Clean Process Data Methodology- Emerging Approaches in the Management and Preparation of Longitudinal Analysis Sets

Poster Submission to the 2004 Meeting of the Population Association of America.

James W. McNally
NACDA
Institute for Social Research
University of Michigan
Ann Arbor, MI 48106
jmcnally@umich.edu
734-615-9520

Abstract

*Clean Process Data* represents a specific philosophy towards the cleaning and organizing complex secondary data that leaves a fully reproducible footprint of each stage of the data preparation process as well as a final analysis product. The organizational approach underlying this process represents a significant contribution to the methodology of data preparation because the storage structure allows complex transformations and reformatting of data to be easily located programmatically and exactly reproduced every time. This is important because the cleaning and preparation of secondary data represents an expensive and time consuming endeavor. The presence of standardized data preparation files not only enhances the efficiency of the research process, it also adds ongoing value to these studies as secondary data. This poster will provide an overview of the Clean Process approach and how it can be applied to the preparation of complex data to enhance use and facilitation analysis of research questions.

## Introduction

This paper presents results from ongoing research to formalize the work initiated by the University of Michigan economist, Lee Lillard in the development of *Clean Process Data* methodologies for archiving complex data collections. Using the Panel Study of Income Dynamics (PSID) as our baseline data, we are currently in the process of archiving, documenting and annotating data management program files and analysis data files underlying the *Clean Process Data* approach. The National Archive of Computerized Data on Aging (NACDA) currently supports the PSID clean process files, donated to the Inter-university Consortium for Political and Social Research (ICPSR) by Dr. Lillard's estate upon his death in December of 2000. During the course of his research career Dr. Lillard used the PSID analysis files generated by these clean process methodologies extensively, generating an impressive series of articles and reports. Unfortunately, while complete in content and in programming application, the process files (numbering 1,487) that underlie the preparation stages that result in clean process PSID data sets remained largely undocumented.

As part of a broader NIA funded initiative the author, NACDA and ISR researchers are working to accomplish three specific tasks to illustrate the value of the Lillard approach to the preparation of longitudinal analysis files. Initially we plan to fully document the Lillard *Clean Process Data* methodology; both as a specific application to aid researcher using the PSID and more generically as an approach that can be applied to other secondary data collections. Secondly, we plan to annotate the key theoretical issues as operationalized by the Lillard methodologies and their application to econometric and social research. As many of the aspects of the Lillard method help the data address standard econometric problems such as simultaneity and endogeneity as well as the organization and structure of the data itself; the underlying rationale for decision rules needs to be documented and annotated. Finally, we plan to expand the principles of *Clean Process Data* methodology to new data collections, creating an interrelated and interchangeable collection of data sets that share commonalties in structure and application.

A key outcome emerging from this research project will be a standard set of decision rules that underlie the *Clean Process Data* approach. These rules if applied to other complex data sets such as HRS, SIPP and NLS, enhancing the standardized use of longitudinal data and creating a library of clean processes, which could be combined in a variety of ways depending on the research topic. Finally, the project will provide a fuller understanding of the methodological contributions of Lee Lillard and his approach to the data processes underlying his analysis and interpretation of results.

## Data and Methodological Issues

*Clean Process Data* represents a specific philosophy towards the cleaning and organizing complex secondary data that leaves a fully reproducible footprint of each stage of the data preparation process as well as a final analysis product. The organizational approach underlying this process represents a significant contribution to the methodology of data preparation because the storage structure allows complex transformations and reformatting of data to be easily located programmatically and exactly reproduced every time. This is important because the cleaning and preparation of secondary data represents an expensive and time consuming endeavor, and this cost increases substantially when processing steps are lost or undocumented. The presence of standardized data preparation files as part of an archived data collection not only enhances the efficiency of the research process, it also adds ongoing value to these studies as sources of secondary data.

The preservation of standardized data collections avoids, not only the costly repetition of the data cleaning process by future users, it also simplifies the analytic process leading to the replication and validation of existing results. Consequently, the opportunity to preserve a fully harmonized collection of PSID analysis files for use by ancillary researchers represents an important contribution to the overall archival process. Equally important to the archival process is the need for all the procedures employed to be annotation so new users can employ the clean data intelligently and verify the validity of the harmonization process itself. Further, this kind of system will encourage the addition of enhancements to the underlying files by users as these files are offered as *open source* tools to be revised and expanded by researchers with an interest in data preparation methodology. An essential first step in accomplishing these archival goals, however, is performing the thorough documentation of the methodology and programming steps that lead to the *Clean Process Data* analysis file.

**Clean Process Described**

Due to the untimely death of Dr. Lillard, the preliminary work underlying *Clean Process Data* is seen as part of an ongoing body of work that was unexpectedly interrupted. Dr. Lillard created this approach as a tool to assist in managing a consistent and replicable body of research that spanned his career and involved a wide array of collaborators. Consequently, while we have a broad and well developed system of data management, this system was still in the process of revision and refinement at the time of Dr. Lillard's death and as such was not documented for the purpose of easy use by uninitiated secondary users. Still, one of the most important parts of Dr. Lillard's scientific bequest lay in his desire to share his ideas for managing, processing, storing and using the data. All his data preparations were done in a consistent way, regardless of the nature of analysis for every data source he worked with. His main idea was to create a library of clean processes, which can then be combined in a variety of ways depending on the project.[1] The application of this philosophy is reflected in the voluminous publications that emerged from analysis employing the *Clean Data Process* PSID files (see bibliography). During the course of his research career, Dr. Lillard defined a series of underlying constructs for the Clean Process Data approach and defined a steps of steps leading up the creation of processed analysis files.

What is a "process"?

*Clean Process Data* represents a structured methodology that takes a variable or a set of variables that can be conceptually thought of as being related to one another and systematically walks the variables through a series of steps that recode, reformat and refine them in according to a set of well defined decision rules. These rules can be as simple as "if sex =1 then male =1; else male = 0;" to set up a consistent gender dummy variable or as complex as allocation algorithms for missing data or the construction or disaggregating labor income from other income streams.

Put simply, a process is a data set (typically in SAS or STATA format), that contains the results of preparing and cleaning the raw data, resolving all inconsistencies, imputing missing values, and performing all tasks associated with the creation of complete analysis file. The primary difference between *Clean Process Data* and other forms of recodes and data manipulations is that it is: 1) based upon a consistent design philosophy and set of decision rules, and 2) is represents "open source code" in that it is publicly archived, available to all interested researchers and open to change. Consequently it represents a template of a working data preparation methodology that is dynamic in structure, able to develop and adapt in the face of new information and advances by researchers and to new additions to the baseline secondary data source.

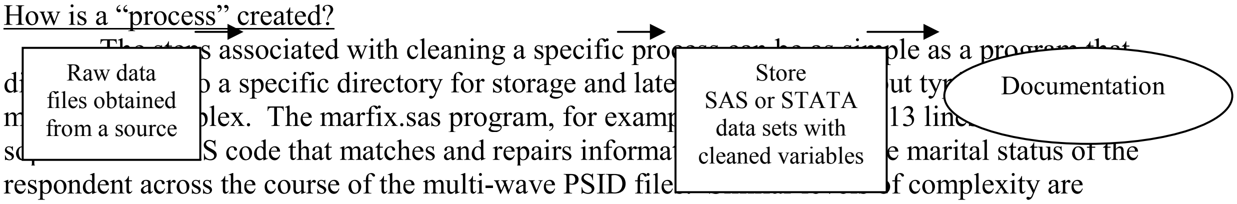
Why are "processes" useful?

There are a number of advantages associated with the use of standardized processes in the creation of data analysis files. A uniform system of handling data makes it easy to switch between components of a complex data set, or switch from one data source to another, and for one user to replicate or build upon the work of another user. Similarly, multiple processes

[1] This statement and much of the following information is based upon the ICPSR Study Documentation #1239.

cleaned in a consistent way are easy to use in analysis, conjointly or independently. The use of processes to separate complex data sets into distinct sets of related variables allows users the option of merging only the files that are directly relevant to a specific research goal. At the same time, having access to the extensive library of processes that makes up the *Clean Process Data* collection allows users to verify hypotheses and ideas quickly and efficiently. Finally, the open source structure of the process files allows users to modify and expand the library of processes while maintaining an ongoing reproducible footprint of the data preparation process.

How is a “process” created?


Raw data files obtained from a source
Store SAS or STATA data sets with cleaned variables
Documentation

The steps associated with cleaning a specific pro ple as a program that d a specific directory for storage and late ut ty m plex. The marfix.sas program, for examp 13 lin s S code that matches and repairs informa e marital status of the respondent across the course of the multi-wave PSID file f complexity are found in the creation of consistent income variable, measures of labor force participation and health. Regardless of the complexity of the specific task or process, however, they all follow a consistent set of processing rules .

How to store a “process”?

One of the strengths of the Lillard methodology is the consistent way that all processes create integrated modules or subsets of the original raw data and that these modules are stored in a reserved location on disc, treated explicitly as being part of a data library. Conceptually it is as if the PSID had been divided logically into a series of related book chapters with the library catalog serving as binding of the book. The use of a standard naming convention for all directories and for the *clean process* modules contained inside a specific library makes it easy to locate relevant information. For example, each SAS program associated with Stage 2 of the process creation always has its corresponding log (program report) file and an output (results) file containing listing of the variable transformations resulting from the program scripts. These are then all associated within the program folder that refers to a specific *clean process* such as “income” or “marriage history”. This logical structure means that another researcher with little direct experience with PSID can simply use the existing *clean process* module for analysis or they can run a process generating program and create their own *clean process* analysis file, but in doing so they can also compare their results to the output listing files associated with the existing process contained in the process library as a consistency check.

C.1.5 What types of processes are there?

While almost any variable in a data set can be considered to be part of a process, some types of data present special challenges that lend themselves to the Lillard approach. As they only have to be resolved once under this methodology, researchers with similar analysis questions can share the resulting processes, saving valuable research time and encouraging rapid developments on the topic addressed by the process. Two different types of data concerns common in the analysis of longitudinal data were of particular interest to Dr. Lillard when he created the PSID file. Dr. Lillard referred to the programming resolution of these concerns as *point in time processes* and *event history processes*.

- *Point in time processes* are repeat measure variables that can be represented as an array of values; each value containing information about the respondent’s status at each observation or survey date.
- *Event history processes* represent a series of arrays that logically describe the beginning and end dates of certain events in respondent’s life such as childbirth, marriage histories or migration behaviors.

## Conclusion

The proposed poster will present an introduction and applications of the Clean Process Methodologies that will enhance the use of longitudinal data for analysis and research development.