CALIBRATED IMPUTATION FOR THE DRUG ABUSE WARNING NETWORK (DAWN)

Andrea R. Piesse and James L. Green, Westat
Andrea R. Piesse, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Imputation, Calibration, Mathematical Programming

1. Introduction

This paper presents an imputation approach used for the Substance Abuse and Mental Health Services Administration's (SAMHSA) Drug Abuse Warning Network (DAWN). DAWN uses a stratified, single stage sample of hospitals to produce national and area-specific estimates of drug-related emergency department visits. In some DAWN areas, data from one or more sampled units are critical for producing the estimates for an area. These units are generally critical because they contain unusually large volumes of DAWN events or are otherwise atypical of other units in their area. Occasionally, all data from a critical unit becomes unavailable for a period (e.g., due to a change in hospital staff). We have developed a specialized imputation approach, outside of the standard survey imputation procedure, to deal with missing data for critical units. The approach imputes case records for the missing data period from case records for the unit in an adjacent period. Time series models are used to predict important drug-use estimates for the missing data period, and then mathematical programming is used to calibrate reported data from an adjacent period to these estimates. The end result is a complete dataset of case records.

2. DAWN Background

DAWN uses a stratified, single-stage cluster sample of hospitals with 24-hour emergency departments. The American Hospital Association's annual survey database is used as the source for the sampling frame and the sample is updated annually. A hospital is eligible[1] for DAWN when it:

1. Is a general medical/surgical unit;
2. Has a 24-hour emergency department;
3. Is located within the coterminous United States;
4. Is a non-federal institution; and
5. Is a short-term stay institution.

The sample is stratified by DAWN area (similar to a Metropolitan Statistical Area, or MSA), size of unit (annual emergency department visits), central city status (located within the central city of the MSA or not), and types of services offered (presence or absence of an outpatient unit or an alcohol/chemical dependency unit). Units not located within the 21 targeted DAWN areas are assigned to a final primary stratum referred to as the national panel. The national panel stratum is needed for producing national estimates.

[1] The DAWN data collection protocol and sample design changed in 2003 (Green et al., 2001). The description given here applies to periods for which this imputation approach was used; however the imputation approach still applies under the new DAWN.

Monthly data have been collected on all drug-related visits to sampled hospital emergency departments since 1988. A drug-related visit is called a DAWN case. Drug-specific estimates are made based on the particular drug(s) identified in the case charts. Estimates are usually counts of specific drugs, e.g., marijuana, alcohol and cocaine. Thus, estimates are generally at the level of individual drugs, while the case record data may involve combinations of drugs. Annual and semi-annual estimates are required for the coterminous United States, for the 21 specific DAWN areas, and for specific drugs. These estimates are used to monitor changes in drug abuse patterns across time.

3. Requirements

The imputed data must satisfy several requirements. First, it is important to produce good national estimates and estimates for the individual DAWN areas. The latter estimates may be at greater risk of bias due to unit nonresponse from a critical unit. Given that such a unit has been identified to be atypical of others in its stratum, there is a risk that handling missing data via the standard nonresponse adjustment might result in an unacceptable level of bias in survey estimates.

Second, the DAWN dataset must be suitable for many different analyses (not all of which can be anticipated ahead of time). Therefore, in order to support any form of special request or data analysis, imputed data are required at the individual case level (combinations of drugs), not at the estimate level (specific drugs). We impute complete case records with a procedure that aims to preserve relationships between the variables collected. In the rest of the paper a specific combination of drugs is referred to as a case type.

Finally, the imputation scheme is required to ensure that no previously unreported drugs or drug combinations are imputed.

4. The Approach Used

Our approach consisted of four main steps:

1. Identify critical units and develop time series models for predicting missing estimates;
2. Exploit empirical relationships between drug-specific estimates and case types;
3. Calculate calibration factors; and
4. Impute the missing data.

First, we identified those nonresponding sampled units (a) that might cause considerable underestimation or overestimation using DAWN's usual unit nonresponse adjustment method, and (b) for which we could develop reasonably precise time series models for drug-specific

estimates. For each such unit, we produced time series predictions at the estimates level for the missing data time periods. We then examined the various combinations of drugs from the reported data in each critical unit to determine a subset of case types that covered the bulk of the cases or contributed in an important way to the key drug-specific estimates. Taking advantage of this reduced set of case types, we used an imputation method similar to raking or calibration estimation to create records that satisfied the time series predictions. This approach used past DAWN cases for each critical unit, in proportions that allowed us to match the predicted estimates for the same unit.

4.1 Identifying Critical Units and Predicting Drug-Specific Estimates Using Time Series

We reviewed sampled units to identify those with data gaps of one or more months in the survey year to be analyzed. We classified such units as critical units if they typically contained unusually large volumes of DAWN events or were otherwise atypical of other units in their area. In these circumstances, the standard DAWN unit nonresponse adjustment (which uses sampling stratum as the default nonresponse adjustment cell) may have been problematic, with the possibility of severe underestimation or overestimation. Typically, data were missing for one or more months at the end of the survey year; however, for some hospitals the missing data periods were "book-ended" either side by reported data. The imputation approach employed was the same regardless of the temporal pattern of the missing data.

For each critical unit for which there was sufficient past reported data, we created separate monthly time series for a set of the most important drug-specific estimates. Typically, this set included most of the following estimates:

- All drugs;
- Alcohol;
- Cocaine;
- Heroin;
- Marijuana;
- Amphetamines; and
- Club drugs.

We then used the Time Series Forecasting System software in SAS to fit different models to each time series and we selected the best fitting model (in terms of R-square). The software tested up to 42 different models, including linear trend, exponential smoothing, seasonal terms, ARIMA, etc. The best models varied by critical unit, and within units, by type of drug estimate. Based on the goodness of fit measure, we identified a subset of critical units that had reasonably precise models for all of their major drug-specific estimates.

By fitting a separate time series model to each series we were able to predict drug-specific estimates for the missing data periods within each critical unit, making full use of all of the reported data available for that unit. The predicted monthly estimates from these time series models were used as control totals in the third step of the imputation procedure, described in Section 4.3.

Figure 1 shows an example of a model fitted to a monthly time series of the count of all drugs reported by one critical unit. The best fitted model is Winters method (additive) with an R-square of 0.8.

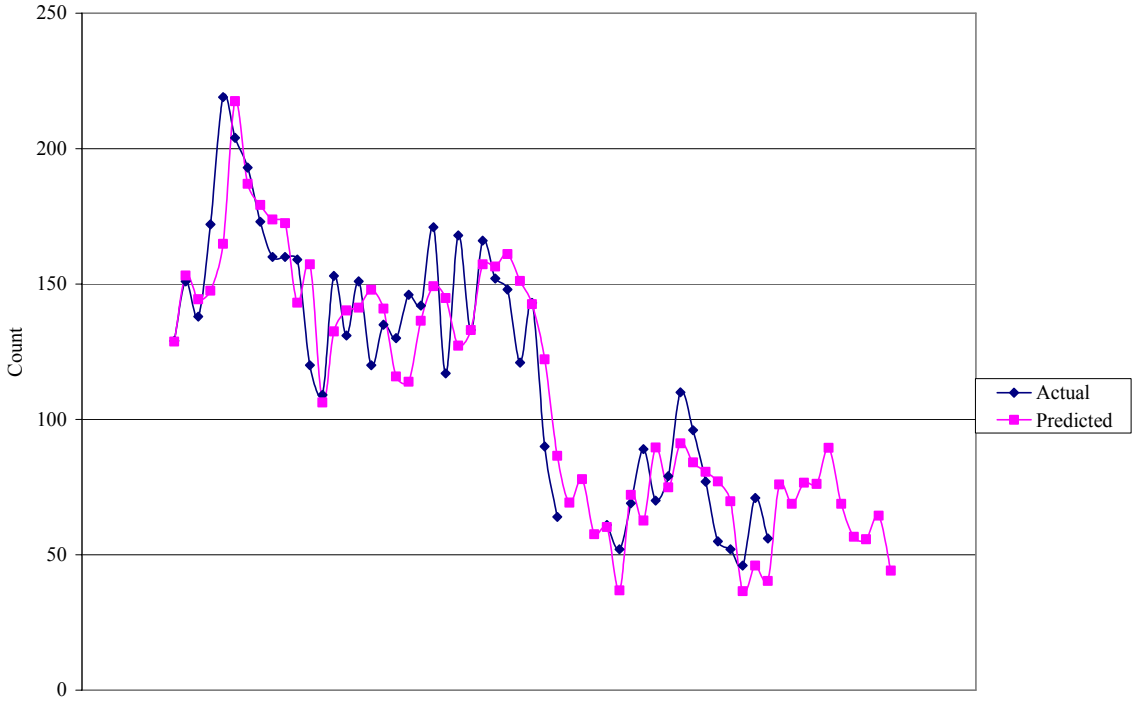

Figure 1. Example of a model fitted to a monthly time series of the count of all drugs reported by one unit

4.2 Exploiting Empirical Relationships Between Case Types and Drug-specific Estimates

In general, the relationship between case types and estimates is many-to-many. In other words, more than one case type can contribute to a specific drug estimate, and a specific case type can contribute to more than one estimate. However an examination of past data indicated that there is almost always a small number of case types (combinations of drugs) that account for (a) the bulk (e.g., 85%, 95%, etc.) of all drugs reported within a given unit and (b) those drug-specific estimates of most interest to SAMHSA. See Table 1 for a simple example of the kind of patterns detected. By taking advantage of this nature of the data with respect to the relationships between case types and individual drug estimates, we were able to reduce the dimension of the imputation problem without significantly limiting its usefulness. This is explained more fully in the next section.

Table 1. Illustrative distribution of case types for one unit

Case type	Count	Percent of cases
Cocaine only	500	63%
Heroin only	100	13
Cocaine and marijuana	75	9
Marijuana only	50	6
Cocaine and heroin	50	6
All others	25	3
Total	800	100

4.3 Calculating Calibration Factors

For each critical unit, the challenge was then to translate the predicted drug-specific estimates (corresponding to missing data time periods) to the case level at which imputed data were required. At this point we exploited the empirical relationships established in Section 4.2, whereby the bulk of reporting of major specific drugs is accounted for by a relatively small number of case types. This reduced set of case types was used in the formulation that follows, along with an "all others" case type that included any other combination of drugs not in the reduced set.

Consider the following representation of standard drug-specific estimates based on reported data:

Let x_j = the survey estimate of the total number of reports of drug type j in a given month

Then
$$x_j = \sum_{k=1}^{K} y_k I_{kj}$$

where

k = the case type index (e.g., cocaine and heroin),
y_k = the count of cases of type k,
I_{kj} = the number of times case type k contributes to estimate j.

When the estimate is of an individual drug, the quantity I_{kj} equals 0 or 1 for all case types except "all others". When the estimate is of the total count of all drugs reported, I_{kj} equals the number of drugs in the case type combination (except for case type "all others"). For the "all others" case type, I_{kj} is typically non-integer.

Taking advantage of the above formulation, we used each critical unit's reported DAWN cases for the previous year, and imputed cases for each current month with missing data based on a method similar to raking or calibration estimation. Specifically, calibration factors were calculated via a mathematical programming approach such that if these factors were used to adjust counts of case level data from the previous year, the difference between predicted drug-specific estimates and estimates resulting from the calibrated past data would be minimized. The optimization procedure was repeated separately for each critical unit and for each month of missing data. The method is illustrated in Tables 2 and 3, and can be expressed as follows:

Minimize:
$$D = \sum_{j=1}^{J} w_j \sqrt{(X_j - x_j)^2}$$

where

j = the drug-specific estimate index (e.g., cocaine),
J = the number of drug estimates for which control is desired,
w_j = a preference weight for estimate j,
X_j = the time series prediction for estimate j,
x_j = the calibrated estimate j

$$x_j = \sum_{k=1}^{K} y_k I_{kj} z_k ,$$

z_k = the calibration factor for case type k.

Subject to: $z_k > 0$

Table 2 presents a simple example of this approach, where instead of the unit's entire past year data, only data from the corresponding, reported period in the previous year were used. In this example, the drug estimates predicted by the time series models for the missing data period (see last row of the table) are exactly twice those of the corresponding time period in the previous year. For example, for the previous period, three case types contributed to the cocaine estimate, with counts of 500, 75 and 50, for a total of 625. The calibrated estimate for cocaine is 1250. The calibration factors indicate that the desired predicted estimates for the current

missing data period can be achieved by doubling the number of each case type from the previous period.

Table 3 presents a more complicated example. Here, the predicted drug estimates for the missing data period have increased for some drugs (marijuana, all others) and decreased for others (cocaine, heroin) relative to that of the corresponding reported period in the previous year. The calibration factors indicate that the desired predicted estimates for the current missing data period can be achieved by taking fractions of some case types and multiples of others (at the rates indicated in the last column of the table).

Table 2. Calibration example 1

	Past period case count	Estimate				Calibration factor
		Cocaine	Heroin	Marijuana	All others	
Case type						
Cocaine only	500	998.50	0.00	0.00	0.00	2.00
Heroin only	100	0.00	198.81	0.00	0.00	1.99
Cocaine & marijuana	75	149.31	0.00	149.31	0.00	1.99
Marijuana only	50	0.00	0.00	99.69	0.00	1.99
Cocaine & heroin	50	99.69	99.69	0.00	0.00	1.99
All others	25	2.50	1.50	1.00	45.00	2.00
Total	800					
Estimate						
Calibrated		1250.00	300.00	250.00	45.00	
Predicted		1250	300	250	45	

Table 3. Calibration example 2

	Past period case count	Estimate				Calibration factor
		Cocaine	Heroin	Marijuana	All others	
Case type						
Cocaine only	500	167.47	0.00	0.00	0.00	0.33
Heroin only	100	0.00	61.59	0.00	0.00	0.62
Cocaine & marijuana	75	193.24	0.00	193.24	0.00	2.58
Marijuana only	50	0.00	0.00	105.87	0.00	2.12
Cocaine & heroin	50	37.07	37.07	0.00	0.00	0.74
All others	25	2.22	1.33	0.89	40.00	1.78
Total	800					
Estimate						
Calibrated		400.00	100.00	300.00	40.00	
Predicted		400	100	300	40	

4.4 Imputing the Missing Case Data

The actual imputed case level data for each critical unit was created by sampling entire reported cases, by case type, from its previous year's data in accordance with the calibration factors determined to be optimal at the previous step. Referring back to Table 3, for example, imputed data would be created by sampling approximately one-third of the cocaine only case types from the previous period, while taking more than double the number of cocaine and marijuana, and marijuana only case types, and so on. We implemented the random elimination and duplication of past case records via systematic sampling. The past records were sorted by race, age, and gender within case type to provide additional implicit control over the distribution of demographic variables amongst the imputed cases. The end result was a complete dataset, satisfying the requirements set out in Section 3.

5. Summary and Relationship to Other Work

The imputation method described here is generally applicable to surveys where reported data may be available at other points in time for a given unit with periods of nonresponse (e.g., panel surveys). The available data can be

used both to predict estimates for the unit in question and to serve as the basis for case level imputation.

The method is especially useful when the relationship between cases and estimates is complex. In such situations, the data should be explored for empirical relationships that can be exploited in order to reduce the scope of the imputation problem without compromising its effectiveness. Additional requirements (e.g., minimizing deviation from historical distributions) can be expressed via the objective function and/or the constraints in the mathematical programming algorithm. Improvements to this method may include borrowing strength in the prediction step from those units that did respond during a given unit's missing data period.

Some related problems and methods are discussed in the literature. The weighting-like imputation method described by Deville (2000), where the total of auxiliary variables is known, has similarities to the imputation method we developed. Mantel, Singh and Yu (1995) describe an approach where calibration is used to adjust imputed second phase data items such that agreement between expansion and small area estimates is achieved.

6. References

Deville, J. (2000), Generalized Calibration and Application to Weighting for Nonresponse. *COMPSTAT—Proceedings in Computational Statistics, 14th Symposium*, pp. 65-76.

Green, J., Baskin, B., and Lee, K.C. (2001), Sample Redesign for the Drug Abuse Warning Network (DAWN). *ASA Proceedings of the Joint Statistical Meetings.*

Mantel, H.J., Singh, A.C., and Yu, M. (1995). Mass Imputation for Two Phase Sampling: Use of Small Area Estimation and Calibration Techniques. *Proceedings of the Survey Methods Section of the Statistical Society of Canada,* pp. 57-62.