DISCUSSION

Paul Burke, U.S. Department of Housing and Urban Development

The moral of this session is that administrative data are not free. The papers show clearly some of the costs and problems in using administrative data. This is not to say administrative data are useless. Quite the opposite, they are usually much cheaper and frequently more accurate than survey data. But costs remain, and these costs must be faced and budgeted for.

The session is split almost equally between papers which describe data bases and the work that goes into preparing them, and papers which describe applications. I will discuss the papers in order, which means starting with three papers that describe data bases.

I. DATA BASES

Cys-Hinkins-Rehula. This paper shows one extreme in the cost of using administrative data. Normal IRS processing of these returns is largely manual, and does not require much computerized or even standardized data. Therefore, the statistical office at IRS must standardize and computerize the data in a complicated process described in the paper. All of the normal costs of running a survey apply to these data except the actual data collection. The costs are clearly worthwhile in this case, because the data are irreplaceable. Many companies would never give this information in a voluntary survey.

Hirschberg-Phillips. This paper shows other costs of checking and exploring administrative data. The researchers have tested the data base to find which variables are present for which companies, how consistent variables are over time, and other issues that affect reliability. One would think these tests would have been needed by the original creators of the data, in doing their credit reports, but they were not done, and since the statistical tail cannot wag the administrative dog, the statistical tail has to pay for these checks itself. The data base clearly is very valuable for many uses. For example, we have several times used it at HUD, and I encourage the Small Business Administration to continue their work and to circulate their findings widely.

Sailer, et al. This paper provides the kind of documentation that the authors of the previous paper would have loved to have. It shows internal procedures at IRS and how these affect the data. The paper also continues the dialogue between IRS and users of the data over what users need. There are serious issues involved in the timing and accuracy of preliminary and revised estimates, handling of returns filed late (even years late) and timeliness of data. In this context, the paper indicates what appears to me to be somewhat too much emphasis on publications, and too little on computer tapes, which serious policy-makers are likely to use, or on clean, steadily lengthening, longitudinal files, which serious researchers are likely to use.

II. APPLICATIONS

Irwin - Herriot. This is the first paper in the session that describes an application of administrative data. There is a better fit than I would have expected between joint returns and married couple households. However, the present paper shows very simple comparisons, and much more study is needed, especially to estimate unmarried individuals. The matching should take into account income, sex, age, and perhaps race (sex, age, and race are available from Social Security files for each income tax return, as described by Word and Zitter in another paper given at these meetings [1].

Further study may show that some household types, particularly in some income levels, cannot be accurately estimated from income tax data, but estimates of other household types will still be very useful for many government and private purposes. At HUD, we have to estimate demand for different sizes of housing units, often by county, and any information on even a few household types would be better than none. The same need applies in many other industries. I encourage the authors to pursue this methodology (Mr. Irwin should also pursue his excellent fiddle-playing).

Smith-Diamond-Orcutt. This paper is another application, involving several data bases. The paper has a rich blending of data to determine how much of wealth is inherited, how much is built up gradually in one's life, and how much comes from chance.

The paper has some weaknesses in merging kinship data and wealth data, since nothing is known about their correlation. The wealth data themselves have weaknesses, being based on a survey where a quarter of all wealth went unreported, and on an assumption that this wealth was distributed like the rest. The kinship data are an area where actual "administrative" data, such as from the Mormon Church or other genealogical experts, could have been better than the computer simulation used.

With these data merged as a starting point, the authors, Smith, Diamond and Orcutt, simulate aging, births, marriages, accumulation of wealth, and inheritances. For example, marriages are simulated by controlling for education, age and race. The authors do not control for wealth, however. This seems important; wealthy people generally marry each other. If wealthy people randomly married less wealthy people, one would expect a rapid diffusion of wealth, which does not happen in the U.S. The authors do find a rapid diffusion in their model, and compensate for it by assuming that random events (oil wells, lotteries, etc.) re-concentrate wealth. The simulation is interesting but seems to have too many weaknesses to use for policy purposes.

Petska. This paper gives an appropriate end to the session, including both a

description of data sources, and some applications of the data. This paper describes how the internal procedures and definitions at IRS determine what data are filed by nonprofit organizations. The paper also reports some of these data to show the variety of the nonprofit sector and the data available. It explains how this sector responds to changes in the law and in overall taxation. One of the most intriguing findings is that large foundations give relatively less of their wealth away than small foundations. Perhaps that is how the large foundations stay large?

III. CONCLUSIONS

Overall the session shows that administrative data are essentially just like survey data. They must be processed the same way: cleaned, documented, and I suppose they offer the same scope for errors if misused. Their advantage is that administrative data can be much cheaper than survey data (at least for the statistical unit).

REFERENCE

[1] Word, David L. and Zitter, Meyer. "Further Developments in Intercensal Population Estimates Using Administrative Records." 1982 American Statistical Association Proceedings, Social Statistics Section.