

NEAREST NEIGHBOR HOT-DECK IMPUTATION FOR MISSING VALUES WITH SAS/IML

Dr. Thomas W. Sager
The University of Texas at Austin

and

James P. Gise

Dr. M.W. Hemphill

Texas Air Control Board
Austin, Texas

ABSTRACT

Dealing with missing values continues to challenge statisticians. In this paper we examine the application of one modern missing value technique, *nearest neighbor hot-deck imputation* (NNHDI), to one large data set in which about 15 percent of the data are missing or incomplete. Extensive computer processing is required, for which SAS IML (Interactive Matrix Language) provides a compact implementation. Each observation in the data set is a vector of related measures, of which one or more components may be missing. NNHDI involves imputing, or filling in, missing components in a target observation from a complete donor observation whose components closely match the nonmissing components of the target. IML code, provided in the text, is used to compute similarity indices between observations, search for close matches, and impute values from the donor observation to the target observation. When applicable, NNHDI avoids the understatement of Type I error probability that regression and mean-based missing value methods are prone to, avoids having to assume a parametric model as in most versions of the EM algorithm, avoids having to assume missing-at-random missing values, and facilitates extreme value analysis by preserving the variability of components. But NNHDI does require a large donor data set of complete observations and it is computationally intensive.

Dealing with missing values continues to challenge statisticians. The statistical literature on missing values is currently enjoying vigorous growth. Little and Rubin [1] present numerous references, as well as the theory underlying the major approaches.

The authors recently collaborated on a study for the Texas Air Control Board (TACB) of a large data set that illustrates how modern statistical methodology for missing data can integrate with intensive computation to meet satisfactorily the challenges of abundant missing data. In the study, it was necessary to create a complete data set prior to addressing the main research question. The methodology of nearest neighbor hot deck imputation (NNDHI) was implemented in SAS IML [2] to supply values for missing data, thus completing the data set.

The main issue in the study was whether there were time trends in ozone from 1980-87 at 21 sites in six Texas areas (Houston, Dallas - Ft. Worth, El Paso, Beaumont - Port Arthur - Orange, San Antonio, and Austin). Ozone is a major urban air pollutant. The first three named Texas areas are in violation of the National Ambient Air Quality Standards (NAAQS) for ozone. Houston is consistently ranked high among U.S. urban areas in the severity of its ozone problem. Concern over the effectiveness of control programs phased in over the years have prompted numerous studies to determine the resultant change (if any) in ozone levels over time.

Analysis of the time series is complicated by the prevalence of missing data. Measurements are taken hourly in each area throughout the day. But equipment breakdowns, maintenance, and other problems often result in one or more hours being missing. Sometimes whole days are missing. The crucial statistic for NAAQS is the maximum hourly ozone measurement for a day: no area should exceed 12 parts ozone per hundred million on more than three days in any continuous three year period. There is no assurance that the daily maximum has been observed if one or more hours are missing, particularly if the missing values occur in the afternoon, when ozone is usually highest.

The eight-year span of the study encompassed 2,992 days. Using 12 hourly measurements per day (9 AM to 9 PM), there were 35,904 possible hourly measurements per site. Altogether, 10 to 20 per cent of the 35,904 possible measurements were missing for most of the sites. Problems with interpretation of the data can occur if the pattern of missing data is related to the magnitude of the ozone concentrations. Even if the data were missing completely at ran-

dom, in a manner independent of the ozone measurements being made, their absence would still make holes in the time series that comprises the data. These interruptions can violate the assumptions that most time series analyses are based on, thereby rendering suspect any analyses that simply omit missing data.

The Statistical Algorithm

To address the problem of missing ozone measurements, a statistical method for determining replacement values was devised. Our selection of this approach was motivated by three factors: (1) most of the observations were complete, so a large donor set was available from which reasonable replacement values could be chosen; (2) previous experience led us to mistrust parametric modelling for these data, on which the EM algorithm [1] could have been employed; and (3) substitution of predicted values from regression or substitution of the mean of non-missing observations both artificially reduce the variability of the data, thus leading to too many Type I errors and too low estimates of the number of violation days.

The NNDHI method developed for dealing with missing ozone values is one of a class of missing value techniques involving imputation, that is, the substitution of other values for the missing values. The method devised falls into the category called *nearest neighbor hot-deck imputation* by Little and Rubin. The term *nearest neighbor* is applied because for each target day with one or more missing hours of ozone measurements, the method finds that complete donor day which is most similar to the target day. The measure of similarity used here is pattern matching between target and donor for their nonmissing ozone hours. The idea is that the missing hours in the target are likely to have been similar to the corresponding hours in the donor if the pattern of nonmissing data in target and donor are similar. Nonmissing values from the most similar donor day are then imputed (substituted) for the missing hours in the target day. The method is called *hot deck* because the data set donating the imputed values is the same data set on which the analysis will be conducted (as opposed to a cold deck, in which the donor is some previous data set which will not be used in the current analysis). The imputation is site-specific. That is, donor values are all measurements taken at the same site as the target.

Although it seems intuitive that NNDHI will impute reasonable values when there are not many missing hours in a day, it is also intuitive that it will do no better than guessing if

most of the hours in a day are missing. Therefore, we split the data at a site into three sets. The first (DONOR) set consists of those days which have a complete set of 12 ozone measurements for every day. NNDHI will impute values from DONOR to the missing values in each of the other two data sets.

The second set (DIRECT TARGET) consists of those days which have relatively complete ozone values. Values for the missing hours in each day of DIRECT TARGET are imputed from DONOR by NNDHI as follows: The non-missing hours of a DIRECT TARGET day are compared with the corresponding hours of every day in the DONOR set and a score is computed for each DONOR day to measure how close it is to the DIRECT TARGET day. The score is a weighted sum of the differences between corresponding pairs of hourly ozone measurements:

$$S_{ij} = \sum_{k=1}^{12} |o_{ik} - d_{jk}| w_k$$

where S_{ij} is the score from comparing DIRECT TARGET day i having measured ozone values o_{ik} , $k = 1, \dots, 12$, with DONOR day j having measured ozone values d_{jk} , $k = 1, \dots, 12$; and the w_k , $k = 1, \dots, 12$ being the weights applied to the 12 hours, the summation excluding those hours for which the ozone is missing in the DIRECT TARGET day. The weights give more emphasis to differences in the 1:00 pm - 4:00 pm time frame when ozone values are more likely to be elevated. The weights are determined adaptively, in proportion to the frequency distribution by hour of the daily maximum ozone value. If a DONOR day perfectly matches the pattern of non-missing values in the DIRECT TARGET day, the score will be zero. The DONOR day with the minimum score is selected and its ozone values are substituted into the corresponding missing ozone values in the DIRECT TARGET day. However, the non-missing values in the DIRECT TARGET day are not replaced. When the scores of two or more DONOR days are identical, the earliest DONOR day is used.

The third set (INDIRECT TARGET) are those days with very few hourly ozone values measured. In fact, the majority of INDIRECT TARGET days have all 12 hourly ozone values completely missing. Attempting to match on ozone would provide little more than random matching. Instead, values were imputed from DONOR to INDIRECT TARGET indirectly, by matching hourly temperature patterns instead of hourly ozone patterns. Temperature is a

useful correlate of ozone. The temperature data are generally more complete than the ozone data for all sites. The temperature values for an INDIRECT TARGET day are matched against the corresponding temperature values of each ozone-complete DONOR day using the same scoring function described above, but with temperature differences replacing ozone differences. The DONOR day which best matches the temperature pattern of the INDIRECT TARGET day is selected, and the ozone values of the selected DONOR day are substituted for any corresponding missing ozone values of the INDIRECT TARGET day.

The classification of a missing value day into DIRECT TARGET or INDIRECT TARGET was based upon an examination of the distribution of target days by number of hours missing and upon our appraisal of the usefulness of temperature as a correlate of ozone. We chose to classify a day as DIRECT TARGET if it were missing 1 - 8 hours of ozone. It was classified into INDIRECT TARGET if it were missing 9 - 12 hours of ozone.

Advantages.

There are several advantages to this approach to missing values. First, as noted above, omitting the missing values from the analysis could impair interpretation of the time series structure of the data. Second, parametric approaches to missing values such as the EM algorithm [1] require confidence in a parametric model for the air pollution data. Previous work [3] has weakened the authors' confidence in such parametric models for this application. Third, regression and other averaging techniques for supplying estimates of missing values suppress variability. Thus, confidence intervals based on analysis of data with "averaged" estimates for missing values will be misleading because of the "regression to the mean" phenomenon. NNHDI preserves variability because actual data are being substituted for missing values. Fourth, there are enough days which are complete so that a close match can probably be found for most patterns of missing data. Fifth, even if the missing data are rather unlike the complete data, this technique is likely to impute relatively unbiased estimates for the missing values. For example, suppose that most TARGET days tend to be high ozone days. Then NNHDI will be looking for high ozone days in the DONOR set to match the pattern of remaining high ozone hours in the TARGET day, and is more likely to find a good match among the high DONOR days, however many there may be, than among the low DONOR days. This conjecture has been checked by simulation. Finally, the computer code for imputation is easily implemented in PROC IML of SAS (Statistical Analysis System).

The SAS Code

This section contains the core SAS/IML subroutine that performs the imputation. Considering what it achieves, it seems fairly compact. The SAS statements are numbered for convenient referral in the discussion that follows.

```

1  START IMPUTE(TARGET, MTARGET, DONOR, MDONOR, WEIGHTS, IMP);
2    ROWTARG =NROW(TARGET);
3    ROWDONOR=NROW(DONOR);
4    DO I=1 TO ROWTARG;
5      RVMFIT=MTARGET(|I,1:12|);
6      MWORK=REPEAT(RVMFIT,ROWDONOR,1);
7      MWORK=ABS(MWORK-(MDONOR(|,1:12|) # (MWORK ^= . )));
8      MWORK=MWORK # REPEAT(WEIGHTS,ROWDONOR,1);
9      ZINDEX=MWORK (|,+) (|>:<,|) ;
10     ZMIN= SUM(MWORK(|ZINDEX,|));
11     EST=TARGET(|I,1:12|) +
           DONOR(|ZINDEX,1:12|) # (TARGET(|I,1:12|) = .);
12     IF I=1 THEN IMP=
           TARGET(|I,|) || DONOR(|ZINDEX,|) || EST || ZMIN ;
13     ELSE IMP= IMP //
           ( TARGET(|I,|) || DONOR(|ZINDEX,|) || EST || ZMIN );
       END;
FINISH;
```

1. The IMPUTE subroutine presumes that the data have already been read into IML matrices. For example,

```

PROC IML;
USE OZONE.COMPLETE;
READ ALL VAR {O1 O2 O3 O4 O5 O6 O7 O8 O9 O10 O11 O12 SDATE}
INTO DONOR;
```

turns the permanent SAS data set OZONE.COMPLETE into the matrix DONOR in which the columns are the 12 hourly ozone measurements and the rows are the days. SDATE is the SAS date of the day (number of days from Jan 1, 1960).

TARGET contains the observations with missing values which are to be replaced by values imputed from DONOR, which should match TARGET in column structure. MTARGET and MDONOR are matrices containing the values used to score the similarities between TARGET and DONOR days, respectively. For direct imputation, MTARGET and MDONOR will both contain ozone values and will be identical to TARGET and DONOR, respectively. For indirect imputation, MTARGET and MDONOR will contain the covariate data (such as temperature) corresponding to TARGET and DONOR and will match those matrices in column structure, and the rows will correspond. MTARGET and MDONOR could be eliminated and the IMPUTE subroutine simplified if there were no need for indirect imputation. WEIGHTS is a vector of scaling weights to be applied in the computation of similarities between days.

2. and 3. Count the number of rows (days) in TARGET and DONOR data sets.
4. Row-by-row (day-by-day), each observation with missing values will be matched against the class of DONOR days.
5. and 6. Build a matrix having identical rows equal to the ozone (direct) or covariate (indirect) values of the current target day. This matrix is conformable to the DONOR matrix and facilitates all-in-one computation of similarities.
7. and 8. Return a matrix (conformable to DONOR) in which the elements of a row are the weighted differences between the ozone (or covariate) values of the target hours and the ozone (or covariate) values of the donor hours. This begins the process of scoring similarities. What remains is to sum the elements row-wise, to yield the set of donor-day similarity scores, and then find a minimum score. Note the use of elementwise multiplication by the matrix of Boolean conditions (MWORK $\rightarrow$ = .) to avoid propagation of missing values.

9. Perhaps the most compact -- and cryptic -- statement in the routine. Sums the weighted hourly similarities row-wise, finds and returns the row number of the row with smallest similarity score. This identifies the donor day best matching the target day.

10. Returns the best similarity score. (Not an essential part of NNHDI, but useful in diagnosing how well NNHDI did.)

11. Imputes the DONOR day's values for the missing hours in TARGET day.

12. and 13. Add the completed TARGET day at the bottom of the others in the IMP matrix returned by subroutine IMPUTE. Note that IMP will return not only the reconstructed day's values (in EST), but also the original data with missing values (in TARGET(|I,I)), and the complete donor day (in DONOR(|ZINDEX,I)), and the best similarity score (in ZMIN). If only the reconstructed data are desired, the horizontal concatenation in 12 and 13 could be eliminated: 12 IF I=1 THEN IMP=EST; 13 ELSE IMP=IMP//EST; and VARNAMES modified appropriately below.

To run the subroutine, a RUN statement can be included within PROC IML, as follows:

```
RUN IMPUTE  TARGET=mytarg1  MTARGET=mytarg2  DONOR=mydonor1
           MDONOR=mydonor2  WEIGHTS=mywghts  IMP=myimp;
```

Here, the matrices in lower-case will have been created previously from SAS data sets read into IML and are passed to the IMPUTE subroutine. Some may be identical. For example, for DIRECT TARGET datasets, mytarg1 = mytarg2 and mydonor1 = mydonor2. For INDIRECT TARGETs, these pairs will not be identical. SAS IML seems not to like duplicate argument names in RUN statements. The myimp (= IMP) matrix returned from IMPUTE can be turned into a SAS data set by a CREATE and APPEND statement:

```
VARNAMES={O1 O2 O3 O4 O5 O6 O7 O8 O9 O10 O11 O12 SDATE
          D1 D2 D3 D4 D5 D6 D7 D8 D9 D10 D11 D12 DSDATE
          I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 ZMIN};
CREATE OZONE.IMPUTED FROM myimp (|COLNAME=VARNAMES|);
APPEND FROM myimp;
```

O1-O12 are the original data from the TARGET data set; D1-D12 are the corresponding data from the most similar DONOR day, and DSDATE is the SAS date of that donor day; and I1-I12 are the reconstructed (imputed) data and are a combination of O1-O12 with I1-I12.

A somewhat more complicated version of this algorithm can be written to return several of the most similar donor days. This would implement multiple imputation [4].

REFERENCES

1. R. J. A. Little, D. B. Rubin, *Statistical Analysis with Missing Data*, Wiley, 1987.
2. SAS/IML User's Guide, Version 5 Edition. Cary , NC: SAS Institute, Inc., 1985.
3. T. W. Sager, M. W. Hemphill, A. D. Vaquias, "Statistical assumptions matter in data analysis for Texas ozone nonattainment sites," *Journal of the Air and Waste Management Association*. (1990) vol. 40, pp. 199-205.
4. Rubin, Donald B, *Multiple Imputation for Nonresponse in Surveys* , Wiley, 1987.