09	1.0002	1.0002	0	1.0003	1.0003	0	0.4066	-0.0022
4660	1	1	-7.45E-09	1	1	-3.73E-09	1.0005	1.0005	0	0.9978	0.9978	-3.73E-09	0.5578	-0.0012
4670	1	1	3.73E-09	1	1	3.73E-09	0.9977	0.9977	0	0.9983	0.9983	-7.45E-09	0.7399	0.0005
4680	1	1	0	1	1	7.45E-09	1.0027	1.0027	7.45E-09	1.0017	1.0017	0	0.6551	-0.0015
4710	1	1	-7.45E-09	1	1	7.45E-09	1.0015	1.0015	0	1.0003	1.0003	-7.45E-09	0.4402	-0.0014
4720	1	1	0	1	1	-7.45E-09	1.0006	1.0006	0	0.9972	0.9972	7.45E-09	0.5575	-0.0012
4730	1	1	0	1	1	-7.45E-09	0.9996	0.9996	0	0.9985	0.9985	0	0.4101	-0.0029
4740	1	1	-3.73E-09	1	1	0	0.9976	0.9976	0	0.9989	0.9989	-7.45E-09	0.6562	-0.0028
4750	1	1	7.45E-09	1	1	-7.45E-09	0.9999	0.9999	0	1.0013	1.0013	0	0.5572	-0.0033
4760	1	1	3.73E-09	1	1	-3.73E-09	0.9990	0.9990	-1.49E-08	1.0050	1.0050	-7.45E-09	0.8202	-0.0024
4810	1	1	0	1	1	7.45E-09	0.9998	0.9998	0	1.0004	1.0004	0	0.4730	-0.0012
4820	1	1	-7.45E-09	1	1	7.45E-09	0.9970	0.9970	-7.45E-09	0.9999	0.9999	-7.45E-09	0.5297	-0.0029
4830	1	1	0	1	1	-7.45E-09	0.9995	0.9995	7.45E-09	0.9990	0.9990	0	0.3180	-0.0022
4840	1	1	7.45E-09	1	1	0	1.0009	1.0009	7.45E-09	1.0021	1.0021	0	0.4995	-0.0038
4850	1	1	0	1	1	-7.45E-09	1.0010	1.0010	0	0.9986	0.9986	7.45E-09	0.4609	-0.0029
4860	1	1	-7.45E-09	1	1	-7.45E-09	0.9996	0.9996	7.45E-09	0.9994	0.9994	0	0.3719	-0.0023
4870	1	1	7.45E-09	1	1	0	0.9994	0.9994	7.45E-09	0.9953	0.9953	-7.45E-09	0.5274	-0.0029
4880	1	1	-7.45E-09	1	1	7.45E-09	0.9996	0.9996	-7.45E-09	1.0017	1.0017	7.45E-09	0.4807	-0.0030
5910	1	1	3.73E-09	1	1	0	1.0011	1.0011	-7.45E-09	1.0014	1.0014	0	0.3641	-0.0025
5920	1	1	3.73E-09	1	1	-7.45E-09	1.0003	1.0003	0	1.0008	1.0008	0	0.3461	-0.0026
5930	1	1	0	1	1	-7.45E-09	1.0007	1.0007	0	1.0019	1.0019	0	0.4322	-0.0022
5940	1	1	7.45E-09	1	1	3.73E-09	0.9987	0.9987	3.73E-09	0.9984	0.9984	-7.45E-09	0.4417	-0.0035
5950	1	1	0	1	1	-3.73E-09	0.9987	0.9987	0	0.9980	0.9980	0	0.4733	-0.0031
5960	1	1	-3.73E-09	1	1	-7.45E-09	1.0010	1.0010	7.45E-09	1.0028	1.0028	0	0.5590	-0.0034
5970	1	1	-7.45E-09	1	1	0	0.9976	0.9976	0	0.9980	0.9980	0	0.6221	-0.0075
5980	1	1	7.45E-09	1	1	-7.45E-09	0.9955	0.9955	7.45E-09	0.9946	0.9946	7.45E-09	0.5680	-0.0005

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 626-631

Markov Modeling of Breast Cancer

Chunling Cong   Chris P. Tsokos
University of South Florida

Previous work with respect to the treatments and relapse time for breast cancer patients is extended by applying a Markov chain to model three different types of breast cancer patients: alive without ever having relapse, alive with relapse, and deceased. It is shown that combined treatment of tamoxifen and radiation is more effective than single treatment of tamoxifen in preventing the recurrence of breast cancer. However, if the patient has already relapsed from breast cancer, single treatment of tamoxifen would be more appropriate with respect to survival time after relapse. Transition probabilities between three stages during different time periods, 2-year, 4-year, 5-year, and 10-year, are also calculated to provide information on how likely one stage moves to another stage within a specific time period.

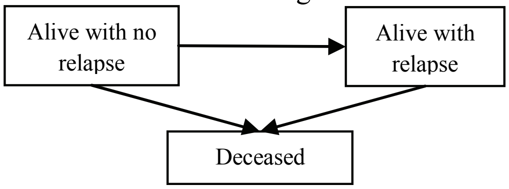
Key words: Markov chain, breast cancer, relapse time, tamoxifen and radiation.

Introduction

The Markov (1906) chain model has been applied in various fields such as physics, queuing theory, internet application, economics, finance, and social sciences among others. As an effective and efficient way of describing a process in which an individual moves through a series of states (stages) in continuous time, homogeneous Markov models have also been extensively used in health sciences where the progression of certain diseases are of great importance to both doctors and patients. In the present study, the main objective is to investigate the progression of breast cancer in patients in three different stages who were given different treatments. One group of patients received combined treatments of tamoxifen and radiation, and the other group received only tamoxifen. Figure 1 shows the three stages of interest in the study are: alive with no relapse, alive with relapse, and deceased. Even though breast cancer patients who have recurrence may be treated and recover from breast cancer to become active with no relapse, due to the fact that the data does not include any observations of that process, we consider the second state-alive with relapse as those patients who once had relapse and are still alive, regardless of whether they have recovered from breast cancer or not.

Chunling Cong is a doctoral student in Statistics at the University of South Florida. Her research interests are in developing forecasting and statistical analysis and modeling of cancer. Email: ccong@mail.usf.edu. Chris Tsokos is a Distinguished University Professor in mathematics and Statistics at the University of South Florida. His research interests are in modeling Global Warming, analysis and modeling of cancer data, parametric, Bayesian and nonparametric reliability, and stochastic systems, among others. He is a fellow of both ASA and ISI. Email: profcpt@cas.usf.edu.

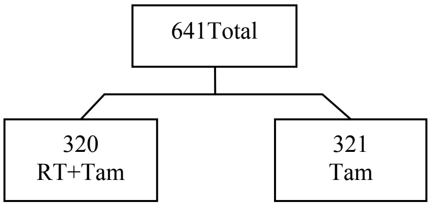
Figure 1: Three Stages of Breast Cancer Modeling



Methodology

Between December 1992 and June 2000, a total of 769 women were enrolled and randomized in the study. Among these, 386 received combined radiation and tamoxifen (RT+Tam), and the remaining 383 received tamoxifen (Tam) only. The last follow-up was conducted in the summer of 2002. As shown in Figure 2, only those 641 patients enrolled at the Princess Margaret Hospital are included: 320 and 321 in RT+Tam and Tam treatment groups, respectively.

Figure 2: Breast Cancer Data



This data was used by Fyles, et al. and was later analyzed by Ibrahim, et al. Analysis was conducted on this data with respect to the treatment effect of the two different treatments using decision tree and modeled relapse time using AFT and Cox-PH model. Mixture models were also applied to compare the cure rate of the two groups.

The Markov Chain Model

The Markov chain is a model for a finite or infinite random process sequence $X=\{X_1,X_{2,...,X_N}\}$. Unlike the independent identical distribution (i.i.d) model that assumes the independency of a sequence of events X_i's, the Markov model takes into account the dependencies among the X_i's.

Consider a random process $X=\{X_t\}_{t\geq 1}=\{X_{1,}X_2,...\}$ of random variables taking values in a discrete set space of stages $S=\{1,2,3,...,s\}$ where X_t represents the state of the process of an individual at time *t*. The transitions possible among the three stages in this study, alive without relapse, alive with relapse, and deceased are shown in Figure 1 indicated by arrows. Consider a realization of the history of the process up to and including time *t*, as $\{X_t=x_t,X_{t-1}=x_{t-1},...,X_1=x_1\}$, where $x_t,x_{t-1},...,x_1$ is a sequence of stages at different times. A random process is called a Markov Chain if the conditional probabilities between the stages at different times satisfy the Markov property: the conditional probability of future one-step-event conditioned on the entire past of the process is just conditioned on the present stage of the process. In other words, the one-step future stage depends only on the present stage:

$$\begin{aligned}&P(X_{t+1}=x_{t+1}\,|\,X_t=x_t,X_{t-1}=x_{t-1},...,X_1=x_1)\\&=P(X_{t+1}=x_{t+1}\,|\,X_t=x_t)\end{aligned}\qquad(1)$$

for every sequence $x_1,...,x_t,x_{t+1}$ of elements of S and every $t\geq 1$.

The transition probability from stage i to stage j at time t and transition intensity are defined by

$$p_{ij}(t)=p(X_{t+1}=j\,|\,X_t=i),\qquad(2)$$

and

$$q_{ij}(t)=\lim_{h\to 0}\frac{P(X(t+h)=j\,|\,X(t)=i)}{h},\qquad(3)$$

where h is the time interval.

If the transition probabilities do not depend on time, $p_{ij}(t)$ can simply be written as p_{ij}, then the Markov chain is called time-homogeneous. If not specified, the following analysis is based on time-homogeneous Markov chain. A transition probability matrix $P(t)$ consisting of all the transition probabilities between stages in a matrix form is given by:

$$P(t)=\left\{\begin{matrix}p_{11}(t) & p_{12}(t) & \cdots & p_{1s}(t)\\ p_{21}(t) & p_{22}(t) & \cdots & p_{2s}(t)\\ \cdots & \cdots & \cdots & \cdots\\ p_{s1}(t) & p_{s2}(t) & \cdots & p_{ss}(t)\end{matrix}\right\},\qquad(4)$$

where probabilities in each row add up to 1. Thus, it is 100% certain that for any individual at time *t* is in one of the stages and the sum of probabilities of being in each stage is 1.

The transition probability matrix can be calculated by taking the matrix exponential of the scaled transition intensity matrix defined by

$$P(t)=Exp(tQ),\qquad(5)$$

where

$$Q = \begin{Bmatrix} q_{11} & q_{12} & \cdots & q_{1s} \\ q_{21} & q_{22} & \cdots & q_{2s} \\ \cdots & \cdots & \cdots & \cdots \\ q_{s1} & q_{s2} & \cdots & q_{ss} \end{Bmatrix}, \qquad (6)$$

and q_{ij} denotes the transition intensity from stage i to stage j.

The exponential of a matrix A is defined by

$$Exp(A) = 1 + A^2 / 2! + A^3 / 3! + \dots, \qquad (7)$$

where each summand in the series is the matrix products. In this manner, once the intensity matrix is given, the transition probabilities can be calculated as shown above.

Next, the intensity matrix and transition probabilities matrix can be obtained by maximizing the likelihood $L(Q)$ which is a function of Q. Consider an individual consisting of a series of times $(t_1, t_2, ..., t_n)$ and corresponding stages $(x_1, x_2, ..., x_n)$. More specifically, consider a pair of successive stages observed to be i and j at time t_i and t_j. Three scenarios are proposed and considered here.

Scenario 1

If the information for the individual is obtained at arbitrary observation times (the exact time of the transition of stages is unknown) the contribution to the likelihood from this pair of states is:

$$L_{ij} = p_{ij}(t_j - t_i). \qquad (8)$$

Scenario 2

If the exact times of transitions between different stages are recorded and there is no transition between the observation times, the contribution to the likelihood from this pair of stages is:

$$L_{ij} = p_{ij}(t_j - t_i)q_{ij}. \qquad (9)$$

Scenario 3

If the time of death is known or $j = death$, but the stage on the previous instant before death is unknown as denoted by k (k could be any possible stage between stage i and death), the contribution to the likelihood function from this pair of stages is:

$$L_{ij} = \sum_{k \neq j} p_{ik}(t_j - t_i)q_{kj}. \qquad (10)$$

Results

The breast cancer patients were divided into two groups RT+Tam and Tam based on the different treatments they received. For those patients who received combined treatments, 26 patients experienced relapse, 13 patients died without recurrence of breast cancer during the entire period of the study, and 14 died after recurrence of breast cancer. For the patients in the Tam group, 51 patients experienced relapse, 10 died without reoccurrence of breast cancer, and 13 died after recurrence of breast cancer.

As can be observed from the transition intensity matrixes for both groups RT+Tam and Tam as shown in Tables 1 and 2, patients who received single treatment have a higher transition intensity form Stage 1 to Stage 2, thus, they are more likely to have breast cancer recurrence. Thus, the probability of that happening in the Tam group is higher than that of the RT+Tam group. For those patients who died without relapse, there is no significant difference between the two treatments as illustrated by the intensity form Stage 1 to Stage 3.

Combined treatment is also more effective than a single treatment with respect to the possibility of death without relapse as can be observed from the transition intensity from Stage 1 to Stage 3. However, for those who already experienced relapse of breast cancer, patients who received combined treatments are more likely to die than those who received a single treatment. Therefore, combined treatment should be chosen over single treatment to avoid recurrence, but for those patients who already had breast cancer relapse, it would be advisable to choose a single treatment to extend the time from recurrence to death.

Figures 3 and 4 illustrate the effectiveness of the two treatments with respect to the survival probabilities and also show the survival curves of the patients who had recurrence and who had no recurrence in each treatment group.

From the above analysis, the proposed Markov chain model provides recommendations for which treatment to choose for breast cancer patients with respect to relapse and survival time. Moreover, it provides patients with very important information on the exact time or possibilities of recurrence and death. Estimated mean sojourn times in each transient stage for patients who received combined treatment are 43.46 and 3.25 in Stage 1 and Stage 2, respectively. Estimated mean sojourn times for patients who received single treatment are 25.53 and 11.72 in Stage 1 and Stage 2. This further confirms that patients with combined treatment will remain in Stage 1 longer than those with single treatment; however, for patients who had relapse of breast cancer, patients with single treatment will stay alive longer than those with combined treatment.

Another goal of this study was to provide a transition probability matrix at different times so that given a specific time period, the probability that a patient in a given stage will transit to another stage could be conveyed. Tables 5a-8b give 2-year, 4-year, 5-year and 10-year transition probability matrixes of patients in RT+Tam and Tam.

Conclusion

Through Markov chain modeling of the three stages of breast cancer patients , it has been shown that combined treatment of tamoxifen and radiation is more effective than single treatment of tamoxifen in preventing the recurrence of breast cancer. However, for patients who had a relapse of breast cancer, single treatment of tamoxifen proves to be more effective than combined treatment with respect to the survival probability. This finding could give significant guidance to doctors with respect to which breast cancer treatment should be given to breast cancer patients in different stages. Transition probabilities between different stages during 2 years, 4 years, 5 years and 10 years are also calculated for predicting purposes. Those transition probabilities could help provide a clearer view of how one stage transits to another stage within a given time period.

Acknowledgements

We wish to thank N.A Ibrahim for supplying us the source of the data that made the subject study possible. We also wish to express appreciation to Dr. James Kepner, Vice President of the American Cancer Society for our useful discussions on the present study.

References

Dynkin, E. B. (2006). *Theory of Markov Processes*. Dover Publications.

Fyles, A. W., & McCready, D. R. (2004). Tamoxifen with or without breast irradiation in women 50 years of age or older with early breast cancer, *New England Journal of Medicine*, *351*, 963-970.

Gentleman, R. C., Lawless J. F., Lindsey J. C., & Yan, P. (1994). Multi-State Markov models for analyzing incomplete disease history data with illustrations for HIV disease. *Statist. Med.*, *13*, 805-821.

Ibrahim N. A., et al. (2008). Decision tree for competing risks survival probability in breast cancer study, *International Journal of Biomedical Sciences*, Volume 3 Number 1.

Jackson, C. (2007). *Multi-State modeling with R: The msm package, version 0.7.4.1 October*.

Kalblfleisch, J. D., & Lawless J. F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, *80*, 863-871.

Kay, R. (1986). A Markov model for analyzing cancer markers and disease states in survival studies. *Biometrics*, *42*, 855-865.

Lu, Y. & Stitt, F. W. (1994). Using Markov processes to describe the prognosis of HIV-1 infection. *Medical Decision Making*, *14*, 266-272.

Markov, A. (1906). Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga, Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete, 2-ya seriya, tom 15, *9 4*, 135-156.

Norris J. R. (1998). *Markov Chains*, Cambridge University Press.

Satten, G. A. & Longini, L. M. (1996). Markov chains with measurement error: estimating the true course of a marker of the progression of human immunodeficiency virus disease. *Applied Statistician*, *45*, 275-309.

Sharples, L. D. (1993). Use of the Gibbs sampler to estimate transition rates between grades of coronary disease following cardiac transplantation. *Statistics in Medicine*, *12*, 1155-1169.

Table 1: Transition Intensity Matrix of RT+Tam

	Stage 1	Stage 2	Stage 3
Stage 1	-0.02301	0.01957	0.0034
Stage 2	0	-0.3074	0.3074
Stage 3	0	0	0

Table 2: Transition Intensity Matrix of Tam

	Stage 1	Stage 2	Stage 3
Stage 1	-0.03917	0.03528	0.003889
Stage 2	0	-0.08533	0.08533
Stage 3	0	0	0

Figure 3: Survival Curves of Patients in RT+Tam

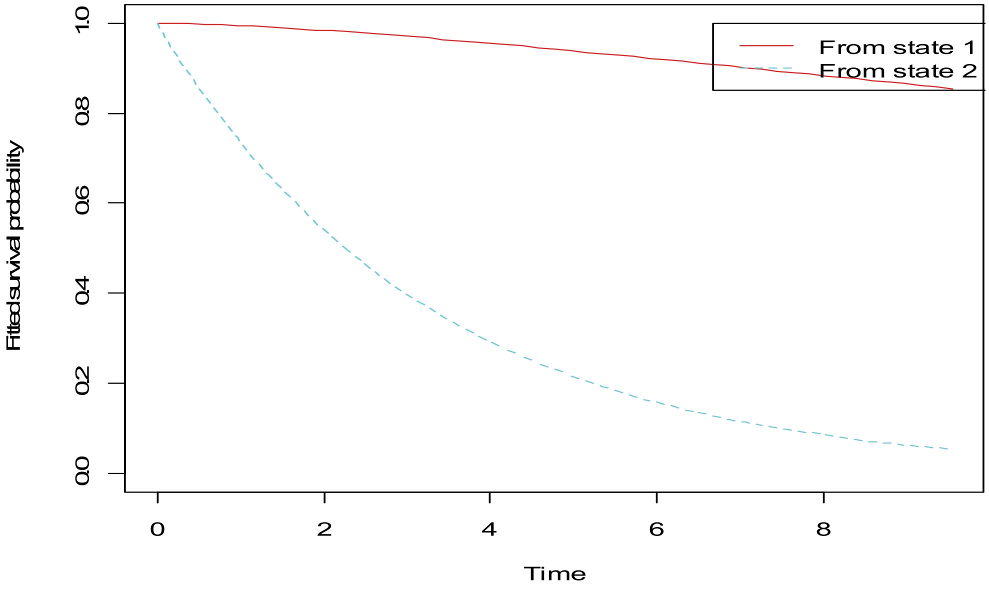


Figure 4: Survival Curves of Patients in Tam

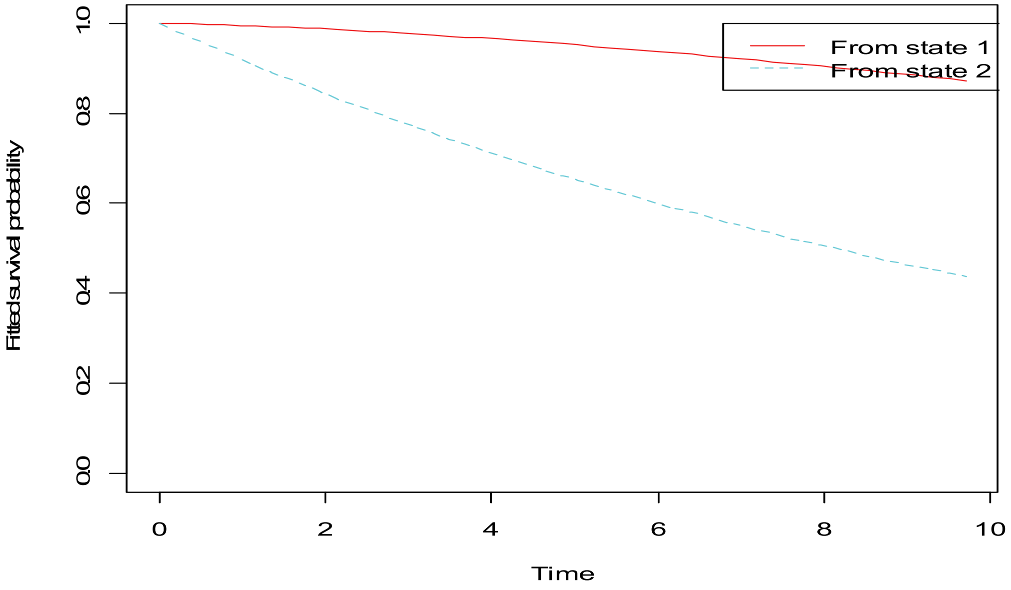


Table 5a: 2-year transition matrix for RT+Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.9550	0.0285	0.0165
Stage 2	0	0.5408	0.4592
Stage 3	0	0	0

Table 5b: 2-year transition matrix for Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.9247	0.0623	0.0130
Stage 2	0	0.8431	0.1569
Stage 3	0	0	0

Table 6a: 4-year transition matrix for RT+Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.9121	0.0426	0.0453
Stage 2	0	0.2925	0.7075
Stage 3	0	0	0

Table 6b: 4-year transition matrix for Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.8550	0.1102	0.0348
Stage 2	0	0.7108	0.2892
Stage 3	0	0	0

Table 7a: 5-year transition matrix for RT+Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.8913	0.0466	0.0621
Stage 2	0	0.2151	0.7849
Stage 3	0	0	0

Table 7b: 5-year transition matrix for Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.8221	0.1295	0.0484
Stage 2	0	0.6527	0.3473
Stage 3	0	0	0

Table 8a: 10-year transition matrix for RT+Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.7945	0.0515	0.1540
Stage 2	0	0.0463	0.9537
Stage 3	0	0	0

Table 8b: 10-year transition matrix for Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.6759	0.1910	0.1331
Stage 2	0	0.4260	0.5740
Stage 3	0	0	0