An assessment of deviations from conditional independence in binary data fusion

Smit, Elsabé
University of the Witwatersrand, School of Statistics and Actuarial Science
1 Jan Smuts Avenue
Johannesburg (2000), South Africa
elsabe.smit@wits.ac.za

INTRODUCTION

In the market research industry, large amounts of data are collected on consumer attitudes and behaviour, via surveys. Despite the fact that there is a wealth of marketing data available from the separate surveys, reports are generally only created for each source individually. No single source of comprehensive information is available for in-depth data mining that can assist in identifying business opportunities (Van der Putten, Kok and Gupta, 2002). As a result marketers often request more detail in their consumer surveys to address all their research needs in a single source.

This need for information in survey research places a large demand on the consumer to provide accurate and detailed information on attitudes and behaviour through the use of longer questionnaires. Consequently the quality of responses is affected through respondent fatigue and even an increase in survey non-response due to refusal to participate in time consuming surveys (Raghunathan and Grizzle, 1995).

One possible solution to this problem of questionnaire overload is to divide the larger survey into smaller parts and administer each part to different samples from the same target population (Raghunathan and Grizzle, 1995). The separate databases would then be combined through data fusion, a technique used for linking multiple data sources through a set of common characteristics (D'Orazio, Di Zio and Scanu, 2006). The information in the individual data sources is collected from different but similar respondents from the same target population. The objective is to estimate the joint distribution of the variables unique to each data source, using the common characteristics. This will enable the analyst to construct a synthetic data file that contains all the information from the separate data sources, as if the entire survey was administered to each respondent.

Data fusion can only be used as a viable solution to the problem of questionnaire overload if it will result in a valid data set that reflects the true relationships between the variables of interest. This largely depends on the link between the set of common variables and the unique variables, i.e. the underlying mathematical model that defines the bridge between the individual data sources. However, since the unique variables are never jointly observed, this link requires certain assumptions that are generally impossible to test in practice. The most common model used to fuse data is based on the assumption of conditional independence (CIA), where the unique variables **Y** and **Z** are assumed to be independent, given knowledge of the set of common characteristics **X**. This implies that all partial correlations $\rho(YZ \mid X)$ are equal to zero. As a result, it imposes the restriction:

$$\Sigma_{YZ} = \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XZ} \qquad (1)$$

If the assumption of conditional independence holds true, the joint density can be completely identified through the conditional and marginal distributions from the separate data sources, through the equation:

$$f(x,y,z) = f_{Y|X}(y \mid x) f_{Z|X}(z \mid x) f_X(x) \qquad (2)$$

The objective of this research is to evaluate the process of data fusion for binary data under the assumption of conditional independence, and to assess how deviations from this assumption will impact the results.

BACKGROUND

Data fusion has been applied and evaluated for many years and in many different disciplines, such as econometric modelling, policy development and market research. Several of the data fusion projects showed promising results, while others were not so convincing.

Many of the initial data fusion applications were in the field of economics. Examples of these include a fusion between the IRS tax file and CPS income data in 1968 (Radner, Allen, Gonzalez, Jabine, and Muller, 1980), the MERGE-66 fusion between the SEO and IRS tax file (Okner, 1972), a planned fusion between the Canadian SCF and FEX surveys (Alter, 1974), the MESP synthetic data file (Wolff, 1977), and more recently, the Pensim2 fusion model (Redway, 2003). These fusions allowed researchers to estimate the size distribution of household or personal income, to compare the relative income distribution internationally, to create more detailed micro-level data to be used for tax policy analysis, or to simulate pension policy scenarios.

Data fusion applications in market research include the UK BARB-TGI database (O'Brien, 1991), the Dutch SummoScanner database (Tchaoussoglou and Van der Noort, 1999), and the TGI-TAM database in Latin America (Soong and De Montigny, 2001), all of which are used for media planning purposes. Becker and Collins (2007) describe a fusion aimed at evaluating the relationship between media and product behaviour, and Internet consumption.

From early on, the technique was not without problems. The initial development of data fusion methods did not involve any strong theoretical basis (Rodgers, 1984). Sims (1972) was the first to highlight the weaknesses in the CIA for the single imputation data fusion approach. Through simulation, Rässler and Fleischer (1998) show that any deviations from conditional independence result in an incorrect representation of the true relationship between the sets of unique variables. This matter continues to be the main concern regarding fusion applications that use the CIA as the underlying model to describe the relationship between variables that were not jointly observed.

Other observations regarding the quality of a fusion centred on the quality of the individual data sources (Alter, 1974), as well as the choice of the set of common variables to ensure the maximum predictive power (Rodgers, 1984). Common to all the fusion applications is the amount of time necessary to fully explore and validate the analysis. In short, there is certainly no quick solution to data fusion.

Recent years have seen a rise in the use of multiple imputation as a way to perform data fusion without the restrictive assumption of conditional independence, based on Rubin (1986). Rässler (2002) formalizes an approach that creates multiple imputations under an explicit Bayesian model. Moriarity and Scheuren (2004) describe a regression-based algorithm to fuse data that assesses uncertainty in matching.

New fusion techniques are constantly evaluated to find models that will improve on existing models such as the non-parametric local linear regression estimator (LLR), introduced by Conti, Marella and Scanu (2008).

METHODOLOGY

In order to address the research question, data are simulated to reflect the distribution of survey-based data, where categorical variables are represented as binary indicators. The simulation is based on a pre-specified marginal distribution and correlation structure, using the binary simulation technique proposed by Alosh and Lee (2001). This algorithm requires the marginal distribution and correlation structure as input, and produces the complete joint probability distribution for *D* binary variables. This algorithm is restricted to positive correlations only, as is generally the case in market research applications.

Each simulated data set consists of four binary variables, namely X, Y, Z_1 and Z_2, with marginal distribution P = (0.7, 0.6, 0.8, 0.5)'. The data are simulated to reflect varying degrees of conditional independence, a property which is captured in the correlation structure. The correlations between binary

variables are restricted by the marginal distribution of the variables. Input correlation structures are simulated as follows: an initial correlation matrix is randomly selected from the valid range of correlations such that conditional independence is absent, i.e. both partial correlations are significantly different from zero. A second, related matrix is then created using Equation (1), thereby enforcing the presence of CIA. The correlations between Y and Z_1, and Y and Z_2 in the two initial correlation matrices provide a range for deviation from CIA. A further eight correlation matrices are selected from this range through either incremental deviation or random selection. The process is repeated 3,000 times, resulting in a total of 30,000 input correlation matrices.

The output from each binary simulation, namely the complete joint probability distribution for four binary variables, is used to create a micro-level data set of size $n = 2000$. The joint probability distribution indicates the proportion of any sample that consists of a particular configuration of zeros and ones. By applying these probabilities to the required sample size, the number of observations in the sample with the specific outcome over four binary variables is created. Such a data set is then seen as the theoretical data that would have occurred if all items in the questionnaire were administered to all respondents.

A key component of this research is to quantify the degree of CIA in each of the simulated data sets. This is done using a function of entropy, called the conditional mutual information (CMI), which indicates how Y and Z are related in the context of a third variable X (Jakulin and Bratko, 2004). It measures the reduction in uncertainty about Y (or Z) due to knowledge of Z (or Y), when X is given. The CMI is always zero or positive. If the CMI is equal to zero, it implies that Y and Z are unrelated, given knowledge of X. Therefore, the association between Y and Z is completely explained by X. This corresponds to the definition of CIA in the context of data fusion, where a zero value indicates complete CIA and a positive value indicates deviation from CIA to some degree. It is defined as:

$$I(Y,Z \mid X) = H(YX) + H(ZX) - H(X) - H(YZX) \qquad (3)$$

where $H(YX)$ is the joint entropy of variables Y and X.

For numerical interpretation, the CMI measure is expressed as a percentile of its valid range. This is referred to as the quantified conditional independence measure (qCIA) and is calculated as follows:

$$qCIA = \frac{I(Y,Z \mid X)}{\min\{H(Y), H(Z)\}} \times 100 \qquad (4)$$

The suitability of the qCIA as a measure of conditional independence is confirmed by comparing the qCIA with the two partial correlations $\rho(YZ_1 \mid X)$ and $\rho(YZ_2 \mid X)$, illustrated in Figure 1. This graph shows that both partial correlations are close to zero for low values of the qCIA, and when at least one of the two partial correlations deviate from zero, then the qCIA also deviates from zero. Therefore, this measure can be used to effectively quantify the level of CIA present in the simulated data.

The simulated data set is divided into two subsets of approximately equal size, subsets A and B. Variables Z_1 and Z_2 are removed from subset A such that it will include variables X and Y only. On the other hand subset B will include the data for X, Z_1 and Z_2. Therefore, Y and **Z** are not jointly observed in the two subsets.

Files A and B are linked together through the common variable X. D'Orazio *et al* (2006) note that the multinomial distribution is a very flexible model for fusing categorical or discrete data. The maximum likelihood estimators of the various components of the joint distribution under the CIA are used to estimate the complete joint probability distribution of the fused data file, rather than creating a respondent-level data file.

Consider the trivariate multinomial distribution (X, Y, Z) with $I \times J \times K$ categories and parameter vector $\Theta = \theta_{ijk}$, such that $\theta_{ijk} = P(X = i, Y = j, Z = k)$, $i \in \{0,1\}$, $j \in \{0,1\}$, $k \in \{00,01,10,11\}$. Under the CIA the joint distribution for the multinomial (X, Y, Z) is given by

$$\theta_{ijk} = \theta_{i..}\theta_{j|i}\theta_{k|i} = \frac{\theta_{ij.}\theta_{i.k}}{\theta_{i..}} \qquad (5)$$

The results from each fused data set are compared to the corresponding original data set, addressing all four levels of Rässler's validity assessment procedure (Rässler, 2002), namely preserving individual values (level 1), preserving joint distributions (level 2), preserving correlation structures (level 3), and preserving marginal distributions (level 4). In practical situations it is only possible to test the fourth level of validity, as the other levels require knowledge of the true distribution $f(y,z)$. These levels are typically assessed with simulation studies. In general, a fusion is seen as successful if the fourth level of validity is attained (Rässler, 2002).

RESULTS

Level 1: Preserving individual values

The proportion of original records that were recreated or retained in the data fusion is the hit-rate of a fusion. Figure 2 shows the relationship between the hit-rate for all simulations and the qCIA. The negative linear trend in this graph suggests that any deviation from CIA leads to a reduction in the hit-rate of the fusion.

Figure 1: Scatter-plot of quantified CIA by partial correlations

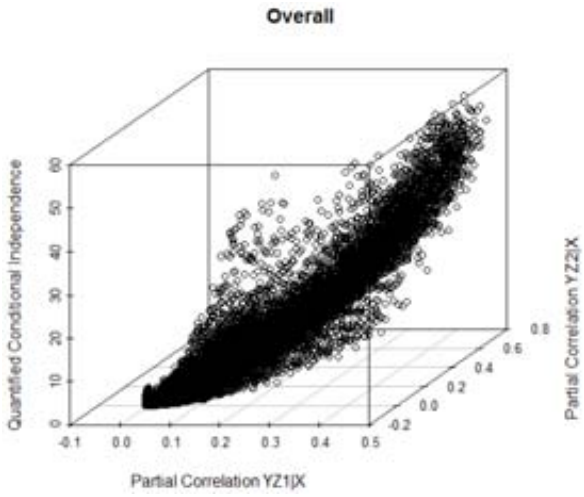
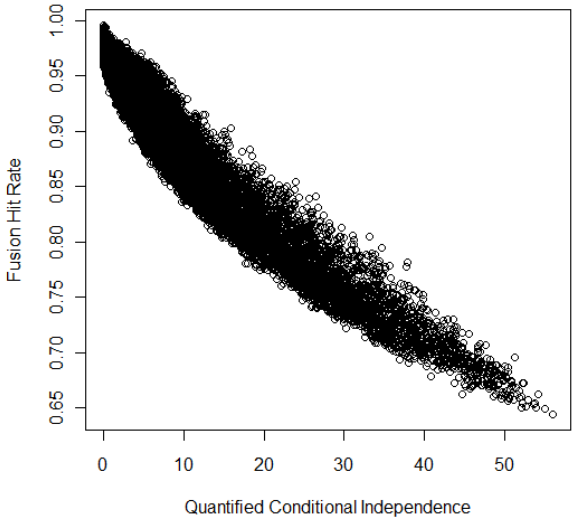


Figure 2: Scatter-plot of fusion hit-rate by quantified CIA



Level 2: Preserving joint distributions

The most important test of a fusion success is the evaluation of the joint distribution of the variables that were never jointly observed. This is done through χ^2 goodness-of-fit tests to compare the fused and original distributions. The p-values of the χ^2 hypothesis tests are grouped into four levels of significance: [0, 0.01], (0.01, 0.5], (0.5, 0.1], and (0.1, 1]. The qCIA measure is categorized into six levels: [0, 1], (1,2], (2,3], (3,4], (4,5] and (more than 5). Figure 3 shows that the fusion deteriorates abruptly as the data deviate from CIA. If the qCIA is ≤ 1, the joint distribution was retained in the fused data for 79.6% of these simulations (p-value > 0.1). Significant differences between the fused and original distributions become apparent when the CIA is no longer a valid assumption. These results indicate that a fusion can only truly be successful if the CIA is true.

Level 3: Preserving correlation structures

The one-sample $\tilde{T}3$ test statistic, proposed by Larntz and Perlman (1985) is used to compare the fused and original correlation structure. The p-values of the $\tilde{T}3$ hypothesis tests and the qCIA are again grouped as described above. Figure 4 illustrates that any deviation from CIA has a negative effect on the quality of the fusion in terms of the correlation structure. The correlation structure was effectively retained in 79.1% of the

simulations for which the qCIA is ≤ 1. For any qCIA > 2 the fusion is unable to accurately reflect the true correlation structure.

Level 4: Preserving marginal distributions

χ^2 goodness-of-fit tests are used to compare the marginal and joint distributions from the fused data with that in the original data. For well over 90% of the 30,000 simulations, the marginal and joint distributions from the individual data sources are retained in the fused files, i.e. the p-values > 0.1. Less than 0.5% of the simulations are significant at the 1% level, about 2% at the 5% level, and approximately 5% at the 10% level. Overall, these results indicate that the minimum requirement for a successful fusion is satisfied.

Figure 3: Mosaic-plot of quantified CIA by χ^2 p-value for (Y, Z)

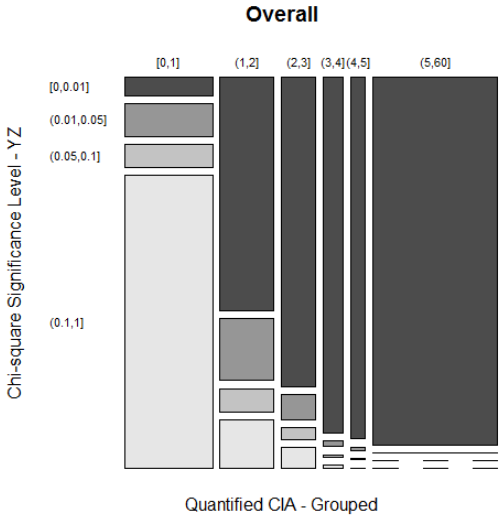
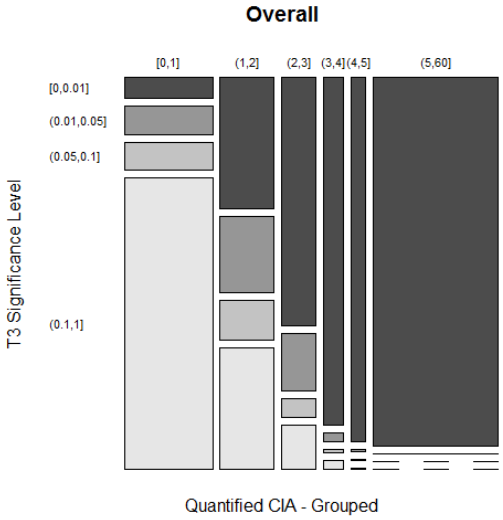


Figure 4: Mosaic-plot of quantified CIA by $\tilde{T}3$ p-value



CONCLUSIONS AND RECOMMENDATIONS

The validity of the CIA in data fusion has raised much concern in the statistical literature. It has been shown by numerous authors that the assumption is perhaps too restrictive to be considered a reliable data fusion methodology. A major drawback of the CIA is that, in practical situations, it is not possible to test whether the assumption is valid or not. Therefore, there is the risk of fusing data based on an incorrect assumption, which leads to a fused data where the true distributions of the data are misrepresented.

When fusing binary data, the fusion is guaranteed to be a success, if the CIA is a valid assumption. All the distributions and data structures are sufficiently retained in the fused data, thereby satisfying all four levels of validity. However, deviations from the CIA have a negative effect on the success of a fusion. Even small deviations do not always produce accurate results.

This really raises the question: how much confidence can the researcher truly have in the validity of a data fusion under the CIA? Although fusion may be the only viable solution to the problem of questionnaire overload, it can only be done if there is sufficient evidence that the required assumptions are satisfied.

Research into data fusion in general, as well as fusing binary data, is far from complete. The CIA is a very restrictive assumption, therefore alternative approaches that do not rely on this assumption should be investigated. In recent years, multiple imputation has been on the forefront of fusion research, with the main focus on continuous variables. Multiple imputation techniques to fuse binary data present important opportunities for further research.

REFERENCES

Alosh, M.A. & Lee, S. J. (2001). *A simple approach for generating correlated binary variates*. J. Statist. Computn Simuln, **70**, 231-255.

Alter, H.E. (1974). *Creation of a synthetic data set by linking records of the Canadian survey of consumer finances with the family expenditure survey 1970*. Ann. Econ. Socl Measmnt **3**(2)**,** 373-394.

Becker, R. & Collins, J. (2007). *Toward total audience – Integrating magazines' hardcopy and internet site audiences using dynamic segmentation fusion*. Presented to the Advertising Research Foundation. www.mediamark.com/PDF/WP%20Toward%20Total%20Audience.pdf

Conti, P.L., Marella, D. & Scanu, M. (2008). *Evaluation of matching noise for imputation techniques based on nonparametric local linear regression estimators*. Computnl Statist. Data Anal. **53**, 354-365.

D'Orazio, M., Di Zio, M. & Scanu, M. (2006). *Statistical matching: Theory and practice*. England: Wiley.

Jakulin, A. & Bratko, I. (2004). *Quantifying and visualizing attribute interactions: An approach based on entropy*. http://arxiv.org/abs/cs.AI/0308002

Larntz, K. & Perlman, M.D. (1985). *A simple test for the equality of correlation matrices*. Technical Report No. 63. www.stat.washington.edu/research/reports/1985/tr063.pdf

Moriarity, C. & Scheuren, F. (2004). *Regression-based statistical matching: Recent developments*. Proc. Surv. Res. Meth. Sect. Am. Statist. Ass. www.amstat.org/sections/srms/Proceedings/y2004/files/Jsm2004-000361.pdf

O'Brien, S. (1991). *The role of data fusion in actionable media targeting in the 1990's*. Marketing & Research Today, **19**, 15-22.

Okner, B.A. (1972). *Constructing a new data base from existing microdata sets: The 1966 MERGE file*. Ann. Econ. Socl Measmnt **1**(3)**,** 325-342.

Radner, D. B., Allen, R., Gonzalez, M. E., Jabine, T. B. & Muller, H. J. (1980). *Report on exact and statistical matching techniques*. Statistical Policy Working Paper 5, Federal Committee on Statistical Methodology.

Raghunathan, T.E. & Grizzle, J.E. (1995). *A split questionnaire survey design*. J. Am. Statist. Ass., **90**, 54-63.

Rässler, S. (2002). *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches*. Lecture Notes in Statistics, 168. New York: Springer.

Rässler, S. & Fleischer, K. (1998). *Aspects concerning data fusion techniques*. ZUMA Nachrichten Spezial **4**, 317-333.

Redway, H. (2003). *Data fusion by statistical matching*. https://guard.canberra.edu.au/natsem/conference2003/papers/pdf/redway_howard-1.pdf

Rodgers, W.L. (1984). *An evaluation of statistical matching*. J. Bus. Econ. Statist., **2**(1), 91-102.

Rubin, D.B. (1986). *Statistical matching using file concatenation with adjusted weights and multiple imputations*. J. Bus. Econ. Statist., **4**(1), 87-94.

Sims, C.A. (1972). *Comments*. Ann. Econ. Socl Measmnt **1**, 343-346.

Soong, R. & De Montigny, M. (2001). *The anatomy of data fusion*. 2001 Worldwide Readership Research Symposium. www.readershipsymposium.org/anatomy-data-fusion

Tchaoussoglou C. & Van der Noort, W. (1999). *Divide and unite – Splitting the SummoScanner and data fusion*. Worldwide Readership Research Symposium. www.readershipsymposium.org/divide-and-unite-splitting-summoscanner-and-data-fusion

Van der Putten, P., Kok, J.N. & Gupta, A. (2002). *Data fusion through statistical matching*. Paper 185, Center for eBusiness@MIT. http://ebusiness.mit.edu

Wolff, E. (1977). *Estimates of the 1969 size distribution of household wealth in the US from a synthetic database*. Conference on research in income and wealth. www.nber.org/chapters/c7448.pdf