

# Introduction to IPUMS and NHGIS

Katie Genadek  
Minnesota Population Center  
University of Minnesota  
kgenadek@umn.edu



The IPUMS projects are funded by the National Science Foundation and the National Institutes of Health



# Minnesota Population Center

Home of the IPUMS, NHGIS, and IHIS



The MPC is one of the world's leading developers of demographic data resources. We provide population data to thousands of researchers, policymakers, teachers, and students. All MPC data are available free over the internet.

## Integrated Public Use Microdata Series



### IPUMS-International

Harmonized data for 1960 forward, covering 544 million people in 238 censuses from around the world.



### IPUMS-USA

Harmonized data on people in the U.S. census and American Community Survey, from 1850 to the present.



### IPUMS-CPS

Harmonized data on people in the Current Population Survey, every March from 1962 to the present.



### North Atlantic Population Project

Complete-count data from 1800s censuses of Canada, Great Britain, Germany, Iceland, Norway, Sweden, and the U.S.

## Other MPC Projects



### National Historical Geographic Information System

Tabular U.S. census data and GIS boundary files from 1790 to the present.



### Integrated Health Interview Series

Annual harmonized data on people in the U.S. National Health Interview Survey from the 1960s to the present.



### American Time Use Survey-X

Annual harmonized data from 2003 forward on how U.S. adults divide their time among activities.



### Integrated Demographic and Health Series

Demographic and Health Surveys integrated for analysis across time and space, 1980s forward.



### Terra Populus

Integrated data on population and the environment, from 1960 to the present.

<http://www.popdata.org/>

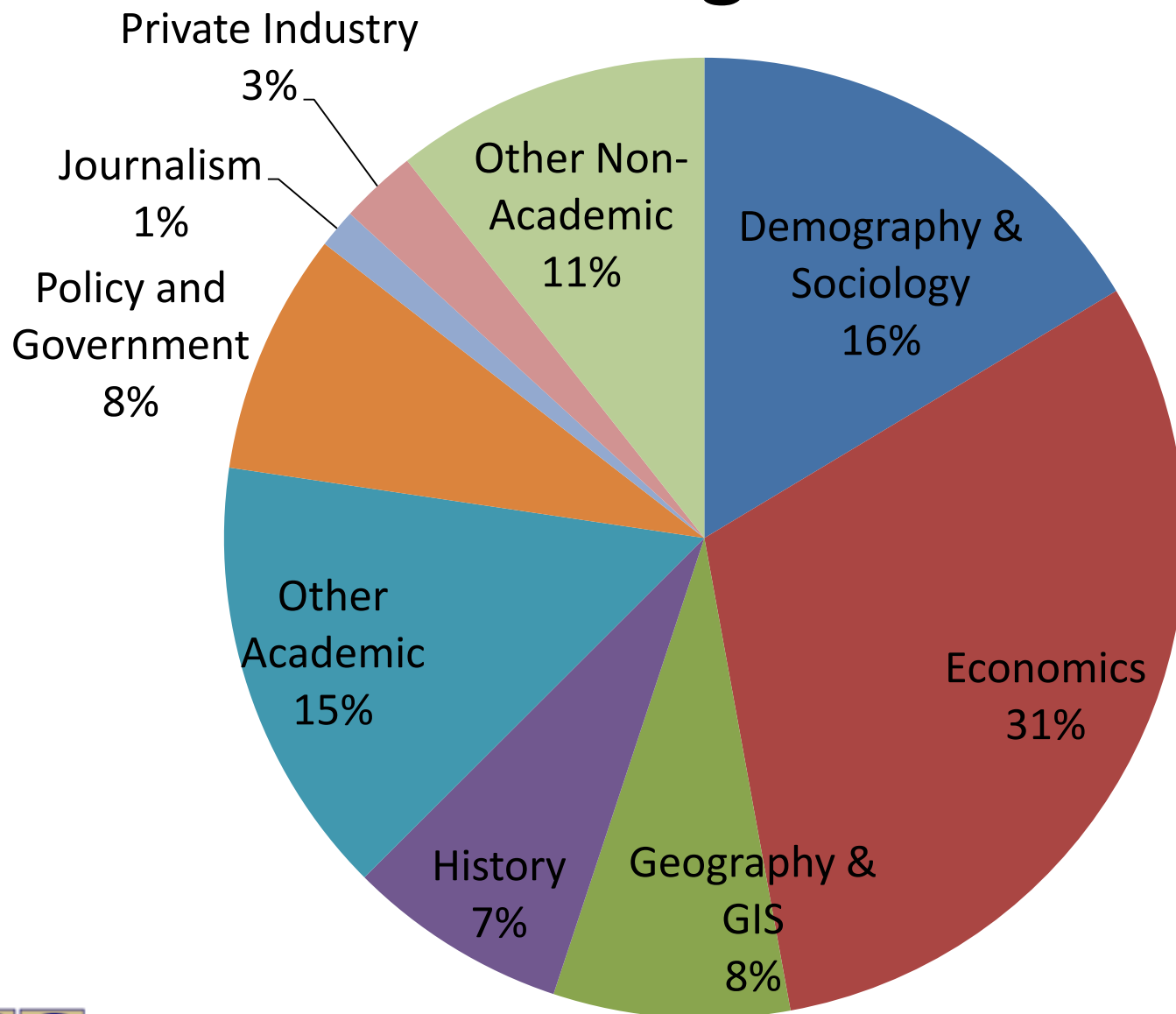
# The Minnesota Population Center

- University-wide interdisciplinary research
  - Members from 23 departments
- Education and Training
  - Demography curriculum
  - Professional and research training
  - Seminars and workshops
- Revolutionary Data Tools
- Home of the IPUMS and other MPC data projects
- More than 200 employees

# What the MPC data projects do:

- Standardizes data across years (integration)
- Provides excellent documentation
- Provides “bonus” calculated fields
- Provides data in multiple formats
- Makes data freely available online
- Provides user support

# Who is using MPC data?



# SUMMARY DATA

# Summary Data

Age	Both sexes	Male	Female
<b>Total population</b>	<b>281,421,906</b>	<b>138,053,563</b>	<b>143,368,343</b>
Under 5 years	19,175,798	9,810,733	9,365,065
5 to 9 years	20,549,505	10,523,277	10,026,228
10 to 14 years	20,528,072	10,520,197	10,007,875
15 to 19 years	20,219,890	10,391,004	9,828,886
20 to 24 years	18,964,001	9,687,814	9,276,187
25 to 29 years	19,381,336	9,798,760	9,582,576
30 to 34 years	20,510,388	10,321,769	10,188,619
35 to 39 years	22,706,664	11,318,696	11,387,968
40 to 44 years	22,441,863	11,129,102	11,312,761
45 to 49 years	20,092,404	9,889,506	10,202,898
50 to 54 years	17,585,548	8,607,724	8,977,824
55 to 59 years	13,469,237	6,508,729	6,960,508
60 to 64 years	10,805,447	5,136,627	5,668,820
65 to 69 years	9,533,545	4,400,362	5,133,183
70 to 74 years	8,857,441	3,902,912	4,954,529
75 to 79 years	7,415,813	3,044,456	4,371,357
80 to 84 years	4,945,367	1,834,897	3,110,470
85 to 89 years	2,789,818	876,501	1,913,317
90 years and over	1,449,769	350,497	1,099,272

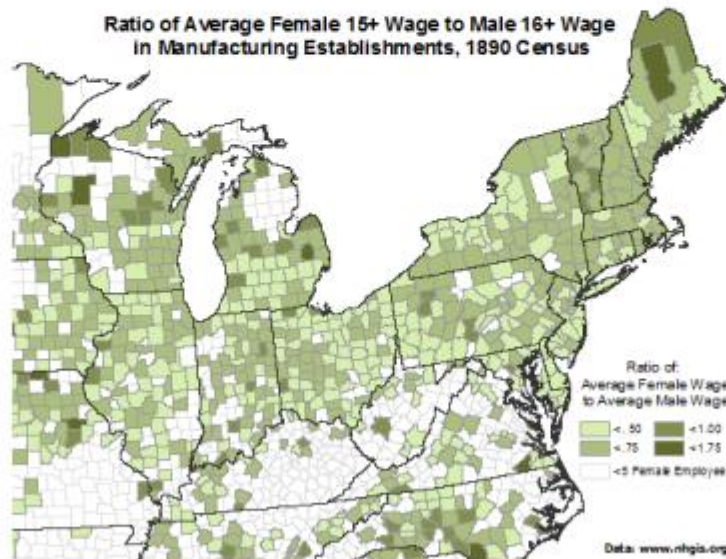
## NHGIS National Historical Geographic Information System

[Home](#) | [Select Data](#) | [FAQ](#) | [Login](#)

### Welcome to NHGIS

The National Historical Geographic Information System (NHGIS) provides, free of charge, aggregate census data and GIS-compatible boundary files for the United States between 1790 and 2012.

Ratio of Average Female 15+ Wage to Male 16+ Wage  
in Manufacturing Establishments, 1890 Census



[Click here to see previous front page maps.](#)

### NHGIS News

[Time series now include  
2008-2012 ACS data](#)

[2012 ACS 5-Year  
Summary File now  
available](#)

[New 2012 & improved  
2000 GIS boundary files](#)

[<read more>](#)

### Other MPC Projects

[IPUMS-International](#)

[IPUMS-USA](#)

[IPUMS-CPS](#)

[NAPP](#)

[IHIS](#)

[ATUS-X](#)

### Data

[Select Data](#)  
[Download/Revise Extracts](#)

### User Resources

[Project Description](#)  
[User's Guide](#)  
[FAQ](#)  
[Data Availability](#)  
[Mapping Options](#)

### Data Documentation

[Tabular Data Sources](#)  
[Time Series Tables](#)  
[GIS Boundary Files](#)  
[Release Log](#)

### Research

[Citation and Use](#)  
[Related Sites](#)  
[MPC Bibliography](#)

### Contact Us

[Technical Assistance](#)  
[NHGIS Staff](#)

**Select  
Data**

Tables & GIS Files



# NHGIS Data

## U.S. Census aggregate data, 1790-2012

Data from population and agricultural censuses, 1790 - 1960

- Nation, state, county, and tract geographic levels

## All electronic summary files, 1970 - 2010

- All tabulated geographic levels available
  - Blocks, Block Groups, Tracts, Places, County Subdivisions, Counties, States, and more

## American Community Survey

- 2010, 2011 & 2012 1, 3 & 5-Year Datasets
- 2009 and older 1 & 3-Year with 5-Year Datasets to follow

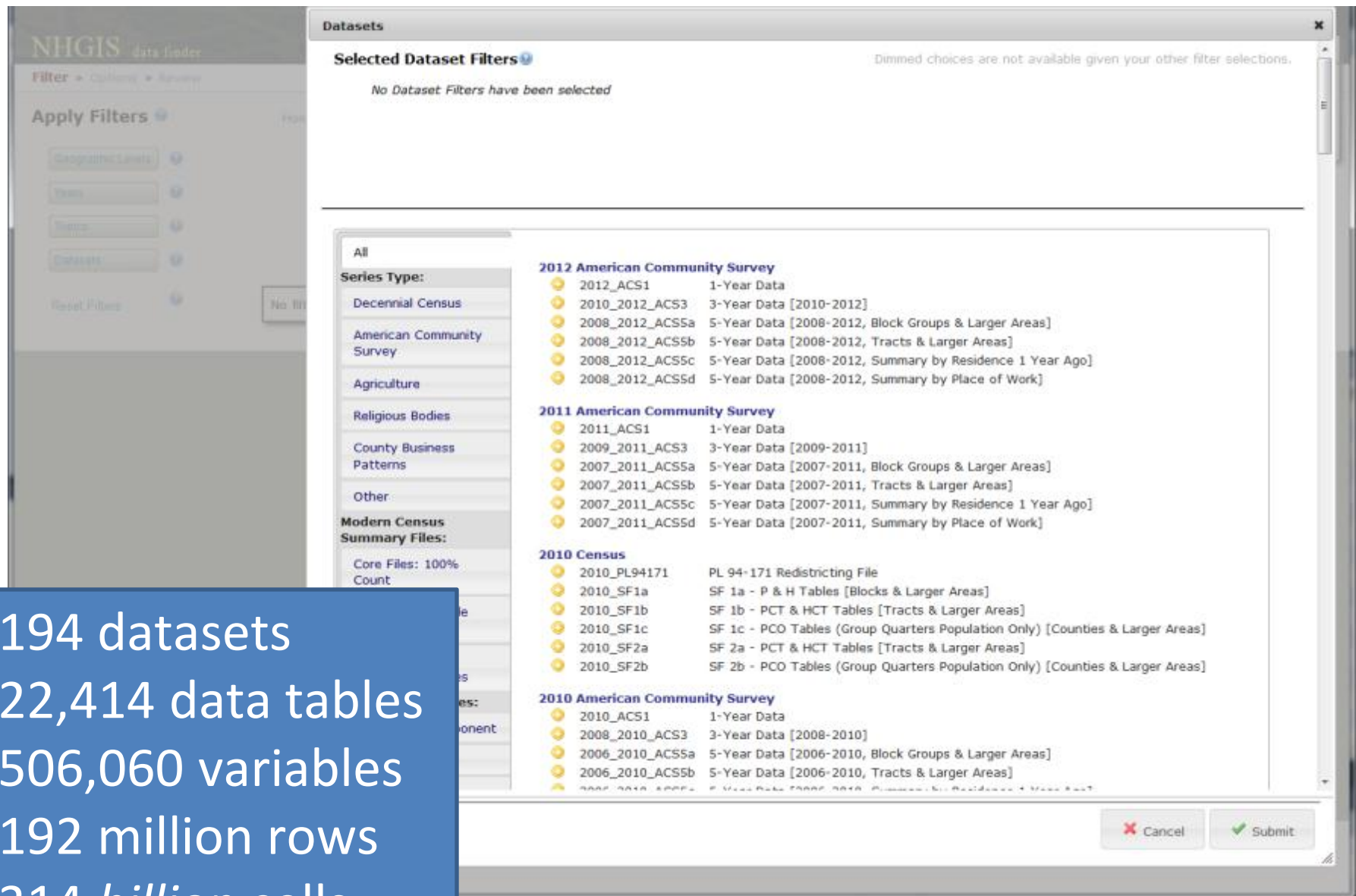
# More NHGIS Data

County Business Patterns, 1970 – 2002

Many other unique datasets

- FDIC bank deposit data, 1920 – 1936
- U.S. unemployment data, 1937
- Censuses of Religious Bodies, Farm Real Estate Values and more

194 datasets  
22,414 data tables  
506,060 variables  
192 million rows  
314 *billion* cells



# GIS Shapefiles

- GIS boundary files, 1790-2012
  - Nation, State and County boundaries, 1790-2012
  - Census Tract boundaries, 1910-2012
  - Metropolitan region boundaries, 1950-2012
  - All other released Census geographies, 1980-2012
    - Blocks
    - Block Groups
    - School Districts
    - Urban Areas
    - Places
    - County Subdivisions
    - Congressional Districts
    - American Indian Areas
    - and several more

# Time Series Tables

- Harmonized census tables
- Multiple years of data in one file
- 197 tables

**NHGIS data finder**

**Filter** » Options » Review

**Apply Filters** [How to use the data finder \(pdf\)](#)

Geographic Levels [?](#)

Years [?](#) ☒ 2010

Topics [?](#) ☒ INCLUDES Race

Datasets [?](#)

Reset Filters [?](#)

**Data Cart** [clear](#)

0 source tables  
0 time series tables  
0 GIS boundary files

Show selections [Continue](#) [?](#)

**Select Data** [?](#)

1642 source tables 65 time series tables 176 GIS boundary files

Table Name	Years	Geographic Levels
Persons by Race [5*]	1970, 1980, 1990, 2000, 2010	state, county, cty_sub, tract, place
Persons by Detailed Race [10*]	1970, 1980, 1990, 2000, 2010	state, county, cty_sub, tract, place
Persons by Race [6*]	1980, 1990, 2000, 2010	nation, region, division, state, county, cty_sub, tract, place
Persons by Detailed Race [17*]	1980, 1990, 2000, 2010	nation, region, division, state, county, cty_sub, tract, place
Persons by Race [7*]	1990, 2000, 2010	nation, region, division, state, county, cty_sub, tract, place
Persons Who Are American Indian and Alaska Native* by Tribe [29]	1990, 2000, 2010	nation, region, division, state, county, cty_sub

# MICRODATA

# Microdata versus Summary Data

## Microdata

- Shows full range of responses for individuals
- Enable custom tables and sophisticated analyses
- Suppression: geography, truncation, and item level suppression

## Summary Data

- Premade or published tables of aggregate characteristics
- Enable examination of small geographic areas
- Suppression: limited content, grouped intervals, and cell suppression

# Microdata

Relation  
to head

Marital  
status

Literacy

Occupation

00009800010010021000079122022080211203900999009900000000001			
00009800010020012000064222022080221203100999009900000000002			
00009800010030004100023110022080221201000832002690000001202			
00009800010040004100013210022080121203300999009900000000002			
00009880010010021000035121022080223112300723005270000000302			
00009880010020012000041221022080221203100999009900000000002			
00009880010030003000017210022080222122300998009900000000002			
00009880010040003000016210022080222123100999009900000000002			
00009880010050004910019221022080222123100999009900000000002			
0000988001006000491002712102208022312100012200011000009602			
0000988001007000410000021002208000000000099900990999999992			
0000988001008000491002211002208022311100061100011000002801			
0001005001001002100004512102208012212100061100011000002002			
00010050010020012000034221022080122213100999009900000000002			
00010050010030003000012210022080122123300999009900000000002			
00010130010010021000065122022070221203430999009900000000002			
00010130010020012000065222022080221203100999009900000000002			
00010130010030003000033235022080222213100999009900000000002			
00010130010040004100015210022080122213300999009900000000002			
00010130010050004100013210022080122123300999009900000000002			
0001013001006000410000211002208000000000099900990999999992			



# What is IPUMS?

**Integrated** - consistent codes, labels, and documentation

**Public Use** - anonymized, downloadable

**Microdata** - individual-level

**Series** - pooled data over time and place

# INTEGRATION

# Variable Harmonization

## Marital Status

### Bangladesh 2011

1 = Unmarried

2 = Married

3 = Widowed

4 = Divorced/separated

### Mexico 1970

1 = Married, civil & relig

2 = Married, civil

3 = Married, religious

4 = Consensual union

5 = Widowed

6 = Divorced

7 = Separated

8 = Single

### Kenya 1999

1 = Never married

2 = Monogamous

3 = Polygamous

4 = Widowed

5 = Divorced

6 = Separated

# Translation Table

## Input

**Bangladesh  
2011**

**Mexico  
1970**

**Kenya  
1999**

1 = Unmarried	1 = Married, civil & relig	1 = Never married
2 = Married	2 = Married, civil	2 = Monogamous
3 = Widowed	3 = Married, religious	3 = Polygamous
4 = Divrc or separated	4 = Consensual union	4 = Widowed
	5 = Widowed	5 = Divorced
	6 = Divorced	6 = Separated
	7 = Separated	
	8 = Single	

# Translation Table

## Harmonized

**Code**      **Label**

1	Single
2	Married or in union
2	▶
2	▶
2	▶
2	▶
2	▶
2	▶
2	▶
3	Divorced or separated
3	
3	
4	Widowed

# Integrating Documentation

- Sample Descriptions
- Variable Descriptions
  - Availability by Sample
  - Universes
  - Comparability
  - Allocation and Imputation Flags
  - Questions and Instructions to Respondents
  - Instructions to Enumerators

# Additional Enhancements to the Data

- Integrated Variables
- Geographic Areas
- Consistent industrial and occupation coding schemes
- Constructed family interrelationship variables

# www.ipums.org/usa

MINNESOTA POPULATION CENTER, UNIVERSITY OF MINNESOTA

# IPUMS<sup>USA</sup>

Home Select Data FAQ Help Login Data Cart

▼ DATA

Browse and Select Data  
Download or Revise Extracts  
Analyze Data Online  
IPUMS Registration

▼ DOCUMENTATION

Variables  
Samples  
User's Guide  
Geographic Tools  
FAQ

▼ RESOURCES

Enumeration Forms  
Published Census Volumes  
Errata and Revisions

▼ RESEARCH

Citation and Use  
Bibliography  
Related Sites

▼ CONTACT US

Help  
User Forum  
IPUMS Staff  
How to Help  
Donate to IPUMS

## Integrated Public Use Microdata Series

census microdata for social and economic research

IPUMS-USA is a project dedicated to collecting and distributing United States census data. Its goals are to:

- Collect and preserve data and documentation
- Harmonize data
- Disseminate the data absolutely free!

Use it for GOOD -- never for EVIL

▼ IPUMS IN THE NEWS

IPUMS is for "Data Addicts"  
1940 Census Archives  
IPUMS and the Civil War  
NYT Interactive Map

▼ IPUMS-USA DATA UPDATES

NEW! 2012 1-year ACS and PRCS  
1850 Complete Count data  
1930 1% and 5% data

▼ OTHER IPUMS DATA

Complete Count Data  
Linked Samples 1850-1900  
1850-1860 Slave Samples

▼ OTHER MPC PROJECTS

IPUMS-CPS  
IPUMS-International  
NHGIS  
NAPP  
IHIS  
ATUS-X



Microdata Data:

**IPUMS-USA**

# IPUMS-USA

- Database includes public use microdata samples:
  - U.S. decennial censuses (1850-2000)
  - Complete-count dataset for 1880
  - Linked Samples 1850 – 1930
  - Samples from Puerto Rico (1910-2012)
  - American Community Survey (2000-2012)
- The first MPC data project
- Most widely used database ~ 30,000 users

# The American Community Survey

- Replaced the long form of the Decennial Census
  - Demonstration stage: 2000 to 2004
  - Full implementation 2005, group quarters added 2006
- Rolling sample design

Microdata samples:

- Full survey responses for 1% of US population
- Yearly samples, multi-year samples

# Census and ACS Household-Level Variables

- Geography
- Group quarters
- Housing characteristics
- Household economic characteristics
- Household appliances

# Census and ACS Individual-Level Variables

- Age
- Race
- Marital Status
- Birthplace
- Nationality
- Family Interrelationship
- Education
- Work
- Income
- Disability
- Migration Status
- Commuting
- Health Insurance

# Geography Limitations

- No confidentiality restrictions for samples prior to 1950 – no geographic limitation
- Samples from 1950-1970
  - Inconsistent geographic identifiers
- Recent samples:
  - State
  - Some Metropolitan Areas
  - County Groups
  - Public Use Microdata Areas (PUMAs)

# What are PUMAs?

- Public Use Microdata Areas (PUMAs)
- Comprised of approximately 100,000 persons
- Boundaries do not always align with jurisdictional boundaries
- Detailed contents and maps available
- GIS shape files for PUMAs available

# Sample Expansion

- 1850 full count dataset – in collaboration with the Church of Jesus Christ of Latter-day Saints
- 1960 data recovery and 5% sample creation
- 1940 full count dataset
- 1790-1930 full count data



Obtaining Data:

# DATA EXTRACTION

# Online Extraction System

- Users create custom data files
  - Pick any samples of interest
  - Pick any variables of interest
- Creates custom syntax for reading the data files into SPSS, Stata, SAS, and CSV
  - Labels variables and values within the data
- Codebook available
- Record of extract is preserved on user account

# Online Extraction System

- Additional Features:
  - Case Selection
  - Attach Variables
  - Custom Sample Size
- Learn one data set, learn how to use all MPC microdata
  - Documentation systems
  - Extract systems
  - Data files produced

# Online Analysis System

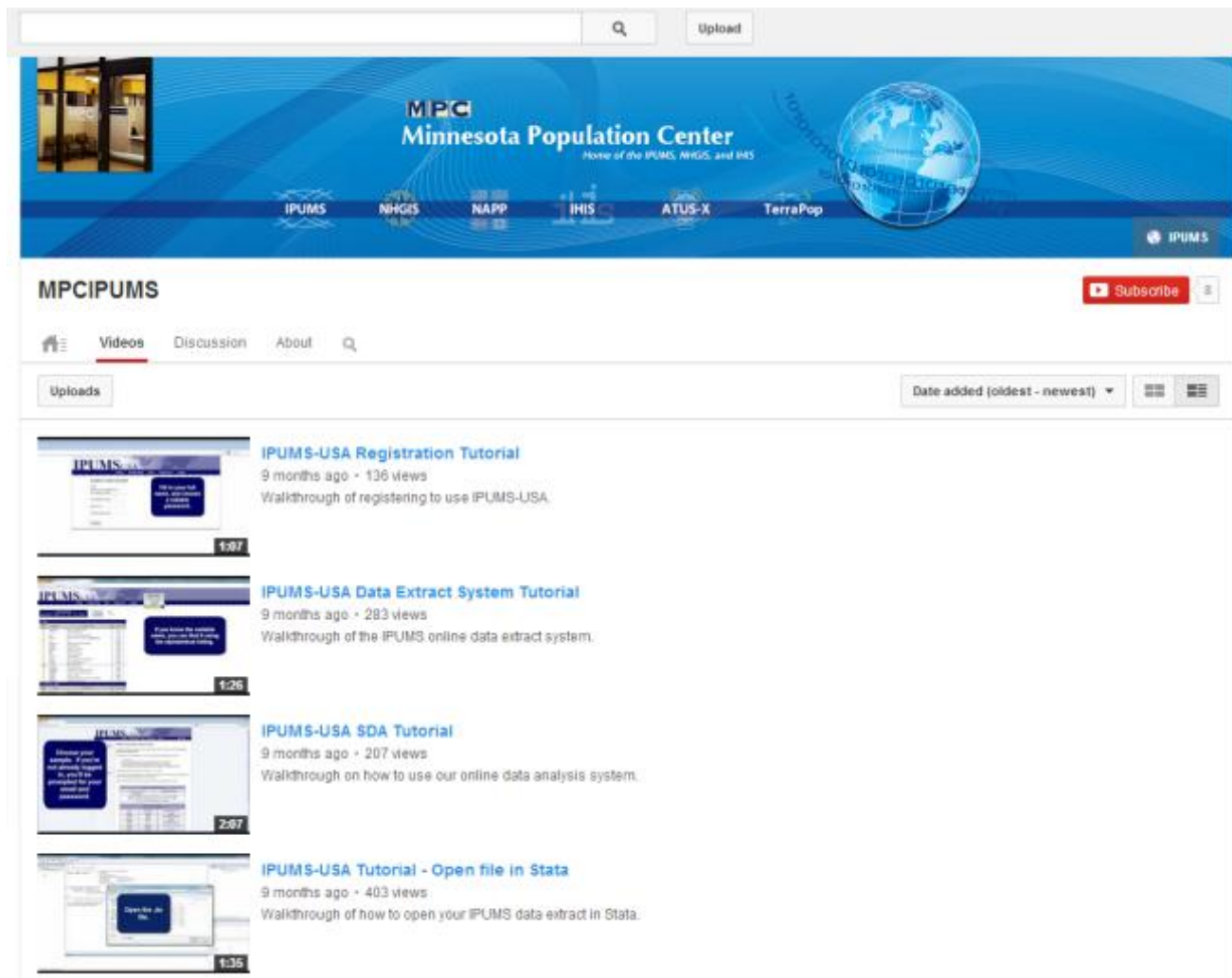
- High-speed tabulation software developed at UC-Berkeley
- Allows for analysis of microdata without statistical package
- All analysis performed online
- Can analyze multiple years of data
- Help guides on webpage

Obtaining Data:

# USING MPC DATA

# Video Tutorials

<http://www.youtube.com/user/MPCIPUMS>



The screenshot displays the MPCIPUMS YouTube channel page. The header features the MPC Minnesota Population Center logo and a navigation bar with links to IPUMS, NHGIS, NAPP, IHIS, ATUS-X, and TerraPop. Below the header, the channel name 'MPCIPUMS' is shown with a 'Subscribe' button. The 'Videos' tab is selected, showing a list of uploads. The videos are sorted by 'Date added (oldest - newest)' and are displayed in a grid format. Each video entry includes a thumbnail, title, upload date, view count, and a brief description.

Video Title	Upload Date	Views	Description
IPUMS-USA Registration Tutorial	9 months ago	136 views	Walkthrough of registering to use IPUMS-USA.
IPUMS-USA Data Extract System Tutorial	9 months ago	283 views	Walkthrough of the IPUMS online data extract system.
IPUMS-USA SDA Tutorial	9 months ago	207 views	Walkthrough on how to use our online data analysis system.
IPUMS-USA Tutorial - Open file in Stata	9 months ago	403 views	Walkthrough of how to open your IPUMS data extract in Stata.

# Why people use MPC data

- Comprehensive online documentation
- Integration makes analyzing change over time possible
- Data analysis system allows you access the data and analyze it online
- All of the data are available for **free** online
- User support is available by e-mail to help you as needed

# Using MPC data

- Cite the data – check homepages for citation information
- Enter your work in our bibliography
  - IPUMS Research Award
- Data Questions
  - User forum (<http://answers.popdata.org/>)
  - Email IPUMS user support (ipums@umn.edu)
- Data Errors
  - Get an IPUMS mug!



# Using the extract system and documentation

- [Extract System](#)
- [Variable Description](#)
- [Online Analysis](#)

# Analyzing IPUMS Data

- Register for IPUMS-USA
- Three research questions
- Hands-on workshop exercise
- Further workshop materials:

<https://www.pop.umn.edu/data-user-resources/data-support>

# Example research questions, methods, answers

- Research Question: How has the racial composition of California changed since 1980?
  - How was race measured in each year?
  - What are options for comparing race over time?

# Example research questions, methods, answers

- Research Question: In Sacramento in 2012, how many people work in government?
  - What variable captures government employees?
  - What are the pumas or metro area code for Sacramento?
  - Which respondents need to be excluded from this analysis?

# Example research questions, methods, answers

- Research Question: How do people in CA commute to work?
  - What information does the ACS have on commuting?
  - Which respondents are asked about commuting?
    - What variables do we need to restrict the sample to these respondents?
  - What is the state code for CA?