PROBLEMS IN CONSTRUCTING TIME SERIES FILES FROM MINCOME DATA:

TECHNICAL REPORT

by

Andy B. Anderson

James D. Wright

Eleanor Weber-Burdin

New England Research Associates

Amherst, Massachusetts 01002

Introduction

The project whose results are reported in this paper was designed to assess and evaluate the kinds of problems a "typical" secondary user of MINCOME data would face in attempting to construct and analyze a time series data file from the MINCOME archive. One of the primary strengths of the MINCOME data, relative to most other social and economic data sets, is that MINCOME data were gathered regularly over a three year period, with each participating family interviewed repeatedly through time. Consequently, the data permit detailed and sophisticated analyses of the social and economic behavior of sample households over time, assuming, of course, that the appropriate time-series data can be extracted in useable form. This project considered the various problems MINCOME users will encounter in constructing such a time series.

In general, our assessment has focussed on two issues: data quality and data useability. The first concerns the question whether the data are sufficiently consistent across waves that time series analysis is warranted. For the most part, the answer to this question is yes. Given that, the second issue concerns the ease with which appropriate time series files can be extracted, and more particularly, the kinds of snags and problems a user will face when attempting a time series extract.

It can be noted initially that MINCOME employed automated data quality checks during the data entry process. (These quality controls are described in detail in one of the Technical Reports on the data base.) These checks assessed data quality <u>within</u> a single interview (or "periodic"), but did <u>not</u> necessarily check for inconsistencies or other data quality problems from one periodic

to the next. Within a single interview, to illustrate, a respondent may give implausible, inconsistent, impossible, or otherwise problematic responses. Most of these problems would have been detected by the error checking routines employed by the on-line data entry system. In like fashion, responses to questions in one periodic may well be inconsistent with responses given in other periodics; and these kinds of "across-periodics" problems have not been extensively examined to date. For a given respondent, two temporally adjacent interviews may each seem perfectly reasonable considered apart from one another, and yet may reveal serious data quality problems when considered as a sequence: for example, jobs may change unaccountably, weeks may be skipped in the reporting of hours worked, the same week may be reported twice, with the second report inconsistent with the first, and so on. Although some measures were taken by MINCOME to reduce these kinds of across-periodics problems, no systematic effort has been made, prior to this project, to evaluate their frequency and character. Such inconsistencies across time, of course, may (at least in principle) greatly imperil time series data analysis.

To be sure, not all of the MINCOME data require analysis within this framework. Many of the survey questions were only asked once or twice; the across-periodics consistency problem only arises in those sections of the questionnaires that were administered repeatedly throughout the run of the experiment. The "continuous" (i.e., repeatedly measured) variables contained within the data base are of two general sorts:

(1) The economic core variables, which contain the basic labor supply data. Variables such as wage rate, hours worked, job, employment status, and related issues were gathered more or less continuously to allow for time series analysis of the labor force response of families to the NIT program.

(2) The family composition variables (age, sex, and related variables describing each member of the sample household). Family composition was also measured more or less continuously in that any change in family composition from one periodic to the next was recorded. As a result, the composition of any given family at any given time can (theoretically) be reconstructed from the interview data. This is important for various reasons. For one, the size of the family's treatment payment (entitlement) is based at least in part on family composition (size). Even though the Income Report Form (IRF) data independently measured family composition, the survey data also make it possible to assess family composition through time. Secondly, family composition is important because it heavily influences both earnings capacity and expenditures of the household, making it an important source of information in any evaluation of the experiences of families in a welfare experiment. Finally, changes in family composition are one possible outcome of the experiment and thus represent a potential dependent variable in a time series analysis.

All things considered, then, our judgment is that the economic core data and family composition data are likely to be of direct interest to most or all secondary users; the following report is therefore focussed on these two segments of the MINCOME data base.

Consistency over time in the economic and family composition variables is a data quality issue; a more general issue addressed in this report is the useability of the data. This part of our assessment is less concerned with actual errors or inconsistencies in the data than with the form in which the data are stored, the extant documentation, and various idiosyncracies in the data that might increase the time and effort required from a user to get the data to an analyzable stage.

In some cases, as we show below, problems of this latter sort may indeed produce errors for the unwary user; in most cases, they simply increase the amount of data manipulation (and ensuing frustration) necessary to produce an analyzable file. Most of the problems of this sort that we discuss are the result of decisions made during the life of the experiment: e.g., to change the interview schedule or the wordings of specific questions, to alter the data coding and entry conventions, to change the data formats, to change the data processing routines used to construct the files, etc. Because of these kinds of changes (and the enormous size of the data base), the data structure is rather unwieldy at best and would certainly discourage all but the most insistent potential user.

Some of the "useability" issues that we discuss have no doubt already been addressed by ISER staff, and satisfactory solutions to them uncovered. Others may be considered as targets for future work at the Institute; still others may be deemed beyond the scope of ISER's work. In all three cases, potential users should at least be made aware of the kinds of difficulties that exist, whether ISER has an "official" solution for them, or not. In any

case, we do not propose solutions in this report; rather, we simply note the various problems and their apparent consequences. Whether the problems we note are best dealt with at the level of the existing archive or in the special documentation that accompanies a data extract tape is another issue that we do not address.

Constraints and Qualifications

In making our assessments, we have tried to assume the position of a so-called "typical user" and to detect the kinds of problems that this typical user will face. This is an unrealistic position for us to assume, in two senses: First, we are far more informed about MINCOME and its data base than almost any user of the data can be expected to be, excepting ISER staff and users working very closely with ISER staff. Moreover, our computer and data processing support is "state of the art." We are, in short, more experienced in working with these kinds of data, and more technically sophisticated, than will be the case for a typical user. In general, it can be fairly assumed that if we encountered a problem on a particular issue, most typical users will too.

It occurs to us that many of the problems we have encountered could perhaps be resolved rather easily, by reference to the proper memorandum from the MINCOME archive, by a telephone call to the right person in Winnipeg, or by digging through tertiary documentation. Even in these cases, where we think the problem is potentially resolvable, we have not gone to great lengths to resolve it. Our feeling is that the public use version of the MINCOME data, and associated documentation, have to stand on their

own; problems whose resolution requires more information than would be available to the normal user cannot help but discourage exploitation of the data base. Further, some of the problems we have come across almost certainly reflect a lack of information, or some misunderstanding, on our part. Again, however, the typical user cannot be expected to understand the experiment and the data base any better than we do; so here too we have simply recorded the problems we encountered, on the assumption that we are not likely to be the only ones who will have difficulty.

Many of the problems we discuss are clearly problems in the available MINCOME documentation, not in the data themselves. The documentation that we received for the data files consisted of (i) the MINCOME data dictionary, and (ii) the SAS control card set-up for each file. Neither of these is particularly informative. We assume that there will eventually be special time series documentation that would accompany time series extracts.

A note of caution must also be introduced with regard to the sample data obtained for this project. We initially agreed to examine a random subsample of the data base, consisting of 250 randomly selected families enrolled and participating in the experiment. The sample of 250 we finally received contained some surprises. First, 57 of the 250 sample families (23%) apparently never participated in the experiment; the data tape contains baseline data for these 57 families but contains no other data from any of the other periodics. Secondly, the sample contains only 15 families from the supplementary sample, many fewer than would be expected in a simple random sample of cases from the master data base. The first of these two "surprises" reduces by about a quarter our effective sample size and consequently

subjects our estimates of parameters such as error rates to greater uncertainty. The second "surprise" severely limits our ability to say much about problems in the supplementary sample. To be sure, data for the supplementary sample were expected to be relatively clean; and in fact, what little data we have from the supplementary sample is consistent with that expectation. Still, with only 15 families to examine, we must be quite circumspect in what we say about the supplementary sample data.

(On the more positive side, it should be kept in mind that the periodic interviews for the main sample from periodics five through nine were identical to the interviews for the supplementary sample. Therefore, unless there are errors generated by the nature of the respondents in the supplementary sample, or by the way in which their data were handled, the error patterns and rates should not be greatly different from those of the main sample during the late stages of the project.)

One final surprise was that the data arrived in a form rather different from what was expected. In our data file, the surveys are simply numbered 1 through 14. Every family in our sample is thus represented by a record on 14 different data files, with dummy codes used to fill data fields for any time prior or subsequent to their participation. Data records for families from the supplementary sample are thus "backfilled" with dummy codes for the first five surveys; they first appear at Survey 6 along with the main sample members. Aside from this pattern, the respondents from the supplementary sample are not otherwise

identified in the data file. Then, starting at Survey 11, only the supplementary sample has data and the data records for the main sample are again padded with dummy codes. Since, as we have already noted, there were only 15 supplementary sample members in our data, we have not used any of the data beyond Survey 11 in this report.

Family Composition

1. Importance of Data. The family composition time series is important for a number of reasons, three of which have already been mentioned. First, family composition affects the entitlement under the experiment because the guarantee level is dependent on family size. Although transfer payment calculations were based on the IRF data, and not on the survey data examined here, family composition information has an obvious relevance in analyzing responses, labor force or otherwise, to the experimental variables. Secondly, it is now well-known that a substantial amount of the movement into and out of the poverty category (at least in the United States) is accounted for by changes in family composition. Earners enter and leave the household and thereby influence the household's earnings potential. Likewise, consumers enter and leave the household and thereby affect the pattern of expenditures. Any examination of data from an experiment with alternative forms of welfare must be sensitive to family composition changes. Thirdly, change in family composition is one possible consequence of an NIT experiment, as has been much-discussed in the various US experiments. The most policy relevant responses to guaranteed income are not necessarily those related directly to labor supply.

At a more technical level, the family composition data provide information needed to trace the people who participated in the experiment. Elsewhere, we have discussed this as the "unit problem." People move into and out of households: they marry, get divorced, remarry and start new families; birth adds household members, while death subtracts them; and so on. As a result of these kinds of changes, any analysis of individual level behavior in MINCOME must be able to locate the individual in the correct family at every point in the experiment; and any analysis of families must be able to state the exact composition of the family at each and every point.

For these and other related reasons, the family composition time series from the MINCOME data base will almost certainly be critical to nearly every secondary user, and it is therefore important to know how well these data meet quality standards and how easily they can be accessed to construct the requisite time series.

2. Limitations on Family Composition Edits. Over and beyond the constraints mentioned earlier, two additional matters must be taken into account in evaluating our findings about the family composition time series. First, since we have worked only with a subsample of the master data file, we are generally unable to trace moves between families. For each of the 250 families in our sample, that is, we know whether or not there were moves-in or moves-out, or both; but we do not know if the moves-in came from other families in the experiment, nor if the moves-out went on to other participating families. We doubt whether the frequency of

moves of this sort is very extensive, but any movement of household members among the experimental families would pose a serious complication in analysis. (For example, the data for Individual X might be appended to the file for Family Y at one time, and to the file for Family Z at some later time.) It would clearly be useful for ISER to estimate the number of moves of this sort that occur in the entire data base, if only to confirm that the number is few; but with the sample data at our disposal, this is a matter on which we cannot draw conclusions.

Secondly, our family composition analysis is based on data from MODULE 2 only. We think it likely that most users interested in family composition variables will also use MODULE 2 data. However, for reasons discussed below, the MODULE 2 data may produce real or apparent problems in some cases that could conceivably be resolved by information that is in the archive (in other MODULES) but that we did not obtain. However, we again feel that most typical users would also have access only to partial information and, therefore, that the potential "resolvability" of some discrepancy by reference to other information (that most users will not have bothered to extract) is rather an academic point. Unless a very substantial portion of the total data base has been extracted, in other words, most secondary studies using family composition data will encounter problems very similar to the ones discussed here.

3. Procedure. There are two sources of information about potential changes in family composition in MODULE 2. First, at each survey, each family has up to 15 data fields to record family member identifications. There are also 15 fields for a relationship-to-head code; and 15 fields to describe the sex of each family member. Although somewhat cumbersome, the family members lists represented in these fields can be compared survey to survey to see if any changes have occurred. Secondly, in addition to these lists, there are questions in every periodic asking if anyone has moved into or out of the household since the date of the last interview (DLI). If any of these questions are answered "yes," then the survey-to-survey list comparisons should confirm the change: a "yes" to the move-in question should be accompanied by the appearance of the new household member on later household listings; and likewise, a "yes" to the move-out question should be accompanied by the deletion of that member on subsequent household lists. (To be sure, more than one person and more than one change may be involved in the reporting in any given periodic.)

In practice, these comparisons are not nearly as straightforward as they appear to be at first glance. At a particular interview (call it N), a respondent is confronted with a list of family members present at the DLI (call it N-1) and asked if there have been any changes in the interim. In the data file, though, the list that would verify the change information is not present until the (N+1) interview. This is because the change that the respondent has reported for the period between (N-1) and

N does not find its way onto the basic household list until the (N+1) interview, at which time (of course!) questions are then being asked about changes occuring in the period between N and (N+1). In other words, changes in the household list detected at a given periodic are not recorded on the list until the subsequent periodic. The comparison between list information and the response to direct questions about changes in family composition must therefore be "out of phase" by one survey.

Once the structure and sequence of the lists and their updating are understood, the comparisons needed to verify the family composition information are not particularly complicated. For any given family, the temporally abutting lists are compared and each discrepancy is noted. Moreover, for each period bracketed by the two temporally adjacent interviews, responses to the moves-in and moves-out questions are examined. Errors are represented by any of the following patterns: (i) a move-in (or moves-in) is reported but no new member (or members) shows up in the appropriate family list; (ii) a move-out (or moves-out) is reported but no member (or members) disappears from the appropriate family list; (iii) a comparison of the two adjacent member ID lists shows one or more moves-out but the appropriate move-out question was answered negatively; or (iv) a comparison of the two adjacent lists shows one or more moves-in but the appropriate move-in question was answered negatively. We have tabulated the errors by type, and we have noted the number of changes involved that were consistently recorded, so that the relative frequency of each error type can be estimated. The

computer program that performs these checks works within a single family as it is followed from interview to interview, over the 11 total interviews.

4. Family Composition Results. To set the context, let us first consider the overall amount of change in family composition over the course of the experiment. For the sample, we examined the family composition data from the first 11 files as received from ISER. Table 1 shows the number of families present at each interview (in our sample), the number of individuals in those families (approximately, as these data are subject to possible errors), the number of families for whom no compositional changes have been detected (refered to hereafter as "stable families"), and the number of families for whom one or more changes in family composition were detected for that period ("unstable families"). The last column reflects a "change rate" and is simply the number of unstable families divided by the total number of families.

[TABLE 1 HERE]

As can be seen, the proportion of families in our sample experiencing some compositional change from one periodic to the next varies across interviews from about 5.5% to about 13.5% and averages 9.3% over the ten entries. There is no apparent secular trend over time in this percentage; rather, it appears to fluctuate more or less randomly from one interview to the next. As shown in Table 2, the changes are accounted for disproportionately by a small number of families experiencing a large number of changes, as would be expected.

[TABLE 2 HERE]

Table 1

Changes in Family Composition

Survey Number	N of Families in Sample	N of Individuals	Stable Families	Unstable Families	Per Cent Unstable
1	235	869	*	*	*
2	177	672	153	24	13.6
3	156	586	137	19	12.2
4	143	543	131	12	8.4
5	137	516	128	9	6.6
6	143**	538	135	8	5.6
7	137	496	127	10	7.3
8	134	491	120	14	10.4
9	129	474	116	13	10.1
10	127	459	112	15	11.8
11	125	455	116	9***	7.2

*No prior data for comparison.

**Data for Surveys 6 and thereafter contain the 15 supplementary sample families.

***This column sums to 133; thus, there are 133 changes in composition overall in this sample of 250 families.

NOTE: There are a few families in the "stables" column whose lists match correctly but who answered the change questions incorrectly. These are discussed in the section on errors.

Table 2

Distribution of Changes in Composition Across Changing Families

Number of Compositional Changes Recorded for the Family:	Per Cent of Total Changes Having Each Specified Number of Changes:
1	28%
2	26%
3	15%
4	10%
5	16%
6	5%

Table 2 shows the per cent of all changes accounted for by families experiencing one change, two changes, three changes, etc., across the eleven interviews. Just under half of all changes (46%) were accounted for by the few families experiencing three or more changes during the course of the experiment. Frequently, it appears, these are families with children that move in and out of the parental household with some regularity.

Clearly, the compositional changes recorded in Tables 1 and 2 do not necessarily represent errors in the family composition time series. Most of these are true changes that are recorded quite consistently and accurately in the data. However, compositional changes do represent the opportunity for error; almost all of the errors we have detected in the family composition time series are associated with compositional changes. The above information on frequency of changes is presented in order to provide some preliminary idea of the size of the "population at risk" for error.

5. Error Report. Table 3 tabulates the errors detected when checking the family member lists against the moves-in and moves-out questions, as discussed above. Families showing no moves-in or moves-out in their basic list and responding "no" to both relevant questions are omitted from the table (the consistent stable families). In the table, Column I indicates the survey period being examined; Columns II and III show the consistent "moves-in" and "moves-out" respectively (where the comparison between household list and the direct questions matches); Columns IV through VII show the numbers of households where some

inconsistency was detected; and finally, Columns VIII and IX show error rates, first with the total number of families as the base, and secondly with the changing families only as the base.

[TABLE 3 HERE]

(The numbers reported in Table 3 do not necessarily correspond to the numbers shown in Tables 1 and 2, mainly because a single error will often be double-counted in our procedures. For example, a move-in recorded in the wrong interview generates two errors rather than one: a missed move-in in the initial interview and an unaccounted-for move-in in the next interview. There are other, relatively minor, reasons for discrepancies between this and earlier tables, none of them sufficiently important to warrant a discussion of the details.)

The error tabulation given in Table 3 suggests that there are definite problems in the family composition time series that will cause difficulties for the typical user attempting to format a data extract or to interpret data contained in such an extract. The calculated error rates are fairly miniscule when expressed as the number of errors detected over total families present in the relevant time period; these rates vary from 0.7% to 3.9%, depending on periodic. However, if attention is focussed on the right-most column of the table, where error rates are expressed as errors detected over total families "at risk" (those experiencing some compositional change), one notes that the rates are surprisingly high. Indeed, in some periodics, errors were detected in over half the families showing some compositional shift. Our estimates of error rates, of course, are not overly reliable simply because of the small numbers involved, but it is clear from the table that some surveys have especially high rates

Table 3

Number of Errors Detected in the Family Composition Data

I	II	III	IV	V	VI	VII	VIII	IX
1-2	8	13	2	0	0	0	.011	.083
2-3	8	9	0	1	0	0	.006	.053
3-4	5	6	1	0	1	1	.021	.250
4-5	2	3	1	4	0	0	.036	.556
5-6	3	5	1	0	0	0	.007	.125
6-7	1	6	1	3	0	1	.036	.500
7-8	7	5	1	1	0	0	.015	.143
8-9	5	7	1	1	0	1	.016	.154
9-10	3	9	1	2	1	1	.039	.333
10-11	3	5	1	0	0	1	.016	.222

I = period over which error is assessed. Note that the 1-2 list comparison actually refers to a change in the interval prior to 1; the 6-7 comparison refers to a change in the 5-6 interval; etc.

II = consistent "moves in," the move-in appearing in both the list and the direct questions.

III = consistent "moves out," as in II.

IV = list indicates a move-in; respondents says "no" to direct question.

V = list indicates a move-out; respondent says "no" to direct question.

(Table notes continue, next page)

Table 3

(continued)

VI = list shows no changes; respondent indicates a move-in on the direct question.

VII = list shows no changes; respondent indicates a move-out on the direct question.

VIII = errors detected (sum of IV, V, VI, and VII) divided by total families present for the interval (from Table 1).

IX = errors detected (as above) divided by the number of "unstable families" (taken from Table 1).

of error for families experiencing a change in composition. In the 6-7 period, for example, there are only ten families in the sample with a change in composition, but half of those have a discrepancy of some sort. Error in the 4-5 period is even higher. For most of these errors, 22 of 28 to be precise, the list comparison indicates a move (either in or out) but the move questions were answered negatively.

The actual error rate in these data could be slightly higher even than the rates reported in Table 3. We have omitted from the table a number of cases (15) in which a move-in or move-out was reported in a previous survey, subsequent to which the family dropped out of the experiment. (In these cases, of course, we have no later family composition list to compare to the earlier list.) Inclusion of these cases would presumably increase the overall error rate slightly.

In general, the error rates detected in our subsample are higher than expected. We suspect that many of them could be reconciled using the data in the full ISER archive. And for some of the errors, we have enough information to make a reasonable conjecture about the series of events leading up to the real or apparent discrepancy. Too, some of these errors have been double-counted, for reasons already described. These provisos notwithstanding, the error rates still seem unexpectedly high. If a secondary user of the data were to take a reasonable shortcut and assume "no change" in family composition on the basis of negative responses to the direct "change" questions, family composition would be seriously mismeasured for a non-trivial portion of the sample and for a quite substantial portion of the families

experiencing some compositional change. In general, we think that for most of the records in error, the list comparison is likely to be the more valid indicator of real change than responses to the direct questions. That, however, is only a speculation that might be considered in additional details by ISER staff. Even if it turns out that many of our detected errors are fully resolvable with the information and experience available to the ISER staff, the fact remains that the problems we have encountered in working with the subsample will also be encountered by other secondary users. Therefore, some action is definitely warranted, even if it is only to be a general warning about the nature and extent of the problem that can be distributed along with any data extract.

6. Useability of the Family Composition Data. In addition to the potential errors discussed above, there are other problems that are likely to make it more difficult than necessary for secondary users to construct a family composition time series from the data base. We think, at the very least, that appropriate documentation should alert users to these potential problems. Some of them prove to be as maddening as they are trivially easy to correct.

To illustrate, in Surveys 1 through 4, and again in Survey 11, the convention is to indicate the end of the list of family member identification numbers by entering a "-2" in the field following the last family member ID. If followed throughout all periodics, this would provide a simple and unambiguous test for the end point of the ID list. However, this convention is not followed

in Surveys 5 through 10. Similarly inexplicable quirks in coding conventions occur throughout the data, appear not to be documented in the data dictionary or anywhere else, and constitute a singular frustration in attempting to analyze the results.

Some other examples: Household member IDs are not always coded in the same order across surveys, even when there has been no change in family composition. In one periodic, for example, the ID for the male head will appear first in the field, and in a later periodic, someone else's ID will appear first. Or, somewhat more commonly, IDs for the various children present in the household will be mixed around. With enough work, it is always possible to figure out who's who, but it should not require such efforts simply to identify the members of a stable family over time.

For most surveys (but not all), a family comprised of N members will have N ID numbers entered in the first N positions in the field, with positions N+1 through 15 padded with "-4" codes. Surveys 1 through 3 and 11, however, use a different convention (denoting the end of the family ID field with a single "-2"; thus, in these surveys, the amount of family ID information is variable.) Survey 4 is different from all the others. In Survey 4, for some mysterious reason, the ID number of the first household member is recorded in position 15. Thus, for a family with N members, the ID number of the second member appears in position 1, the ID number of the ith member is in position i - 1, and the ID number of the first member is in position 15.

Compounding the confusion, relationship-to-head and sex of the ith member of the household are coded in the correct ith position in subsequent fields.

Variations in coding conventions of the sort just illustrated can easily lead to errors on the part of a secondary user not fully familiar with the nature of the data and the evolution of the data base. Even a careful and well-informed user will encounter the same kinds of problems we did, if not warned in advance about the coding conventions followed. Most of what we have learned and reported about these problems results from "eyeballing" the data, going through a formatted hard copy version one case at a time. Nothing in the documentation we received alerts the user to these issues. As such, we spent an inordinate amount of time just deciphering the data file structure, discovering the coding conventions, figuring out which variables appeared in which columns and fields. Later users should not be forced to repeat these tasks. (We should also emphasize: the number of apparently undocumented data changes of the sort we have described was in fact small given the size of the data base. However, the discovery of even a single such instance causes the user to wonder just how widespread the problem is and results (at least in our case) in a process of carefully reviewing each and every data set for any possible further quirks. Advance knowledge that there are a few, but only a few, such quirks would have greatly simplified the process.)

The Economic Core

Since within-survey consistency errors were presumably detected and corrected in the MINCOME on-line data entry system, most of our effort has been spent in checking for consistencies across surveys. However, in the case of the economic core,

especially the job grid, we have also undertaken some additional within-survey checks, this for two reasons: First, some of the apparent across-survey errors may in fact be due to errors committed within a single interview. In order to detect the true across-survey errors, it was first necessary to understand the nature and magnitude of problems occurring within each periodic. Secondly, it seemed useful to verify the data quality within interviews, if only to confirm the correct operation of the quality controls built into the data entry system.

The job grid and associated components of the economic core present a large number of opportunities to check on the internal consistency of responses. To illustrate, the multiplication of wage rate by hours worked over a defined period gives a calculated amount of total earnings, which can be compared with the respondent's answer to the "total amount earned since DLI" question to check for consistent reporting. Reported periods of unemployment should likewise correspond to gaps in the job record for the specific weeks in question. A reported job change should correspond to some change in the occupation, industry, and wage rate codes; and so on.

The internal consistency checks that we performed involved the following variables: (Questions shown here are as they appear in Survey 7, but similar question wording and order are found on all other periodics.)

Q-1: Have you worked at any jobs since DLI?

Q-2: (IF YES): Name of the employer.

Q-3: Type of business.

Q-4: Occupation.

Q-5: Wage rate when job was started.

Q-6: Starting date for each job.

Q-7: Any job(s) started since DLI?

Q-15: Still employed (at each job)?

Q-16: (IF NO): Date the job was left.

Q-29: Current wage rate now (or when job ended)?

Q-30: Change in wage rate since DLI?

(IF YES) Most recent change: Date and amount.

Second most recent change: Date and amount.

Q-39: JOB GRID: Number of days worked each week since DLI.

Q-43: Total amount earned since DLI.

Q-44: Calculated total earnings since DLI.

Q-53: Unemployed since DLI?

(IF YES): Dates of unemployment.

As is apparent from the above list, the available variables allow for a large number of consistency checks; most of the possible checks are fairly obvious from the variable list and do not require elaboration here.

These consistency checks were examined for all male heads present in Surveys 3 through 11. The analysis is for male heads rather than all heads because the existing documentation of the data led us to believe--wrongly, as we learned later--that if the family had only one head, whether male or female, job data for that head would be recorded in the first occurence of MODULE 1. Such is not the case: Male heads are always listed first. If there is no male head and only a female head, the first position is "padded" with a dummy code and the job data for the female head are entered in the second occurence of MODULE 1. Thus, in reading MODULE 1 data from the first position only, we assumed we were picking up all heads but in fact picked up only male heads, a

mistake that was discovered too late to undertake additional analysis. Quick spot checks for female heads show, predictably, a much lower employment rate among females. It is, however, our impression that the internal consistency of the job grid data among the female heads is at least as good as that for the males.

[TABLE 4 HERE]

Results of these consistency checks are shown in Table 4. The first column shows the survey number; the second column indicates the number of males heads in that periodic whose job grid was scrutinized. Column 3 reports the count of participants who did not work during the relevant period. In column 4, we report the number of male heads who held more than one job during the relevant period; in a few cases, these are multiple jobs held simultaneously, but in most cases, they are different jobs held at different times in the relevant interval. Columns 5, 6 and 7 report the counts of male heads who (i) started a job, (ii) ended a job, and (iii) were unemployed at some time during the period.

Generally speaking, there are only a few errors that we were able to detect, and most of them are quite minor:

(1) In Survey 5, one person reported two simultaneous periods of unemployment, and another had a gap of three days between unemployment periods which cannot be accounted for.

(2) In Survey 7, one person has two periods of unemployment out of correct chronological order, that is, the dates for the second unemployment period are earlier than those for the first.

(3) In Survey 8, a person reported that he did not work at all during the period but has two unemployment periods listed; and there were two more people each missing three days for which we cannot account.

Table 4

Results of Consistency Checks on the Job Grid

Survey	N	No Work	2+ Jobs	# Started Work	# Ended Work	Number Unemployed
3	105	22	9	--	18	--
4	99	13	16	17	17	39
5	89	10	8	9	11	42(**)
6	89	11	9	11	10	17
7	82	13	6	7	13	12(*)
8	79	14(*)	7	12	10	10(**)
9	76	15	5(*)	7	8	9(*)
10	73	17	2	4	4	6
11	72	11	9	15	13	12

No Work = respondent has not worked any weeks since DLI.

2+ jobs = number of jobs held since DLI is greater than 1.

Started work = one or more current jobs was started since DLI.

Ended work = respondent quit one or more jobs since DLI.

Number unemployed = number reporting any instance of unemployment since DLI.

Asterisks indicate locations where errors were detected; see text for details.

(4) In Survey 9, we found one person holding two jobs simultaneously but the jobs were out of correct order; that is, the starting date for job 2 is earlier than the date for job 1. Another respondent in the same survey has three unaccounted-for days.

(5) In Survey 11, one respondent has two periods out of correct order; that is, the period with the earlier dates is listed second. For another respondent who held 2 jobs simultaneously, the data show a three-day overlap of the jobs, but both occupation and industry codes match so we cannot be sure exactly how this pattern is to be interpreted.

An additional and rather extensive set of checks was made using the dates appearing in the job grid for each periodic. The procedures we followed need not be described in detail. In general terms, we used the dates DLI and DI (dates of last and current interview) from MODULE 2 to calculate the total number of weeks in the target period and then we used the job grid in an attempt to account for each of those weeks. Some problems arose because of the treatment of partial weeks and of short periods of unemployment within the period covered by the interview. In almost all of these cases, however, it was possible to resolve the apparent discrepancy. In the vast majority of cases, in fact, all the entries in the job grid could be reconciled against the target interval. The only consistent problems resulted from zero entries in the job grid, indicating zero days worked for a particular period, which were, nevertheless, not accompanied by any indication of a period of unemployment. We suspect these entries

may reflect vacations or leaves of some sort but are unable to confirm this speculation. In any case, while this does occur with some frequency, the attending discrepancies are not very large.

In short, the internal consistency checks built into the MINCOME data entry system appear to have been quite successful. By far the largest share of the job data appearing in the job grid is internally consistent within a given interview. There are only a few overlapping periods reported, and only a few days for which we cannot account. (These days, interestingly, occur always in groups of three, suggesting, perhaps, that they result from things like long weekends or some similar ambiguity, but here too we can only speculate.) A couple of cases were found with unemployment periods out of correct order. But as a general conclusion, we are confident that the job data contain few errors that would cause problems in reconstructing and analyzing an employment time series.

Job Consistency Across Interviews

In addition to the within-survey checks just described, a series of checks on consistencies across all temporally adjacent interviews was also performed. Given our procedures for these comparisons, there are five general states of job consistency into which any given case may fall. These are reported below in Table 5.

First, a person may report not working in both interviews, an obviously consistent match the count of which is reported in column 3 of the table.

Secondly, the respondent may report working in both survey periods on the same job or jobs. If this is the case, the respondent should also meet the following four conditions:

(i) R reports "still employed" at the end of Survey N;

(ii) At Survey N+1, R reports that he did not start a job since DLI;

(iii) Occupation and industry codes match for N and N+1; and

(iv) Wage information is consistent (current wage on interview N matches starting wage on interview N+1).

The count of respondents who reported working at the same job(s) at both Survey N and Survey N+1, and who also passed all four of the above checks, is shown in column 4 of Table 5.

A third type of consistency can occur when the respondent reports not working in Survey N but says he is working at Survey N+1. For these cases, again, a series of conditions should be met: First, the reported date of the job start in Survey N+1 must exceed DLI. Secondly, the direct question about starting a job since DLI should be answered yes. Respondents showing this employment pattern for any given interval and passing both the above conditions are enumerated in column 5 of the table.

Fourth, a respondent may report that he was working in Survey N but not in Survey N+1. If so, there are again conditions that should be met: First, the direct question about "still working" should be answered "no" in Survey N; secondly, the ending date for the job should be less than the date of the interview. The count of persons with this job experience for any given interval and who passed both these conditional checks is reported in column 6 of the table.

Finally, true job changes may have occurred between Surveys N and N+1. "True job changes" can occur in several ways, and for each of them, a set of conditions must be met for the record to be considered consistent. The first type of true change occurs when the respondent has two jobs at Survey N. At Survey N+1, the first job has ended but the respondent remains employed in the second job. In Survey N+1, that second job will now be entered as job 1. This may well represent a source of some confusion for the naive user of the data (and is thus a matter that should be reported in the documentation), but it does not represent an inconsistency so long as job 2 at Survey N matches job 1 at Survey N+1 with respect to the occupation and industry codes and the reported wage rate. Cases showing this pattern and successfully "matching" between N and N+1 are recorded in column 7 of the table.

Another type of true change occurs when job 1 reported in Survey N ends within the interval, and another job (recorded as job 1 in Survey N+1) is started within the same interval. Here, the two job 1 entries are in fact two different jobs. If this is the case, then the following four conditions should also be met in order for the record to be considered consistent:

(i) The job ending date reported at Survey N should be less than the date of the interview and the question about still being employed at that job (job 1 at Survey N) should be answered "no."

(ii) The reported job start date for what is now job 1 in Survey N+1 must be greater than DLI, and the question about starting a job since DLI must be answered "yes."

(iii) The occupation and industry codes for the two jobs are (probably) different.

(iv) The reported wage rates are (probably) different.

The count of male heads showing this type of job change and successfully passing all four conditions is shown in column 8 of the table.

A final type of true job change is similar to the changes recorded in column seven, but rather more complicated (and, thankfully, quite rare in the data). Imagine a person who quits a job and is then rehired in the same job during a given survey interval. In the survey that occurs at the end of the interval, two jobs would be recorded--job 1 and job 2--since there were two distinct periods of employment, but both job 1 and job 2 would have identical occupation and industry codes. In Survey N+1, assuming the respondent was still employed at the same job, what was previously job 2 at Survey N would then become job 1. Since occupation and industry codes would be identical in all three entries, this poses no general problem. However, the wage rate information would typically not be identical in all three; and in this case, the proper check for consistency of wage rate information is between job 2 at Survey N and job 1 at Survey N+1. In a word, users of the data must be attentive to what changes in the wage rate information actually mean. The few cases where this type of change has occurred are recorded in column 9 of Table 5.

[TABLE 5 HERE]

Table 5

Consistency in Job Data Across Interview Periods

Survey Period	Common N	Columns* III	IV	V	VI	VII	VIII	IX	X
1-2	124	15	62	2	8	--	--	--	--
2-3	101	15	51	5	7	--	--	--	--
3-4	98	12	66	9	1	4	2	0	2
4-5	89	8	74	2	2	2	0	0	1
5-6	81	6	66	2	1	3	1	1	1
6-7	80	9	59	0	4	2	3	0	3
7-8	79	12	55	1	2	5	2	1	1
8-9	74	13	52	0	3	3	0	1	2
9-10	72	14	50	1	2	2	0	2	1
10-11	71	11	53	5	0	2	0	0	0

*See text for definitions of the III - X columns. Common N = the number of male head respondents present in both surveys. Note: The analysis reported in the text derives from results for Surveys 3 through 11 only. Prior to Survey 3, the economic core data were gathered in a quite different way, and we cannot examine the consistency of these data in the same way as followed for Surveys 3 to 11. We did make some limited checks on the data from Surveys 1 and 2 and detected no errors.

Finally, errors and inconsistencies may occur, that is, any given respondent could fail to pass the conditional checks appropriate to his job situation. The count of "errors" detected in this analysis is reported in column 10 of the table.

As is apparent from the table, the extensive cross-surveys consistency checking detected only a few errors. In fact, only 11 problem cases were uncovered. Nine of these eleven are of the same general form. A respondent at Survey N+1 says he is still employed at the same job as at Survey N. At Survey N+1, there is a job 1 entered and the respondent says he has not started any new job since DLI. But the occupation and industry codes, and/or wage rates, do not match for job 1 at Survey N and N+1, as they otherwise should. (Wage rate produces the inconsistency in five of these nine cases.) It is not at all certain that these are errors. It is conceivable that the "same job" could have changed, either in wage rate or job description. The confusion here may well be in the mind of the interviewer in phrasing the question, or in the mind of the respondent as he interprets it. But on the surface at least, these appear to be real changes that are simply not picked up by the change questions. One additional error was caused by two jobs that reverse order between N and N+1 (perhaps simply an interviewer transcription error).

One final error occurred because a respondent reported at Survey N+1 that he was still employed in job 1 from Survey N but also said at Survey N+1 that he had started the job since DLI. In this case, the two jobs (job 1 at Survey N, job 1 at Survey N+1) have different occupation and industry codes, and different wage rates;

thus it is likely that a new job had indeed been started in the interval and that the error was in the respondent not reporting that the old job had ended.

In summary, the cross-survey comparisons for the job grid/economic core data show little error and most of the detected problems could probably be easily resolved by a knowledgeable researcher. However, users do need to be made aware of the sorts of problems that will sometimes occur. While the overall rate of error is low, shortcut methods for tracing a job history through the data from a partial extract could result in a lot of different jobs being spliced together and considered as a single job. The major problem with these data appears not to be error or inconsistencies in the records, but the rather erratic coding conventions followed, and the complex data structure. Whether or not it is worthwhile for ISER staff to undertake a one-shot reconstruction of the job grid data (removing errors, enforcing a consistent set of coding conventions, etc.) is a question that lies beyond the scope of our work. But it must be emphasized that as things now stand, any user intending to analyze the labor supply data must become thoroughly familiar with the data and its numerous complexities and must plan for a fairly extensive visual inspection of the records, prior even to attempting to construct the time series.

Useability Issues

In addition to the problems discussed above, there are characteristics of the economic core data that are likely to increase the effort required of a typical user to get the data into a form suitable for statistical analysis. We raise these issues in no particular order and with no assessment of the relative seriousness of each one. Quite simply, they are problems we encountered in working with the data that are, in our judgment, serious enough or frustrating enough to warrant consideration by ISER staff.

1. When a respondent did not work at all during the interval covered by a particular survey, he or she should answer "no" to the question about working since DLI. In Surveys 3 to 6, the dates of unemployment for such people contain "-4's," indicating that the section is not applicable. For these surveys, one must calculate periods of unemployment from DLI and DI; examining the data in the unemployment section is no help because it looks just like the unemployment section for an employed person (i.e., is coded as not applicable). The unemployment sections in Surveys 7 to 11, in contrast, contain the actual dates (DLI to DI) indicating the period of unemployment. Having the data in this form makes it far easier to account for periods of unemployment. This inconsistency in format and convention across surveys could produce errors on the part of a naive user and it certainly increases the pre-analysis processing work for all users, naive or not.

2. In the job grid data, one is usually interested in the number of days worked since DLI. On all but one of the surveys, the coding of the first week's worth of "days worked" information begins in the first field of the job grid array and continues seriatum for n fields, where n is the number of weeks since DLI. Survey 5 is an exception, and a seriously confusing one. In Survey 5, the starting position of codes in the job grid begins in a variable position in the field. To illustrate, in our sample file, the first case begins in position 6 in the field, the second case begins in position 7, and the third case begins in position 13. This pattern does not correspond to any sort of calender pattern that we can find. Indeed, we have no idea as to why this variable pattern occurs or what it means. We have confirmed, however, that once the correct starting position has been ascertained, the job grid information contained in Survey 5 is good (that is, the number of weeks' worth of data entered is consistent with the elapsed time between DLI and DI).

3. Many inefficiencies in using the data result from the ordering of variables in the data file. This is especially true for MODULE 1 data. The extract tape with which we worked consists of 15 separate files, each containing data from a separate survey. Of course, the questions themselves varied across surveys and so the total number of variables in each survey will necessarily vary as well. However, the order in which variables are entered in the file does not correspond across data sets even within blocks of common questions. To give a single, simple example, all surveys contain the following six variables: Did R start a job since the

last interview (for three jobs); and the date on which the job was started (another 3 variables). The six variables in question appear in the above-stated order on Surveys 2-5. In Surveys 6-11, the two sets are entered in the reverse order. This change of order does not appear in the actual questionnaires, but does appear in the data base (or at least, in the extract available to us). The result of this and many similar instances that can be found throughout the data base is that each data file (15 in all) has its own unique data definitions and structure. The consequence is that it is very time-consuming even to do a simple listing of something as straightforward as sequential job information. FORMAT statements had to be tailor made (and tested) for each of the 15 data files, since we of course could not use the SAS setups that ISER provided.

4. We are more than a little puzzled by the variable blocked format of the data tape. Conceptually, the MINCOME data are often variable length. Households have differing numbers of members, heads of households have varying numbers of jobs, and so on. The MINCOME data transmitted to us, however, have been "squared off," with dummy codes entered to pad out otherwise blank fields. To illustrate, each family module has data on 15 members; unused fields are filled with "-4" codes. Thus, the data that we received are already in FIXED length format, and should therefore be stored and transmitted in fixed block format because every record has exactly the same field width. In contrast, the data we received are in variable block format, for no apparent reason.

We cannot see the advantage in distributing data in the variable blocked format. On the contrary, there is a considerable disadvantage. Specifically, variable block format is an IBM-specific tape format. Users working on equipment other than IBM therefore face a great deal of complexity and expense just in attempting to read the ISER data tape.

5. In addition to the variable blocked format, our extract tape was not written in character data, but rather in an internal IBM-specific format--a floating point binary format--which is a data representation that cannot be read on CDC (or most other) equipment without a great deal of trouble and expense (namely, a lot of custom programming). We understand the advantages of such a format for users working on an IBM operating system: FORMAT statements are not necessary and a complete codebook identifying the column locations of the variables is not needed. Once off IBM equipment, however, the advantage disappears: most computers simply cannot process these kinds of tapes. We were able to proceed in spite of this only because UMASS has a software package called FORM, a CDC-specific package that deciphers IBM-formatted tapes (gobbling up huge amounts of CPU time in the process). An average user who took the tape we received from ISER to the UMASS computer center would have been told by the consultants there to send the tapes back, as the data could not possibly be worth the expense necessary to read it. Even using FORM to decipher the tape requires extensive knowledge not only of the actual tape format, but also of the operating system on which the tape was

written, the operating system of the CDC installation where the tape is to be read, and the tape format requirements of the CDC operating system. Most users would encounter a great deal of difficulty working with a tape formatted in this manner.

Converting to character data is complex and time-consuming under these circumstances. Each variable is a different width (varying from 2 characters to 6 characters in our data). We suggest putting all extract data in a fixed field width (I6) that is a bit wasteful but makes it much easier for users who must reformat the data.

6. Users must be extremely cautious in using the arrays that contain variable amounts of information. At the present, the only way to be certain of getting all the appropriate data on, say, all family members or all jobs held in a given period, is to search through the total number of fields of information available. The "real" information (as opposed to padded dummy codes) does not always commence in the first position in the array. There may, or may not, be a "-2" end-of-information marker denoting the ending point of the data within the array. In some instances, no "-2" appeared at the correct spot; in a few cases, the "-2" was followed by additional substantive data. Specifically, this occurred in MODULE 2 of Survey 4, where data on the first member was stored in the 15th position; here, for some unfathomable reason, the entire array was shifted left-around by one position.

7. Job start and end dates are a source of confusion. If a job started or ended within a survey period, the date of the start

or end is entered on each survey. But if the job did not start since DLI or if the person is still working, the coding conventions differ from one survey to the next. On Surveys 3 through 6, the start and/or end dates are coded "-4," indicating "not applicable." On Surveys 7 through 11, in contrast, these dates for job start and job end are entered as DLI and DI. Here as in many other cases discussed above, users simply must be warned in advance about the differing coding conventions followed.

8. The coding of a simple matter such as dates also varies across surveys, and so what should be a straightforward calculation (one date compared with another) turns out instead to become a chore. In Surveys 1 through 4, the DDMMYY convention is followed (that is, two fields to record the day of the month [1 - 31], two fields to record the month [1 - 12], and two fields to code the last two digits of the year). In Surveys 5 through 10, a different convention is followed, the so-called YYDDD convention (two fields to encode the last two digits of the year, and three fields to encode the day of the year [1 - 366]). In Surveys 11 through 14 and 51, the original DDMMYY convention reappears. Nothing in the documentation alerts the user to these differences! (Inside a given survey, however, the conventions are followed consistently.)

9. Finally, the job grid is inconsistently formatted, the format followed in Surveys 4 and 11 being completely different from that used in all other surveys. In Surveys 4 and 11, the order is "days worked, first week, first job; days worked, first week, second job; days worked, first week, third job..." In all

other surveys, it is "days worked, first week, first job; days worked, second week, first job; days worked, third week, first job..." Ideally, ISER should reformat the grids in Surveys 4 and 11 to conform to those from other surveys. At minimum, some attention must be called to this problem when extracts are distributed.

Summary

Generally speaking, error and inconsistencies over time in the data records do not represent a serious barrier to time series analysis of the MINCOME data. Although the error rates for family composition fields are surprisingly high for families showing some compositional change, there are relatively few such families in the data, and for the most part, the data are relatively clean. We suspect that most of the errors detected in our analysis could be resolved with other information present somewhere in the archive. Thus, in principle, data quality is more than sufficient to sustain time series analyses.

In practice, inconsistent and apparently undocumented coding conventions make it difficult for a user to decipher the contents of the data file. Moreover, a non-IBM user is at a serious and potentially insurmountable disadvantage. None of the problems discussed here is completely unsolvable given sufficient diligence, a detailed familiarity with the MINCOME data base, and essentially unlimited computer time. But a typical user would probably meet none of these three criteria. If the concern at present is whether time series analysis of these data is possible, the answer is clearly yes. If the concern is whether ISER might do additional work on the master file to facilitate time series analysis, the answer is also clearly yes.