An Approach to the Synthesis and Analysis of Incomplete Marketing Data[1]

Akihiro INOUE*

Abstract

In this paper we present a more general approach to the synthesis and analysis of incomplete marketing data (or fusion technique) that allows missing information to be both discrete and continuous and extends the general location model by relaxing the "identically distributed" assumption. We show the application study of the technique to TV rating and panel data sets in Japan. The validation study demonstrates that the proposed technique produces highly satisfactory results. In the last part of this paper we also conceptually compare our new technique with the mixture model proposed by Kamakura and Wedel.

1. Introduction

We confront the issue of incomplete data analyses for various reasons in marketing applications. A traditional possibility is that long time-consuming questionnaires are likely to bore respondents, resulting in missing data. Another occasion, considering the new development of the Internet and the internet-related environment, is web surveys where interactive questionnaires are used, allowing some questions not to be answered. Alternatively, simultaneous analyses of surveys implemented in different areas, over different time-periods, to different respondents are inherently missing data since they do not have common observations. Even though we have several sources of incomplete data analyses, we would be better off using complete data. Such an example is single-source data.

Single-source data, containing both consumer individual information and various marketing activity information, are extremely useful for investigating the effects of marketing actions on consumer behavior, particularly choice. A marketing researcher, for example, wants to examine the effect of ads on the choice of their brands. She therefore needs to acquire the information of ads exposure, such as what kinds of ads exist in the market and how much they are implemented, and consumer surroundings and behavior, such as how

* Associate Professor, School of Business Administration, Kwansei Gakuin University.

[1] This paper modifies Inoue (1999) so as to take into account the more rigid validation tests discussed in section 4 of this paper.

often each consumer sees the ads, and which brands she purchases. The former type of data usually requires designated devices (*e.g.* People Meter©) and is collected by specialized organizations (*e.g.* Video Research, Inc.). The latter type of data also needs certain instruments and recently we usually rely on panel data. Single-source data could be available either in the case where an organization has the accessibility to both kinds of devices and is able to gather the two kinds of data or in the case where multiple organizations concerned agree to cooperate so as to consolidate the two kinds of data.

In spite of its usefulness, the availability of single-source data becomes limited because of its substantial cost in organizational and financial terms. To cope with this difficulty, we often try to combine two sorts of data that are collected individually and independently. Fusion (sometimes we call it file concatenation in statistics, *e.g.* Woodbury 1983; Rubin 1983; 1986; Rodgers 1984; Barry 1988) is a technique we mainly use for this purpose. Fusion techniques basically deal with the incomplete or missing data problem (*cf.* Little and Rubin 1987; Rubin 1976; Liu, Taylor, and Belin 1997; Schafer 1997). Since each dataset is collected independently, one dataset does not have any information of the other and *vice versa*, except for some common variables which are ordinarily demographics. Thus, the crux of the fusion technique is how to fill in the incomplete information of the first data given the common variables and the second data and that of the second data given the common variables and the first data.

In this study, we propose a more general approach to synthesizing and analyzing incomplete marketing data (or fusion technique) that allows missing information to be both discrete and continuous, as opposed to Kamakura and Wedel (1997) where only continuous variables are admissible to be incomplete. We first review the past studies from the non-parametric and parametric perspectives. We criticize the heuristic-based non-parametric fusion procedures with respect to their *ad hoc* and subjective decisions and the impossibility of statistical tests. In the third section, we show our more general fusion technique. Next, we apply the technique to the consolidation of TV rating data and panel data in Japan and assess its validity empirically.

2. Past Fusion Procedures

We can categorize the past fusion techniques into two categories: non-parametric and parametric techniques. We review them respectively. The fusion procedures in the first group heavily depend upon *ad hoc* and subjective decisions. There are three well-known studies. Santini and his colleagues (Antoine and Santini 1987; Baker, Harris, and O'Brien 1989; Adamek 1994; Raimondi and Santini 1997) have developed a fusion technique, called FRF ('Fusion sur Referentiel Factoriel' in French). The FRF algorithm first applies multivariate correspondence analysis to the common variables shared by the two data sets to be consolidated. This first step gives us a restriction of this method in the sense that the variables under study have to be categorical. The second step is to define the set of the K closest points to a subject in the first data based upon the subject scores resulted from the preceding correspondence analysis. Then, the FRF method attempts to marry a subject in the first database with a subject in the second database based on a fairly subjective decision. Santini and his colleagues name these marriage criteria Love at First Sight, Childhood Sweethearts, Adultery, Attentiveness, Convenience, and Shotgun Marriages. Since all of these marriage criteria are quite *ad hoc* and subjective and also the selection does require some experienced skills, we must criticize them because of the lack of rigidness and robustness. Also, this method does not take into account any statistical aspect in density nor sampling, so we cannot perform any statistical tests on data sets consolidated by the FRF method.

The second and third non-parametric procedures are quite similar to the FRF algorithm. The second methods utilize either the ANOVA (the RSMB algorithm by O'Brien 1991) or the regression analysis (Roberts 1991), instead of the multiple correspondence analysis, so as to calculate the distances between the subjects in the first and second databases. Then, the remaining steps basically follow the same paths as the FRF. Wiegand (1986) introduces a third procedure where the cluster analysis is substituted for the multiple correspondence analysis and the consequent groups, instead of distances *per se*, are used to the rest of the FRF paths. Because these two methods rely on similar marriage criteria as the FRF, they lack rigidness and robustness and similarly do not permit us to run any statistical tests.

We can categorize the other fusion methods into the parametric procedure group. In statistics there have been the methods referred as statistical matching. We can perform the statistical matching methods in two ways: constrained matching and unconstrained matching. In the former we do not allow means and variances to be changeable, as opposed to the latter where means and variances are able to vary substantially. Recently, Kamakura and Wedel (1997) introduced a fusion technique. Their approach is based upon a mixture model (*e.g.* Titterington, Smith, and Makov 1985; McLachlan and Basford 1988; Kamakura and Russell 1989; Kamakura and Mazzon 1991; Kamakura and Novak 1992; Wedel and DeSarbo 1994; Kamakura and Wedel 1995). Both the statistical matching and the Kamakura-Wedel approaches overcome the disadvantages of the non-parametric fusion techniques, *i.e.*, the lack of rigidness and the impossibility of statistical tests. However, the Kamakura-Wedel method confines variables to be categorical. This is a tangible and solid restriction since the variables subject to or related to marketing decisions are not necessarily discrete, rather continuous. On the other hand, against the statistical matching techniques the Kamakura-Wedel procedure has an advantage that we can use the multiple imputation method (*e.g.* Little and Rubin 1987;

Rubin 1987; Meng and Rubin 1992).

In this study we propose a new fusion procedure where both continuous and discrete variables are allowed to be incomplete and we can run the multiple imputation. We show the detail of the procedure in the next section.

3. A New Fusion Procedure

We address a new fusion technique that permits that two data sets to be fused contain both continuous and discrete variables. The procedure is based upon the general location model (Little and Schluchter 1985; Cooper, Klapper, and Inoue 1996) and extends it so as to relax the identically-and-independently-distributed assumption. First we specify the general location model, then the new fusion procedure.

Suppose that the two data sets under study consist of K continuous variables and V categorical variables. Categorical variable j has I_j levels, so that the categorical variables define a V-way contingency table with $C = \prod_{j=1}^{V} I_j$ cells. Let x_i be the K-vector of continuous variables and y_i the V-vector of categorical variables for subject i. C-vector w_i, constructing from y_i, equals E_m in case subject i belongs to cell m of the contingency table, where E_m is a binary C-vector. The general location model is defined as follows: 1) The w_i are iid multinomial random variables with cell probabilities $\Pr(w_i = E_m) = \pi_m$ where $m = 1, ..., C$ and $\sum_{m=1}^{C} \pi_m = 1$. 2) Given that $w_i = E_m$, $(x_i \mid w_i = E_m) \overset{iid}{\sim} = N_k(\mu_m, \Omega)$.

The new fusion procedure simply relaxes the "identically distributed" assumption stated in the conditional multivariate normal distribution, while holding the "independently distributed" assumption. That is, in our new fusion approach, given that $w_i = E_m$, $(x_i \mid w_i = E_m) \overset{iid}{\sim} = N_k(\mu_m, \Omega_m)$.

4. Application to Consolidation of TV Rating and Panel Data in Japan

We attempted to impute TV rating values into panel data by the new fusion technique mentioned above[1]. We describe the data sets respectively.

The panel data consist of 2652 members and we employed only the third week of May in 1997. The data, having been collected through the questionnaire, cover the various aspects

[1] We have also considered fusing the variables of the panel data into TV rating data set but decided not to do so for two reasons. First, the number of variables of the panel data and their diversity were far greater than those of the TV rating data. Second, it seemed more natural to conceive the conditional density of the TV rating variables given the panel data such as demographic variables, rather than the conditional density of the panel data variables given the TV rating data set.

ranging not only purchase behavior but also demographics, involvement, usage behavior, possession status, and so on. Our primary concern was with the involvement toward diverse product categories and to construct the different TV rating sets according to the degree of involvement even though we must confine the result. For example, it is notably common to argue about TV ratings by gender and/or age groups (*e.g.* F1 for 20-29 year old female, M2 for 30-44 year old male). However, with the current infrastructure, it is almost impossible to calculate TV ratings by involvement levels. It would be more efficient to design TV ad plans based upon the involvement-wise TV ratings than the regular gender-age TV ratings. Since the results contain miscellaneous confidential aspects, we cannot show them in this paper. We present only how valid our new fusion technique was.

The TV rating data are composed of 1489 households and a device (People Meter©) allows us to collect TV viewing data individually. We selected the same week as the panel data. The TV rating data contain 840 TV viewing variables (5 TV stations, 7 days in the third week of May in 1997, and 24 hours) and demographic variables (m = 840).

We attempted to consolidate the TV rating variables into the panel data based upon the common demographic variables: gender, age, occupation, and marital status. First, we categorize the age variable into 6 groups (-20, 21-29, 30-39, 40-49, 50-59, 60-). By processing the age in this way, we could treat all the shared variables as discrete variables, leaving the TV ratings as continuous, and simplify the estimation algorithm based upon the "independently distributed" assumption and the factorization of the likelihood (*e.g.* Little and Rubin 1987). Since we had 2 gender classes, 6 age groups, 10 occupations, and 2 marital statuses, we theoretically had 240 cells but put together some cells so as to hold a certain substantiality of observations in a cell with a reasonable justification for the union. It turned out to be 33 classes (C = 33). It should be noted that we did not employ the gender and age variables solely nor the complete information of these variables[2]. Since we modeled the conditional multivariate normal distribution given a certain class to be independently and non-identically distributed, we had only to estimate the mean and variance parameters for each class by the following E-M algorithm (*e.g.* Dempster, Laird, and Rubin 1977; Little and Rubin 1987):

E-Step:

$$E\left(\sum_{i=1}^{n} y_{ij} \mid Y_{obs}, \theta^{(t)}\right) = \sum_{i=1}^{n} y_{ij}^{(t)},\ j = 1, \ldots, K$$

$$E\left(\sum_{i=1}^{n} y_{ij} y_{ik} \mid Y_{obs}, \theta^{(t)}\right) = \sum_{i=1}^{n} (y_{ij}^{(t)} y_{ik}^{(t)} + c_{jki}^{(t)}),\quad j, k = 1, \ldots, K$$

[2] This point is absolutely important because later we run the validation study based on the gender-age groups.

where

$$y_{ij}^{(t)} = \begin{cases} y_{ij} & \text{for observed } y_{ij} \\ E(y_{ij} \mid y_{obs,\, i},\, \theta^{(t)}) & \text{for missing } y_{ij} \end{cases}$$

$$c_{jki}^{(t)} = \begin{cases} 0 & \text{for observed } y_{ij} \text{ or } y_{ik} \\ Cov\,(y_{ij},\, y_{ik} \mid y_{obs,\, i},\, \theta^{(t)}) & \text{for missing } y_{ij},\, y_{ik} \end{cases}$$

M-Step:

$$\mu_j^{(t+1)} = n^{-1}\sum_{i=1}^{n} y_{ij}^{(t)} \; j = 1, \dots, K$$

$$\sigma_{jk}^{(t+1)} = n^{-1}E\left(\sum_{i=1}^{n} y_{ij}y_{ik} \mid Y_{obs}\right) - \mu_j^{(t+1)}\mu_k^{(t+1)}$$

$$= n^{-1}\sum_{i=1}^{n}[(y_{ij}^{(t)} - \mu_j^{(t+1)})(y_{ik}^{(t)} - \mu_k^{(t+1)}) + c_{jki}^{(t)}],\ j,\ k = 1, \dots, K$$

As the above equations imply, we can estimate the parameters easily.

We varied the number of multiple imputations from 20 to 80 by 30. However, the results with only 20 imputations were surprisingly competent, so that we do not show the validation results with 50 imputations but those with 20 and 80 imputations. Theoretically, we know that the relative efficiency of a point estimate based on m imputations to one based on an infinite number of imputations is approximately $\sqrt{1+\lambda/m}$ where λ is the missing proportion (Rubin 1987).

The validation scheme was as follows. First, we performed the consolidation of the TV rating variables into the panel data by the new approach mentioned above. The fusion procedure was based upon the 33 demographic cells shared by the both data sets and the 840 TV rating variables (5 TV stations, 7 days in the third week of May in 1997, and 24 hours). The imputation was replicated either 20 or 80 times.

Second, we run the 840 statistical tests with respect to the null hypothesis where the mean value of an imputed TV rating variable of the panel data is equal to the mean value of the original TV rating variable. We carried out the 10 cases composed of 2 gender-classes and 5 age groups for each of the 840 tests. It should be remarked that this kind of statistical test would not be appropriate if we were to run it based upon the same condition as used in the fusion because, considering the nature of the parameter estimation, the fused variables apparently tend to produce the same marginal densities and frequencies as the original variables. For example, we would not expect to reject the null hypotheses if the fusion were implemented based upon the particular number of gender-age-occupation classes and the statistical tests were run based on the same gender-age-occupation classes. However, in the fusion step we imputed the TV rating values on the basis of 33 cells constructed not with the 10 age-gender classes but with gender-age-occupation-marital status classes and the partial information of the age-gender classes. Thus, we can verify the validation tests base upon 10 age-gender cases.

Third, we implement the 840 equal-mean-null-hypothesis statistical tests at the aggregate versus target levels with respect to the parameter-free (PF), expectation (Reg), and our models.

Table 1 Validation Test at Aggregate Level

	PF Model	Reg Model	Our Model
20 Imputations	**90.4%**	**99.6%**	**100%**
80 Imputations	**85.2%**	**99.6%**	**100%**

Table 1 shows the validation result at the aggregate level. Each cell exhibits the proportion of the statistical tests accepting the null hypothesis out of 840 cases. As an instance, with regard to the expectation (Reg) model with 20 imputations, 99.6% of the 840 TV rating variables did not reject the null hypothesis, indicating that the almost all mean values of the fused variables of the panel data are equal to those of the original TV rating variables. As Table 1 implies, we can comprehend the high (perfect) validity of our new fusion procedure. This could be easily predicted. Since we estimate the parameters of the proposed model based upon the all observations, one might argue, the likelihood of the data fitting would be innately high. However, it should be noted that all three (PF, Reg, our) models are calibrated and their validity compared under the same conditions using the whole dataset.

Next, we conducted a more rigorous test by executing at the target level composed of gender and age classes. Table 2 shows the results of male (upper) and female (lower) classes. As well as Table 1, each cell of Table 2 exhibits the proportion of the statistical tests accepting the null hypothesis out of 840 cases. There exist some points to be remarked on. First, the proposed model performs much better than the other models, especially compared to the expectation model that works as well as our model at the aggregate level. As we mentioned above, the scheme we use at the target level is not same as that we employed in estimating the parameters of the models. Hence, we can confirm the superiority of the proposed model on the ground that the test at the target level is more rigid than the one at the aggregate level.

Table 2 Validation Test at Target Level

	Male	PF	Reg	Our
		Model	Model	Model
20 Imputations	10's	.817	.856	.986
	20's	.705	.760	.988
	30's	.777	.823	.989
	40's	.848	.896	.998
	50's	.816	.841	.991
80 Imputations	10's	.833	.854	.985
	20's	.700	.767	.991
	30's	.782	.829	.994
	40's	.856	.900	.999
	50's	.818	.838	.993

	Female	PF	Reg	Our
		Model	Model	Model
20 Imputations	10's	.810	.841	1.00
	20's	.841	.877	.996
	30's	.916	.932	.998
	40's	.898	.912	.995
	50's	.856	.871	.989
80 Imputations	10's	.826	.846	1.00
	20's	.833	.879	.995
	30's	.925	.929	.996
	40's	.895	.913	.994
	50's	.861	.869	.988

Second, it should be noted that, regarding all three models, we do not see any substantial difference between 20 and 80 imputation cases. As the number of the multiple imputations gets larger, the subsequent application analyses based on the imputations become more laborious and require more computational efforts. As far as this study is concerned, it seems that 20 imputations are sufficient.

5. Summary and Discussion

We presented a more general approach to synthesize and analyze marketing incomplete data (or fusion technique) that allows missing information to be both discrete and continuous and extends the general location model by relaxing the "identically distributed" assumption. We showed the application study of the technique to the TV rating and panel data sets in Japan. The validation study demonstrated that the proposed technique produced highly satisfactory results.

In the last part of this paper let us conceptually compare our new technique with the mixture model proposed by Kamakura and Wedel. In our procedure we do not use any latent class, unlike Kamakura and Wedel. Instead of the mixture, we implicitly take into account the independent classes of 33 cells composed of gender-age-occupation-marital status. We conjecture that these tacit classes functioned in a similar way as the latent classes would do. Thus, in a case where we could mold reasonable discrete classes with sound *a priori* criteria, we may not need to perform a mixture model that has some shortcomings such as multi-modality of parameter space.

References

Adamek, J. (1994), "Fusion: Combining Data From Separate Sources," *Marketing Research: A Magazine of Management and Applications*, 6 (Summer), 48-50.

Antoine, J. and G. Santini (1987), "Fusion Techniques: Alternative to Single Source Methods?" *European Research*, 15 (August), 178-87.

Baker, K., P. Harris, and J. O'Brien (1989), "Data Fusion: An Appraisal and Experimental Evaluation," *Journal of the Market Research Society*, 31, 2, 152-212.

Barry, J. T. (1988), "An Investigation of Statistical Matching," *Journal of Applied Statistics*, 15, 3, 275-283.

Cooper, L. G., D. Klapper, and A. Inoue (1996), "Competitive-Component Analysis: A New Approach to Calibrating Asymmetric Market-Share Models," *Journal of Marketing Research*, 33(May), 224-38.

Dempster, A. P., N. M. Laird, and R. B. Rubin (1977), "Maximum Likelihood for Incomplete Data via the EM-Algorithm," *Journal of the Royal Statistical Society, Series B*, 39, 1, 1-38.

Inoue, A. (1999), "A General Fusion Technique and Its Application to Creating Quasi-Single Source Data," *28th EMAC Conference Proceedings*. Berlin, Humboldt University.

Kamakura, W. A. and G. J. Russell (1989), "A Probabilistic Choice Model for Market Segmentation and Elasticity Structure," *Journal of Marketing Research*, 26 (November), 379-90.

Kamakura, W. A. and Jose Alfonso Mazzon (1991), "Value Segmentation: A Model for the Measurement of Values and Value Systems," *Journal of Consumer Research*, 18 (September), 208-18.

Kamakura, W. A. and Thomas P. Novak (1992), "Value-System Segmentation: Exploring the Meaning of

LOV," *Journal of Consumer Research*, 19 (June), 119-32.

Kamakura, W. A. and M. Wedel (1995), "Life-Style Segmentation With Tailored Interviewing," *Journal of Marketing Research*, 32 (August), 308-17.

Kamakura, W. A. and M. Wedel (1997), "Statistical Data Fusion for Cross-Tabulation," *Journal of Marketing Research*, 34 (November), 485-98.

Little, R. J. A. and D. B. Rubin (1987), *Statistical Analysis with Missing Data*. New York: NY: John Wiley and Sons.

Little, R. J. A. and M. D. Schluchter (1985), "Maximum Likelihood Estimation for Mixed Continuous and Categorical Data with Missing Values," *Biometrika*, 72, 497-512.

Liu, M., J. M. G. Taylor, and T. Belin (1997), "Multiple Imputation and Posterior Simulation for Multivariate Missing Data in Longitudinal Studies," working paper, the department of biostatistics, University of California, Los Angeles.

McLachlan, G. and K. E. Basford (1988), *Mixture Models: Inference and Applications to Clustering*. N. Y.: Marcel Dekker Inc.

McLachlan, G. J. and T. Krishnan (1997), *The EM Algorithm and Extensions*. New York, NY: John Wiley & Sons, Inc.

Meng, X. and D. B. Rubin (1992), "Performing Likelihood Ratio Tests with Multiply Imputed Data Sets," *Biometrika*, 79, 1, 103-11.

O'Brien, S. (1991), "The Role of Data Fusion in Actionable Media Targeting in the 1990s," *Marketing and Research Today*, 19 (February), 15-22.

Raimondi, D. and G. Santini (1997), "Just In Time Data Modeling," in *Worldwide Readership Research Symposium 8*. Vancouver.

Roberts, A. (1991), "Media Exposure and Consumer Purchasing: An Improved Data Fusion Technique," *Marketing and Research Today*, 22 (August), 159-72.

Rodgers, W. L. (1984), "An Evaluation of Statistical Matching," *Journal of Business and Economic Statistics*, 2 (January), 91-105.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581-92.

Rubin, D. B. (1983), "Discussion of 'Statistical Record Matching for Files' by M. A. Woodbury", in *Incomplete Data in Sample Surveys*, vol. 3. W. G. Madow, H. Nisselson, I. Olkin (eds.), 203-205.

Rubin, D. B. (1986), "Statistical Matching and File Concatenation with Adjusted Weights and Multiple Imputations," *Journal of Business and Economic Statistics*, 4, 1, 87-94.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons, Inc.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*. New York, NY: Chapman & Hall. (http://www.stat.psu.edu/~jls/)

Titterington, Rubin, D. B., A. F. M. Smith and U. E. Makov (1985), *Statistical Analysis of Finite Mixture Distributions.* New York, NY: John Wiley and Sons.

Wedel, M. and W. S. DeSarbo (1994), "A Review of Recent Developments in Latent Class Regression Models," in *Advanced Methods of Marketing Research.* R. P. Bagozzi, ed. Cambridge, MA: Blackwell Publishers, 352-88.

Wiegand, J. (1986), "Combining Different Media Surveys: The German Partnership Model and Fusion Experiments," *Journal of the Market Research Society*, 28, 2, 189-208.

Woodbury, M. A. (1983) "Statistical Record Matching for Files", in *Incomplete Data in Sample Surveys*, vol. 3. W. G. Madow, H. Nisselson, I. Olkin (eds.), 173-181.