INTERACTIVE FUSION: STEP TWO

Uwe Czaia, Immediate GmbH, Bremen

At the last symposium in Hongkong we presented our first attempt at an interactive, 'real time' fusion model for data derived from independent samples.

Then as well as now, we aimed for the development of an expert system, a tool for the user, i.e. a marketing person or a media planner, meeting increased requirements of extremely complex data and with multifold inter-connexions.

There is a demand for intra- and, to a lesser degree, inter-media runs to optimise media schedules with a large number of existing and new advertising vehicles. The target groups will not only be defined through product consumption data and/or socio-demographics, but also through a complete range of often inter connected qualitative criteria, media usage patterns and more.

Rolf Speetzen will present one part of these complex requirements in another session of the symposium. It's called 'APX - The New Deutsche Mark in Media Buying' and shows what has been accomplished within the German MA.

Starting-point and goal have not changed during the past years, despite new research techniques such as CATI and others. Single source data meeting the complex requirements cannot be obtained at a realistic cost.

20 TV stations, 30 radio stations, approximately 600 daily newspapers, roughly 200 major magazines plus numerous city magazines and even more local free sheets on the one hand, and a large number of target group definitions on the other, are far too much for individual single source interviews. It would result in one day seminar rather than interviews, just to stem the deteriorating effect on the quality of the data. It might also get the attention of Big Brother, who has sofar not concerned himself with media data. And any way, do we want it?

One solution was to instead fuse data from the different data bases, using mathematical-statistical models. This becomes easier and easier due to the rapid developments of micro electronic and the great efficiency of modern PCs, now comparable to main frames.

Step I of our Interactive PC Fusion technique was presented in Hongkong. We would like to, however recap a few important points. Real time fusion can be achieved with standard hardware, for example MS DOS/Windows compatible 486 PC with a 500 MB hard disc drive. Dependent on individual needs, selected criteria will be fused under optimum conditions and treatment of the structure of the information to be transferred.

Selected criteria from various data bases can be matched as required, under optimum structural conditions for the data selected.

Our software, developed by 'immediate', Bremen, is similar to the media plan evaluation programs, however it uses highly condensed original data. This ensures justifiable cpu units.

For the following example we needed less than 20 minutes on a standard PC equiped as described earlier. In total some 150 million distance measurement comparisons were necessary. Those criteria with the smallest difference were fused in each case. Donors once used were marked, and only used a second time, if no other donors were available anymore. The transfer frequency is defined through the donor/recipient ratio.

Step I was based on the Mahalanobis distance for similarity measurement, thus excluding possible correlations between individual variables. All links are getting the same weight for the fusion process. And in addition, the Mahalanobis distance measurement is scale invariant.

This creates the basis for an optimum and automated fusion, which is not influenced by whether criteria Z from Data Base A is transferred to Data Base B, or criteria Y the other way round.

For our example we used - comparable to our presentation in Hongkong - criteria from the German VA (Consumer Analysis), 12,000 respondents, and fused it into the print media portion of the MA, with it's 20,000 respondents.

TABLE 1:

Consumption of wine	Donor %	Recipient %
daily/nearly daily	1.9	1.8
several times a week	8.7	8.9
approximately once a week	15.0	14.8
approximately every two weeks	11.0	11.6
approximately once a month	12.8	13.2
'One should enjoy life and spend ones money rather than saving it'		
Agree ...		
... completely	10.3	10.8
... mostly	35.4	34.9
... not quite	38.5	38.0
... not at all	15.8	16.3

There are small differences between the two sets of results. They do not stem from inadequate transfer, but rather from differences in the structure of the common criteria (sex, age, residential area, size of city, education, profession). The structure of the recipient data base determines the structure of the fused data.

The fusion technique in Step I allowed the construction of individually matched data bases, however, one must bear in mind that this can lead to the creation of different data bases for similar planning tasks within an agency. If this does occur it can be due to the fact that in one instance an additional variable, such as 'beer consumption' was been used in *one* run but not in the other.

In addition, by using a variable-related real-time fusion, we must be aware that the results may not be exactly repeatable, e. g. different runs may not show identical results, even if just the sorting of the data has been changed from fusion to fusion. While the differences of the results are marginal and only visible as decimal fractions, comparability is however limited.

For a statistician this is a rather peripheral problem. None the less we are presently working on a solution by changing algorithms for the optimisation.

For planners, however, this is different. In their environment of competing agencies, he or she requires absolutely comparable data, and he or she cannot refer to tolerance ranges of efficiency scores, even though this also occurs in a single source survey.

The focus therefore is more on the construction of highly complex 'quasi' single source data bases rather than on a flexible integrative use of a number of sources simultaneously.

This is the starting point for Step II: an additional Interactive Fusion program for PCs, however, constructing complete and uniform sets of data for common use under equal conditions, instead of creating parts of data bases in an individual planning situation.

This extension led to a modification of the methods. As an alternative to the Mahalanobis distance, one can also choose a process based on the Euclidean distance measurement. The Mahalanobis distance was adequate for the fusion of a smaller number of variables, because the data set could be prepared according to the requirements of the process. With highly complex criteria dimensions of a large number of variables to be transferred, the Mahalanobis distance proved less optimal, because the theoretical conditions for use were only met in exceptional cases, and could in practice not be controlled sufficiently well.

The matrix inversion became a problem in cases that lacked independence of individual columns of a matrix. The then rather large values of the inverted matrix led to inaccuracies in the distance calculation. This reduced the quality of the fusion, for example, the transfer of probabilities for individual radio stations and time intervals, where the sample size was very small.

The following overview demonstrates the different designs of Step I and Step II of the PC Interactive Fusion program.

Original data bases $A_1 \ldots A_n$

Analysis of common criteria/compatibility/recoding

Rearrangement of data for basic software system "M"/"F" for media planning

Step I		**Step II**
Step I: real time fusion of some variables		Step II: construction of a quasi single-source data base
	Designation of recipient data Designation of donor data	
	Test: fusion by comparison of samples/survey methods/donor:recipient ratio etc	
Selection of variables for transfer		Analysis of scope and topics for transfer
Calculation of basic structure in view of common criteria		Cluster analysis over common and specific variables within the donor data base
Matrix inversion/distance measurement		Coordination of common variables of recipient data base and clusters
Coordination of donors and recipients		Retransfer into donor data base, because differences in the structure could move the cluster centers
Data output		Construction of a similarity matrix for the individual clusters
		Calculation of similarities and construction of a coordination matrix for each individual field to be transferred. Iterative testing each coordination whether another combination would reduce the residuals in total and swapping if needed
		Data output

The second method to constructing 'quasi' single-source data looks very complicated on first sight. However, the PC user version is menu driven which makes it a lot easier to handle. In addition we are working on a next version under windows, which will even be easier to use.

We assume that the individual user is at least 'expert' in terms of the data structures of the data bases to be fused. This will ensure that the selection of alternatives and the combination of different methods for the transfer of the information will lead to an optimum result namely, a new and extended planning data base.

In either case (Step I and Step II) can the resulting data be delivered for immediate application on a PC, such as planning, evaluation and the like.

The need for generally accepted data bases has increased considerably in Germany over the past two years. This summer, the existing two consumer analyses 'Verbraucher-Analyse' (sponsored by Axel Springer, Bauer Verlag, SAT-1 and PRO 7) and 'Typologie der Wünsche' (sponsored by Burda) will be enhanced by adding detailed TV people meter data on TV usage.

Step II of our Interactive Fusion program, as outlined before, stood the test when it was used to transfer original amount read data from the parallel wave of the German MA into the print portion of the MA, as a preliminary step towards the calculation of APX. The base was data from two independent samples.

In addition, Interactive Fusion (Step II) was successfully used to complete lacking data within a sample, thus creating a data base which could be taken as single-source. Let us look at the Austrian Consumer Analysis. Last year, 'immediate', Bremen, fused consumption data from a self-administered questionnaire which was a part-sample of the Austrian MA, into the total data base:

Total sample of Austrian MA: 14,237 interviews

- Interest in product information
- Media consumption
  - Daily newspapers
  - Regional weekly newspapers
  - General interest magazines
  - Business-/Cultural magazines
  - Women's magazines
  - Special interest magazines
  - TV program supplements
  - Radio
- TV
- Investment
- Household equipment
- Pets
- Car ownership
- Psychographics
- Socio-demographics

Part sample: Consumer Analysis 8,114 interviews

Specific variables

* Sweets, snacks
* Beverages
  - Non-alcoholic
  - Alcoholic
* Dairy products
* Pets
* Personal care
* Cosmetics (Men)
* Decorative Cosmetics (Women)
* Treatment (Women)
* OTC drug products
* General buying habits
* Leisure time, travelling, vacations
* Car ownership
* HiFi, TV and video equipment
* Convenience products
* Cleansers
* General attitudes, consumption habits, concern for environment

A first step was a cross comparison of common criteria of various parts of the samples, 8000 donors, 6000 recipients resp. 14000 respondents in total, versus donors and recipients, to check for comparability.

As a next step we ran 20 product specific cluster analyses within the donor data base, separately for men and women, common as well as specific variables. This ensured, that individual consumption patterns of the households remain intact, the same way as in Step II for a smaller number of variables.

The centroids of the 40 cluster analyses with 10 to 30 clusters were transferred into the recipients data set. The fusion took place - separate for men and women - if there were enough donors within the individual clusters. Donors of adjacent clusters were used only in very few exceptional cases. This turned out to be one of the most important advantages of an Interactive Fusion on a PC: results can be evaluated and checked immediately, especially if media schedules are evaluated.

Finally a sort statement on the quality of the fusion by using chi-squares. We ran 3153 comparisons each for men and women: On a 5% resp. 10% level, no significant deviations could be measured!