

CONTEXTUAL DATA FILES FROM THE 1995 NATIONAL SURVEY OF FAMILY GROWTH: ACCESS AND ANALYSES

Linda J. Piccinino and William D. Mosher
National Center for Health Statistics

Contextual Data and the NSFG

The National Survey of Family Growth first constructed a contextual data file to supplement its 1982 survey of women of childbearing age. The motivation for this was that up until then, most of the knowledge about fertility and contraceptive behavior, two main thrusts of the NSFG, was at the microlevel. While microlevel data are useful for many things, these kinds of data, when used alone, bias possible analyses. Because community level factors were not usually measured in individual level survey data, community context was often omitted when fertility rates were being explained.^[1] However, now that there are datasets and statistical techniques that allow us to measure community level variables as well as individual variables, there is a growing recognition that the characteristics of places in which women and couples live shapes their reproductive decisions by influencing individual alternatives and their associated social, economic and psychic costs.^[2]

Contextual data are collected to provide information on the context in which individual attitudes, behavior, or other experiences take place. Data on the characteristics of the communities where survey respondents live are useful to address issues relevant to policy and program planning. Contextual data from the National Survey of Family Growth 1995 (NSFG95) are now available, on approval, to qualified users. These data include variables at up to four levels of aggregation (state, county, census tract, block group) and for calendar years (1990, 1993, 1995). The richness of the NSFG95 data file is increased by the ability to link NSFG95 main study files to geocoded data compiled from various government and private sources.

Research with contextual data has clear policy relevance and can be used to determine whether policies should be aimed at individuals, or at their neighborhoods, or both. Examples of policy research topics are: the effects of state welfare policies on teen sexual activity and childbearing; the effects of local economic conditions on infant mortality and birth rates; and the effect of the availability of county health and family planning clinics on sexually transmitted diseases (STDs) and contraceptive use.

The Need for Confidential Access Systems

Legal restrictions prevent NCHS from releasing contextual data files directly to the general public. The dictates of Section 308(d) of the Public Health Service Act make it illegal for the National Center for Health Statistics (NCHS) to release data to anyone outside of NCHS that could be used to identify survey respondents. These restrictions are necessary because contextual (neighborhood, community, macrolevel) data are particularly sensitive and also can be used to show the characteristics of very small and easily identifiable areas. The confidentiality pledges, made by NCHS and other researchers, could not be possible if data were released that allowed respondents to be identified.

NSFG Contextual Data Access Procedures

Although confidentiality disclosure limitations apply, access to NSFG contextual data is open to all *bona fide* researchers for a reasonable fee. First, the researcher must apply to NCHS for access. They will then receive a

dummy data file complete with full documentation to familiarize themselves with the data, and to set-up and test their computer programs. To better accommodate the needs of researchers wishing to use the NSFG contextual data, NCHS has devised three options for accessing the data, thereby optimizing the opportunity for analysis of the data while preserving confidentiality strictures. These options include remote access, on-site access, and access through collaboration.

At this time, these are the three access options that are available: remote access, on-site access, or access through collaboration. If the researcher is able to use SAS, remote access can be used. If other software is indicated, then on-site or collaborative access is suggested.

Remote access

Under the remote access option, the NSFG95 contextual data files are available to qualified researchers through a remote system called ANDRE (ANalytic Data Research by E-mail).

SAS programs written by researchers to analyze the NSFG95 contextual data are submitted directly in ASCII format via e-mail to ANDRE to be run in-house 24 hours a day, 365 days a year. Certain SAS procedures and SAS functions are not allowed. For example, users may not use PROC TABULATE or PROC IML or functions that can produce listings of individual cases such as the commands LIST and PRINT. Functions which select for individual cases also are not allowed, and minimum cell sizes must be maintained in all frequency tables. For the NSFG, cells containing fewer than 5 cases will be suppressed and, if necessary, additional cells will be suppressed (complementary suppression) to disallow calculation of suppressed cell sizes from marginals. The analyst will be informed of all prohibited functions and commands when they register for remote access. All output and job logs are scanned for violations. If the program does not break any of the confidentiality rules, the output is returned via e-mail to the researcher. Researchers are notified if there is a problem, and told how to remedy it. This procedure, which is similar to the one used for the Luxembourg Income Study,^[3] is now in full operation.

Prospective users of these data files will sign a confidentiality agreement with NCHS, and then move through the remote access process:

- > Client writes computer programs; tests programs on the >dummy= data set.
 - > Client e-mails >finalized= programs to ANDRE or to contact person at NCHS.
 - > NCHS runs programs.
 - > ANDRE (or contact person) checks for confidentiality breaches.
 - > If problems, NCHS e-mails program back to client for modification.
- If no problems, NCHS provides client with output.

On-site access

Contextual data file users may choose to do their analytic work on-site at the new Research Data Center (RDC), a facility located at NCHS in Hyattsville, Maryland. Researchers with approved projects are granted access to the data files, but must complete all work within the confines of the RDC. Electronic or hardcopies of data files or file documents may be removed from the RDC after being reviewed by RDC staff for disclosure risk.

Since resources and physical space are limited, researchers are advised to contact the RDC well in advance of their planned work visit to ensure that their research proposal has been processed and accepted, and that all necessary data files, computer hardware, software, and staff are available at the time of their arrival. Researchers may only work at the RDC under the supervision of NCHS-RDC staff during normal working hours. There is a fee charged for use of the RDC facility. This fee helps to defray part of the cost for equipment, workspace, NCHS staff time for monitoring, technical assistance, disclosure limitation review, and file management. Files needing special handling and set-up, or custom file formats, can be accommodated for an additional charge.

Collaboration

NSFG staff welcome the opportunity to work with researchers from outside agencies, academic institutions, and non-profit organizations, etc. to analyze data and publish reports using the contextual data files. Collaborations are formed on an *ad hoc* basis and depend on the joint interests of the prospective researchers. While these collaborations are encouraged, the availability of the NSFG staff is limited. The NSFG staff will request joint authorship on all collaborative projects and will be required to submit all manuscripts to NCHS for clearance.

Staff programming is another alternative for users who need a lot of technical assistance with large complex analytic projects or who do not have access to programming support at their home institution. Researchers are advised to contact NCHS-RDC to perform programming tasks necessary for their research project.

Contextual Data Analysis Software

NCHS currently supports use of several software packages for analysis of the NSFG contextual data files. SAS is the primary and most popular package, and is the one that can be used with our remote access system, ANDRE. Other packages, including LIMDEP, HLM (Version 4, for Windows) and STATA, are available for use but must be run manually by an NCHS staff person on-site. SAS PROC MIXED is also being made available.^[4]

Origins and Types of Contextual Data in the NSFG

A contextual data file is the set of characteristics of the environment in which the respondents (women interviewed) live. The 1995 NSFG contextual data file is actually a set of three main files, one file for variables describing the area in which the respondent lived at one of three dates: 1990, 1993, and 1995. Each contextual data file contains 10,847 observations, one for each respondent in the NSFG. The number of states (including the District of Columbia), counties, tracts and block groups for the respondent residence years are shown in Table 1.

Table 1. Number of Geographic Areas Represented by NSFG95 Respondents= Residences

Year	States	Counties	Tracts	Blocks
1990	51	972	5591	6208
1993	51	815	5605	6228
1995	51	890	6250	7000

Source: Battelle, 1997.

Source files

Variables were culled from several existing files, e.g. census data, area resource files and private databases. Data on where the women lived are shown by state, county, census tract, and block group. Hundreds of variables are included, with variables on topics ranging from the availability of family planning service providers per capita to median Hispanic household income to unemployment rates by race and sex.

Each variable in the NSFG95 contextual data file was derived from one or more of the source files listed below. When multiple sources were used to calculate a variable, the numerator was used to code the data source of the constructed variable.^[5]

- > US Census Bureau files:
 - Population Estimates
 - State Government Finances
 - United States Summary
- > Area Resource File, 1995
- > Census of Population and Housing, 1990: Summary Tape File 3A
- > HCFA File
- > Dept. of Labor ETA Form Files
- > CDC STD File
- > NCHS Natality Tapes
- > USA Counties, 1994
- > Alan Guttmacher Institute Files
- > Uniform Crime Reports, US-FBI
- > The Green Book

Time points

The NSFG contextual file contains a rich set of information for the respondents' residences at three key points in time:

- 1990 -- decennial census year
- 1993 -- year of NHIS from which NSFG household sample drawn
- 1995 -- NSFG survey year

Levels of aggregation

Data at different geographic levels are shown for the state, county, census tract & block group in which the sample women resided in 1990, 1993 and 1995. Note that many variables are only available at the state or county level, and that many variables are highly correlated with each other. While these may be a limitation for some analyses, both of these factors reduce the chances of disclosure.

Examples of variables available

- > Percent of population that is black, white, Hispanic
- > Median rent; median value of homes
- > Median family income; median household income
- > Percent receiving public assistance
- > Average value of public assistance
- > Unemployment rate
- > Percent below poverty level
- > Rates of gonorrhea, chlamydia, syphilis
- > Crime rates (violent, property, and total)
- > AFDC payment per family, or per recipient
- > AFDC income cut-off
- > Infant death rates by race
- > Family planning providers per capita
- > Medicaid payments per capita
- > Abortion rates

Analyses Using the NSFG Contextual Data

Contextual items cover a broad range of topics, from poverty and income data to information on health services and sexually transmitted disease infection rates. A variety of variables from some of these subject areas, as well as others, illustrate the kinds of research results that can be generated with these data files.

Test researchers

The NCHS remote data access system was first tested with the help of a group of researchers^[6] that were interested in doing contextual data analyses with NSFG data. Although extensive testing was done in-house by NCHS staff, testing by outside researchers was also desirable to develop a product and system that was usable by a larger research community with its unique interests and needs. The test researchers enabled us to refine our remote automated system, and alerted us to additional types of software that might be requested as add-ons to our manual submission process. They also gave us practice in merging outside data sources with our contextual data files, and helped us streamline our disclosure review process.

Current research

In an effort to stimulate research on the NSFG contextual data, a paper session was held at the NCHS National Conference on Health Statistics, in August, 1999, in Washington, DC. At this session four groups of researchers presented their analyses that demonstrated a variety of methodologies, software, and research questions. Some of their work is discussed below.

In their paper on economic incentives, teenage sexual activity, and contraceptive use, Argys et al.^[7] merged in their own state-level policy variables with NSFG95 data and submitted programs remotely to create a SAS output analysis data file of contextual and individual variables. They then worked on-site at the NCHS RDC and used LIMDEP to jointly estimate the probability of being sexually active and, conditional on being sexually active, the probability of using contraception, using a series of bivariate probit models. Specific contextual variables from the NSFG file that were of interest in their study included county-level abortion providers and family planning clinics, median household income, population density, the male/female ratio, the AIDS death rate, and others. State-level variables included per-capita state government expenditures on health and hospitals, and syphilis and gonorrhea rates.

They found that while some contextual variables had effects, some did not. The availability of family planning services (number of clinics per 10,000 women in the county of residence) had no effect on the probability that a teen was sexually active, which was encouraging from a policy perspective. They also found strong evidence that family planning availability was linked to an increased probability of contraceptive use. Additionally, they indicated that neighborhood context (as measured by median income in the census tract) influenced both the decision to become sexually active and the decision to use contraception.

Brackbill and Piccinino^[8] looked at two important outcome variables -- condom use versus non-use, and dual use of condoms and other methods -- with the general objective of explaining effects of individual and contextual variables.^[9] This approach was intriguing in that it additionally used contextual factors to explore outcome behaviors that are generally only described in terms of individual characteristics. Wherever possible they used block level contextual variables assuming that the measured population level characteristic was more likely to encompass the environment in which the respondents lived. When correlation matrices of county and block level variables were compared, it was shown that block level measures had a higher correlation with individual level variables. Multilevel logistic regression models were used in an exploratory approach to assess possible contextual influences on behaviors. These model types are limited, however, because of the variance constraints on the higher level variables in a mixed (fixed and random effects) level regression. This research, too, found strong effects of family planning service providers on the outcome variables. Analysis runs were done with a combination of on-site and remote SAS submissions.

Kanaiaupuni and Fomby^[10] multidimensional approach placed an emphasis on policy variables, though the impact of contextual variables in this paper was not as clear cut as in the Argys et al.^[11] These researchers analyzed variables at the individual, county, and census tract level using SAS and HLM, a software package for analyzing hierarchical linear models. In this instance, SAS programs were submitted remotely for early exploratory work. Later, SAS programs were e-mailed to an NSFG staff person where they were used to

create a set of intermediate confidential output files. These files, because they contained identifiers, remained in-house and were used as input for the series of HLM runs performed by staff of the RDC.

In the first stage of a three-part proposal, Manlove and Terry merged in state-level data from 1980 to 1995 generated from vital statistics, the Department of Education and census data to run multivariate hazard models to analyze the effects of family, individual, partner and community-level characteristics on the risk of teen motherhood and non-marital teen motherhood across three cohorts of teens.^[12] Another researcher, Zavodny of the Federal Reserve Bank of Atlanta, Georgia, currently is working with the RDC to run nested logits in STATA for her project on characteristics of teens and their partners, and resolution of their pregnancies. Contextual variables of interest in this analysis are abortion providers per capita in the state, whether the state allows Medicaid-funded abortions, whether the state has a parental consent law, as well as variables measuring the generosity of welfare benefits.

Appendix

Hierarchical Linear Models (HLM)

NCHS now has on-site HLM Version 4.04 by Anthony Bryk, Stephen Raudenbush and Richard Congdon, Jr. This software package can read data from SAS and STATA, among other packages. It allows estimation of two- and three-level models, e.g. models using data observations within persons, persons within communities, and communities within states.^[13] This software is available for purchase and license through Scientific Software International, Inc., 7383 N. Lincoln Avenue, Suite 100, Chicago, IL 60646-1704.

LIMDEP

LIMDEP is a general econometrics program for estimating linear and non-linear regression models, primarily for cross-sectional, time-series and panel data. It is available through Econometric Software, Inc., 15 Gloria Place, Plainview, NY 11803.

SAS PROC MIXED

This procedure is offered as a SAS System feature through the SAS Institute Inc. It's basic function is to fit linear models, and can sometimes used in the analysis of multilevel models.

More information may be obtained from the SAS Institute Inc., Box 8000, Cary, NC 27511-8000.

STATA

STATA is a powerful general purpose package for data analysis and data management, with graphics capabilities and a graphics editor. STATA covers a wide range of statistical techniques and is programmable, allowing the user to add new commands. For more information contact STATA, 702 University Drive E., College Station, TX 77840.

^[1] WR Grady, DH Klepinger and JOG Billy, 1993, The Influence of Community Characteristics on the Practice of Effective Contraception. FPP 25(1), pp. 4-11.

^[2] Battelle, 1997, User Documentation for the National Survey of Family Growth, Cycle V, Contextual Database, Final Report.

^[3] Journal of Official Statistics, 1993, Supplement. See also S Fienberg and L Willenborg, 1998, Introduction to the Special Issue: Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data, Journal of Official Statistics.

^[4] These packages are described briefly in the Appendix.

^[5] Ibid. Battelle, 1997, Appendix B.

- [6] From academic and government research environments.
- [7] LM Argys et al., 1999, The Impact of Economic Incentives on Teenage Sexual Activity and Contraceptive Use. Paper presented at the National Conference on Health Statistics, August 2-4, 1999, Washington, DC.
- [8] RM Brackbill and LJ Piccinino, 1999, The Influence of Individual and Contextual Variables on Condom Use to Prevent Sexually Transmitted Diseases. Paper presented at the National Conference on Health Statistics, August 2-4, 1999, Washington, D.C.
- [9] Much of the early analysis for this paper was done using remote submissions through the automated system ANDRE.
- [10] S Kanaiupuni and P Fomby, 1999, Assessing the Impact of Family Planning Services on Individual Contraceptive Use and Pregnancy: A Multilevel Analysis. Paper presented at the National Conference on Health Statistics, August 2-4, 1999, Washington, D.C.
- [11] LM Argys et al., *ibid.*
- [12] J Manlove and E Terry, 1999, The Effect of Contextual Factors on Demographic Trends in Teen Fertility. Poster presented at the National Conference on Health Statistics, August 2-4, 1999, Washington, D.C.
- [13] A Bryk et al., 1996, Hierarchical Linear and Nonlinear Modeling with the HLM/2L and HLM/3L Programs. Chicago, IL: Scientific Software International, Inc.