POPULATION ESTIMATION FROM ADMINISTRATION RECORDS

C. D. Palit and P. R. Voss, University of Wisconsin
H. C. Krebs and B. D. Kale, State of Wisconsin

INTRODUCTION

We all have our individual definitions of administrative data. To us, the authors of this paper, administrative data are data which are routinely collected by some other agency for non-demographic purposes. Others have written more extensively about administrative data (see for example Lee and Goldsmith editors, Population Estimates: Methods for Small Area Analysis, Sage publications, 1982, Beverly Hills).

Perhaps the most bothersome feature of administrative data for demographers engaged in population estimation activities is our lack of control of quality. Particularly bothersome is the variation in quality over time. This variation is most apparent perhaps in the classification process wherein the data are classified back to geographic area of origin. Minor changes of emphasis in the collection process can have significant effects on the geographic classification results. Nevertheless, since the administrative data set is usually almost a "free good" to the demographer it provides a cheap if not always convenient access to information about population and its change. The challenge for the demographer is to design procedures which will enable him/her to extract the information hidden in the administrative data. Over the years demographers have devised intricate and clever schemes to do this: ratio correlation, censal ratio, component, composite, ratio difference, etc. Since, population estimation is a practical art with an external source of validation (i.e., the periodic census), much of the art has outstripped the theory. For demographers engaged in the estimation endeavor the ultimate proof of the pudding is how well the procedure works. Many person hours are spent testing and refining procedures for different situations. Demographers have long since learned that the properties of an administrative record data set are a function not only of the type of data collected, but also of the collection apparatus/agency, and the culture from which the data are collected. Thus the behavior of automobile registration counts in Wisconsin may well be different from the behavior of automobile counts in Illinois or Nevada.

Administrative data are only of interest for population estimation, if they are symptomatic of the presence or absence of population, i.e., data which yield statistics, usually counts, whose values move more or less systematically in response to changes in population. Examples are drivers licenses, tax returns, etc. Naturally some data sets are more weakly connected to population movements than others.

ERROR STRUCTURE

From our point of view the error structure of administrative data can be split into the following components:

1. natural variation stemming from the relationship of the administrative data set to the population;
2. allocation or classification error;
3. error accruing from changes over time in the relationship between the administrative data and the presence of population;
4. error due to the lag between the change of the real events in the population and their registration in the administrative record data set.

The amount of error and the nature of the source of error will impact on the performance of the various population estimation procedures. Our understanding of the error structure should and does influence the choice of estimation strategies. Before discussing this further we have some more to say about this classification scheme.

1. Natural Variation

For some data sets, e.g., drivers licenses (DL), the counts (of drivers licenses) are in fact counts of the status for members of the population. In the case of drivers licenses we can consider each person in the population to have one of two states; either they have a drivers license or they don't. If we code the "don't have a drivers license" state as a zero and the "has drivers license" state as a 1, then we can view the count of driver licenses in an area as a realization of a random variable with a binomial distribution, whose parameters are the true population size, N, and the probability of having a drivers license in that area, p. The expected value for the number of drivers license's is Np and the variance is Np (1-p).

If the variance was small and we knew p, then the relationship

$$\hat{N} = \frac{\text{Number of DL's in the area}}{p}$$

would be a reasonably accurate estimate of N. The variance of this estimate would be

$$N(1-p)/p$$

and the coefficient of variation for this estimate would be

$$\sqrt{\frac{1-p}{Np}}\ .$$

As p approached 1 the variance on the estimate would approach zero. As p approached zero the variance on the estimate would go to infinity. Thus for symptomatic data with small p (for example births and deaths) the natural variation induced in the estimate by this binomial process would tend to be the dominant error source. For other symptomatic data sets, (for example tax filers), allocation errors or changes in the relationship between the symptom and the population would more likely be the dominant source of error. The value of p can be extremely important for those estimation procedures which seek to estimate the change in the population directly. It would not, for example, be a good idea to use the change in the number of deaths to produce a direct estimate of the change in the population. On the other hand changes in Medicare data can be used quite reliably to directly estimate the change in the number of 65 plus persons, because for this dataset and population value of the parameter p is very close to one.

We have taken to calling p the "capture ratio". Not surprisingly we prefer to deal with symptomatic data which has a high capture ratio.

There are of course symptomatic data sets with more complex relationships to population. A good example of this is the dollar value of exemptions claimed on Wisconsin tax returns. We can consider this as a compound variable consisting of a binomial process coded "one," if a person files a return, or "zero," if a person does not file a return, and another process which is conditional on the person having filed a return in the first place. Both process contribute to the natural variation component of error with the binomial parameter still being considered as the "capture ratio". This more complex variable has the same behavior as the simpler variable considered earlier and, like it, will yield more reliable estimates if the capture ratio is large than if it is small. We do, however, have to consider the variation of the secondary component as well.

2. Allocation or Classification Error

Many estimation procedures use a base period, or periods for which census counts are available to perform a calibration of the relationship between the symptomatic data and population. Allocation error may become a serious problem when the administrative data collection process changes in a way as to substantially change the nature (pattern if you like) and probabilities of misallocation from what they were in the base or calibration period. Error due to misallocation is probably always present at some background level. The background level varies with the organization and purpose of the agency collecting the data. If either were to change after the base period the performance of the estimation procedure may be impaired -- sometimes with a bias, and sometimes with a decrease in the precision of estimation but with no malice.

3. Relationship Between Symptom and Population

Administrative records are related to people, but the relationship of the symptom to population can change over time for a variety of reasons: changes of the basic law that is the reason for gathering the data, administrative decisions changing policy about the collection or about the content of the file, compositional changes in the population which change the relationship between the reported event and population, and, finally, behavioral changes in the population resulting in a change in the relationship.

The sensitivity of the estimation procedure to changes in the relationship can become a serious issue if the data set is seen as being subject to such changes differentially.

4. Lag

Administrative records do not record a change in a person's status until this change is registered. The recorded values for symptomatic data will therefore often lag behind its true value. Persons moving from one area to another for example may take considerable time to change their address with the appropriate agency.

As far as the effect of lag is concerned, we know of little that we can do. If changes in the relationship between symptom and population are expected then it would be best to steer clear of ratio-correlation, and other procedures which rest on a strong assumption of constant relationship, and to use censal ratio or other procedures which do not rest as heavily on such an assumption.

FILTERS

For error sources discussed here, we believe that it may well be possible to construct filters which will take advantage of the auto-correlation present in an administrative data set across time to reduce the total variation of the observed symptom value around its true value.

Smoothing or filtering procedures of this type are not new. Censal ratio estimates using births and deaths have routinely used a three year average centered over the estimation year to reduce the variability of the birth or death count. As an illustration of the potential effect of the use of filters on the error distribution of the population estimate we have used 65+ deaths to estimate the 65+ population of 71 Wisconsin counties; and births to estimate the 18-44 female populations of 71 Wisconsin counties. Tables 1 and 2 show the error distribution for censal ratio estimates based on unfiltered data (i.e., single year), and for two estimates based on filtered data. The two filters used were a three year average centered over the year of estimation, and the predicted value of a first

order auto-regressive model fitted to data which included the year of estimation.

Similar results are observed if we apply an autoregressive filter to the dollar value of exemptions data for the municipalaties of Wisconsin. Table 3 shows the pattern of accuracy by Minor Civil Division (MCD) size of MCD estimates made using raw dollar value of exemptions data and filtered dollar value of exemption data. In all but one size category the estimate based on the filtered symptom out performs the estimate based on the raw data. The index of accuracy used is m. This index has been previously discussed by the authors but for completeness we will briefly review its properties. Algebraically,

$$m = 1 - \sqrt{\frac{\text{Sum of the squared errors in estimating changes}}{\text{Sum of the squared changes in population}}}$$

Some properties of m are:

1. m is negative if the sum of the squared error is greater than the sum of the squared changes in population.
2. m=0, if the sum of the squared error is equal to the sum of the squared population changes. Thus if the estimation rule is to use the base-line census then m is always zero.
3. m is between zero and one if the squared error is less than the sum of the squared change.
4. m=1 if the sum of the squared error is zero, i.e., we make no estimation errors.

In general m can be interpreted as the proportion of the total change in population captured by the estimation procedure and the higher m, the better the estimation procedure.

From these results we conclude that, indeed, we can, by the construction of appropriate filters improve estimates.. . . sometimes. But only sometimes. When we try the same filtering process on the number of tax return filers per MCD we find the opposite results. The estimates based on the filtered data show no improvement over the estimates based on the raw data. These results are shown in Table 4.

To explain these results we advance the following hypothesis. For simple symptomatic data sets with low capture ratio the underlying binomial process produces MCD totals which are realizations of random variables with relatively high variance. We have shown this algebraically above. Because this variance is relatively high the information contained in the historical series can be used together with the current data to construct an estimate of the expected value of the total which has a smaller variance. This would seem to be the case, for example, for estimates based on the number of deaths. However, when the value of the capture ratio is high, i.e., approaching one, the variance around a given MCD total is relatively small (again this is an algebraic property), and the amount of additional information which can be gleaned from the historical series is too small to produce any significant improvement in the estimate. Indeed there may be a small penalty to pay because the filtering process may remove a critical quantity of the information in the current data. This seems to be the case for the estimates based on the filer data, where the capture ratio is substantially higher than for the death data.

If this hypothesis is true then can we explain the improvement wrought in the exemption data by the filtering process? The answer is yes the explanation is as follows:

Even though the exemption data have the same capture ratio as the filer data, the structure is different. The exemption file is, as we have previously pointed out, a complex data set. This means that MCD estimates by this symptom have a process variance which arises from two sources; (1) the binomial part of the process, and (2) the other process which determines the size of the exemption claimed. As a consequence, the relative variance for the MCD's dollar value of exemptions is higher than for it's filers, and the improvement which can be obtained by incorporating information from the historical series of dollar value of exemption values is more than the penalty incurred by using the filtering process.

SUMMARY AND CONCLUSION

In this paper we have proposed a way of looking at administrative data which are symptomatic of population which we believe can provide a useful framework for deciding how to use the data for making population estimates. We believe that it can and will be a fruitful path to follow.

TABLE 1.

Frequency Distribution Error Measures for 71 Wisconsin County Populations, 65 Years or Older, by Method, Using 65+ Deaths

Percent Error	Method		
	Single Year	3 - Year Average	First Order Auto-Regressive
- .30		1	0
- .25	2	0	0
- .20	1	1	1
- .15	1	0	3
- .10	7	2	2
- .05	14	20	19
.00	19	18	23
.05	11	19	17
.10	10	6	4
.15	3	4	2
.20	2	0	0
.25	1	0	0
Average Error	348	272	259
Average Percent Error	7.13	5.30	4.74
Mean Square Error	369,116	299,123	202,814

TABLE 2.

Frequency Distribution Error Measures for 71 Wisconsin County Populations, Females Age 18 - 44, by Method, Using Births

Percent Error	Method		
	Single Year	3 - Year Average	First Order Auto-Regressive
- .25			
- .20			
- .15	5	3	2
- .10	7	9	10
- .05	11	7	9
.00	6	9	12
.05	15	15	13
.10	6	14	13
.15	11	6	6
.20	4	4	3
.25	3	3	0
.30	2	1	1
.35	0	0	0
.40	1	0	2
Average Error	901	861	750
Average Percent Error	10.74	9.46	9.77
Mean Square Error	3,598,466	3,902,002	2,509,254

TABLE 3.

M Values by Size of MCD for Estimates of Wisconsin MCD's Based on Dollar Value of Exemptions, 1980

Population Size	Estimate from :	
	Raw Dollar Value of Exemptions	Filtered Dollar Value of Exemptions
Less than 250	.323	.351
250 to 499	.413	.402
500 to 749	.458	.587
750 to 999	.580	.590
1,000 to 1,499	.284	.656
1,500 to 2,499	.420	.586
2,500 to 4,999	.552	.587
5,000 to 9,999	.455	.470
10,000 to 19,999	.546	.527
20,000 to 49,999	.250	.332
50,000 or more	.749	.775
TOTAL	.690	.723

TABLE 4.

M Values by Size of MCD for Estimates of Wisconsin MCD's Based on Number of Tax Filers, 1980

Population Size	Estimate from :	
	Raw Number of Filers	Filtered Number of Filers
Less than 250	.318	.300
250 to 499	.405	.411
500 to 749	.477	.473
750 to 999	.480	.475
1,000 to 1,499	.541	.527
1,500 to 2,499	.609	.587
2,500 to 4,999	.570	.568
5,000 to 9,999	.494	.478
10,000 to 19,999	.333	.274
20,000 to 49,999	.246	.206
50,000 or more	.394	.411
TOTAL	.394	.405