# Texas Census Research Data Center: Opportunities for Research Using Non-Public Data

Mark Fossett
Director, Texas Census Research Data Center
College of Liberal Arts, Texas A&M University
College Station, Texas

State Data Center/Business and Industry Data Center
Annual SDC/BIDC Conference
Austin, Texas
May 23, 2012

# **Good News from NSF & the US Census Bureau**

Texas A&M University has won US Census Bureau approval and National Science Foundation funding to establish the Texas Census Research Data Center (TXCRDC) at Texas A&M University in College Station.

This research facility will be one of only about twelve (12) such centers in the nation. It is the only CRDC in a region extending 800 miles in any direction from Texas A&M University.

Texas A&M University initiated and led the consortium proposal effort. Four institutions in all are involved additionally including:

- The University of Texas at Austin
- Baylor University
- The Texas A&M University System

The TXCRDC Consortium invites interested institutions to join.

Active conversations are ongoing with several interested institutions (e.g., Rice, UTSA, UTMB-Galveston, Texas State)

# Thanks for Help in Winning NSF/Census Approval

#### **US Census Bureau**

**Dr. Robert Groves, Director** 

Dr. Ron Jarmin, Associate Director, Research and Methodology

#### **TAMU Division of Research**

Dr. Jeffrey R. Seemann, Vice President for Research

Dr. Charles A. Johnson, Senior Assoc. Vice President for Research

### **TAMU College of Liberal Arts**

Dr. José Luis Bermúdez, Dean

Prior Deans Dr. Ben M. Crouch and Dr. Charles A. Johnson

#### **Consortium Institutions**

Texas A&M, Texas A&M University System, UT-Austin, and Baylor

Four Dozen Researchers from Eight Institutions & Ten Disciplines Texas A&M, UT-Austin, Baylor, TAMU Health Science Center, UT-San Antonio, Sam Houston State, Houston, UTMB-Galveston

# What is a Census Research Data Center (CRDC)?

Census Research Data Centers (CRDCs or RDCs) are unique research facilities based on contractual relationships — Joint Statistical Partnerships (JSP) — between the Census Bureau and leading research institutions.

In essence, the RDC is a census "outpost" at TAMU.

The RDC has a computing lab that is linked with the Census Bureau computing network via a secure network connection.

The national RDC network is managed by the Center for Economic Studies (CES) at Census Headquarters which hosts non-public federal data sets on their servers.

RDC computing labs are managed by an on-site federal (Census Bureau) employee – the RDC Administrator – who serves as a liaison with the research community.

# The Claim to Fame for CRDCs

The contractual relationship (JSP) with the Census Bureau makes it possible for "qualified" researchers who have approved projects to access non-public data from within the RDC facility.

Based on this, RDCs can provide researchers access to a wealth of non-public data maintained by the Census Bureau and other agencies (e.g., NCHS health data sets) in the federal statistical system.

Note that researcher access is strictly limited.

- Researchers must be "qualified" based on having "special sworn status" which makes them an unpaid Census Bureau researcher.
- Research projects must be approved under a rigorous Census Bureau review process.
- Projects must provide direct benefits to the Census Bureau.
- Data access is limited to the minimum required to accomplish the approved project.

# Why are RDCs Needed?

Federal Laws and Regulations Place Stringent Confidentiality Restrictions on the Detailed Data in Federal Statistical System

- It is illegal to disclose confidential data
- The restrictions are grounded in important ethical and practical considerations and are unlikely to change

Federal Employees Can Access Non-public Data (Under Strict Rules)

RDCs Extend This Opportunity to Researchers by Giving them Special Sworn Status (SSS)

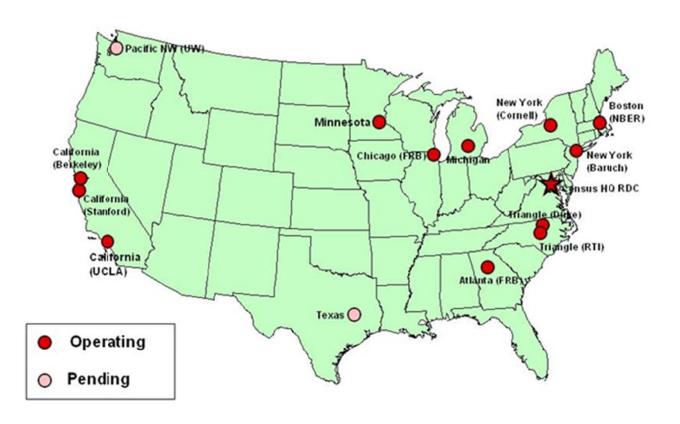
- This in effect "deputizes" the researcher as an unpaid census bureau employee
- SSS is conferred under a review process in which several specific conditions must be met (no exceptions)

# **The National RDC Network**

```
Existing RDCs
 Atlanta – Federal Reserve Bank (opened Fall 2011)
 Center for Economic Studies (CES) at Census HQ (Suitland, MD)
 Boston – NBER (Cambridge)
 Chicago – Federal Reserve Bank
 University of Michigan
 Duke University / Research Triangle
 UC-Berkeley / Stanford
 UCLA
 Cornell / Baruch (New York City)
 University of Minnesota
Coming Soon
 TXCRDC – opening Fall 2012
 University of Washington – also opening Fall 2012
```

# **The National RDC Network**

# **Census Research Data Centers**



# What Will the TXRDC Ultimately Offer and When?

#### What will the TXCRDC Offer?

- The TXCRDC will provide qualifying researchers access to nonpublic data in the federal statistical system.
- This access is only possible on site from within the physical facility of the <u>secure computing lab</u> housed in the TXCRDC.

### Some Secondary, But Nice, Things

- Free on-site parking
- Easy driving distance Dallas/San Antonio/Austin/Houston
- Easy airport access
- Relatively inexpensive accommodations for extended stays

# When will the Facility be Open for Business?

- Establishing RDC involves a complex collaboration with the US Census Bureau. It takes 12-15 months.
- Our goal is to have the secure lab open for researchers to use in Fall 2012 shooting for as early as possible.

### What Research Communities Do RDCs Serve Best?

#### **RDCs Serve Well ...**

- Basic Science Research
  - discipline-based substantive research
  - statistical and methodological research
- Planning and Policy Science Research
  - research evaluating program effectiveness
  - research assessing program impacts
- Research programs with long time horizons and "high-level" research products (e.g., results of regression analyses)

#### **RDCs Serve Less Well ...**

- Research with short time horizons
- Detailed descriptive research
- Research generating special tabulations, maps, and lists
- Projects to produce data products for public use or distribution

# Fall 2012 is a Long Time Away, But It's Not

Researchers seeking to use the TXCRDC secure computing lab in Fall 2012 should begin developing proposals by Spring 2012.

Only researchers with projects that are officially approved by the Census Bureau can use the secure computing lab.

The process of getting projects approved for research conducted in RDCs can be elaborate and time consuming.

Starting now gives interested researchers a decent chance of obtaining Census Project approval with time to spare.

Note some advantages of getting projects approved early.

- The research could begin in other existing RDCs.
- The researcher may gain priority access to the benefits of Virtual RDC resources (more on this below).
- The researcher has better chances of securing external funding to support the research. (Official Census approval can help in funding review.)

# **Advantages of RDC Access to Researchers**

Access to microdata not available publicly

Access to more detailed information than that available in public use data sets

Access to full population counts or full samples

Access to original variables before processing (e.g., income not top-coded)

Access to finer geographic detail

Ability to link with external data (e.g., based on geography)

Ability to merge multiple *internal* data sets via non-public link keys (personal identification codes)

# **RDC-Based Research is Common in Many Fields**

### Leading researchers in many fields conduct research in RDCs

- Business, Trade, Finance, and Management
- Crime and Crime Victimization
- Demography, Population Distributions and Trends, Migration, and Immigration
- Economics, Labor Markets, Entrepreneurship, Employment and Industry
- Education and Education Policy
- Hazard Reduction, Environmental Impact Assessment,
   Pollution Abatement
- Health and Well-Being, Health Insurance, and Health Policy
- Housing, Housing Markets, and Residential Patterns
- Poverty and Social Welfare Policy
- Transportation Analysis and Planning
- Urban and Regional Economics and Planning

# **Examples of Available Data Sets**

- 1. AHRQX MEPS Extract (2005)
- 2. American Community Survey (1996-2008)
- 3. American Housing Survey (1993-2005)
- 4. Annual Capital Expenditures Survey (1993-2005)
- 5. Annual Retail Trade Survey (1999-2006)
- **6. Annual Survey of Manufactures (many years)**
- 7. Auxiliary Establishment ES9200 (1977; 1982; 1987; 1992; 1997; 2002)
- 8. Business Expenditures Survey (Every 5 years, 1992-2002)
- 9. Business Register Bridge (2001-2004)
- 10. Census of Construction Industries (Every 5 years, 1972-2002)
- 11. Census of Finance, Insurance, and Real Estate (Every 5 years, 1992-2002)
- **12.** Census of Manufactures (1963; Every 5 years, 1967-2007)
- 13. Census of Mining (1987; 1992; 1997; 2002)
- 14. Census of Retail Trade (Every 5 years, 1977-2002)
- 15. Census of Services (Every 5 years, 1977-2002)
- 16. Census of Transportation, Communications, and Utilities (many years)
- 17. Census of Wholesale Trade (Every 5 years, 1977-2002)
- 18. Commodity Flow Survey (1993; 1997; 2002)
- 19. Compustat-SSEL Bridge (2005)
- 20. Current Industrial Reports (1993-2008)

# **Examples of Available Data Sets – Continued**

- 21. Current Population Survey March Supplement (1967-2008)
- 22. Decennial Census Long Form Sample (Every 10 years, 1970-2000)
- 23. Decennial Employer-Employee Database (Every 10 years, 1990-2000)
- 24. Economic Census of Puerto Rico (1992 and 1997)
- 25. Employer Characteristics File (1989-2004)
- 26. Employment History Files (1985-2004)
- 27. Enterprise Summary Report ES9100 (large company) (many years)
- 28. Exporter Database (1987; 1992; 1996-2006)
- 29. Foreign Trade Data Export (1992-2007)
- **30. Foreign Trade Data Import (1992-2007)**
- 31. Form 5500 Bridge File (1992-2006)
- 32. Geocoded Address List (2004)
- 33. Individual Characteristics File (1985-2004)
- 34. Integrated Longitudinal Business Database (Every 5 years, 1977-1992)
- 35. Longitudinal Business Database (1976-2005)
- **36. Manufacturing Energy Consumption Survey (many years)**
- 37. Medical Expenditure Panel Survey Insurance Component (1996-2008)
- 38. National Center for Health Statistics Data Extract (2005)
- 39. National Employer Survey (1994; 1996; 1997; 1998)
- 40. National Longitudinal Survey (many years)

# **Examples of Available Data Sets – Continued**

- 41. Ownership Change Database (1963; Every 5 years, 1967-1997)
- 42. Quarterly Financial Report (1977-2008)
- 43. Quarterly Survey of Plant Capacity Utilization (2007-2008)
- 44. Quarterly Workforce Indicators (2004)
- 45. Services Annual Survey (2002-2006)
- 46. Standard Statistical Estab. List non Name and Address File (1974-2005)
- 46. Standard Statistical Estab. Listing Name & Address File (1974-2007)
- **47.** Survey of Business Owners (many years)
- 48. Survey of Income and Program Participation (SIPP) Panels (many years)
- 49. Survey of Income and Program Participation Longitudinal (many years)
- 50. Survey of Industrial Research and Development (1972-2007)
- 51. Survey of Manufacturing Technology (1988; 1991; 1993)
- 52. Survey of Plant Capacity Utilization (1974-2005)
- 53. Survey of Pollution Abatement Costs and Expenditures (many years)
- 54. Unit-to-Worker (1990-2004)

### More Data Sets are Being Added Every Year

### See the TXCRDC website for detailed compilations

- Compilations by subject area
- Links to data sets and other compilations

# **Health Data Sets?**

Health-related research in CRDCs has recently grown dramatically. Currently about 30% of projects in CRDC are health related, up from under 10% five years ago.

The website for the National Center for Health Statistics (NCHS) RDC provides a description of the restricted variables that are available in CRDCs by NCHS data set.

http://www.cdc.gov/rdc/B1DataType/Dt100.htm

The website for the US Department of Health and Human Services Agency for Healthcare Research and Quality (AHRQ) RDC provides a description of the restricted variables that are available in CRDCs by AHRQ data set at the following link.

http://www.meps.ahrq.gov/mepsweb/data stats/onsite datacenter.jsp

# What are the Trends in RDC Activity?

### The Number of RDCs in the National Network is Increasing

- About a dozen RDCs are now operating
- Adding about 1 new RDC every two years
- Organization under of the Consortium Funding Model makes
   RDCs more feasible and cost effective

### **Usage is Increasing**

- Data set availability is increasing; making RDC use more attractive
- Protocols are being improved and streamlined
- In many fields, RDC access is crucial to leading research

#### **Patterns and Trends**

- Traditionally, most research was in economics/business
- In recent years, health research has risen dramatically
- Demographic research and policy research has held steady

# **Benefits to the Census Bureau**

### **Many Potential Benefits**

- Ability to benefit from consulting with leading researchers, effectively increasing the pool of expertise they draw on.
- Identifying measurement problems and potential solutions
- Deriving new estimates from existing data
- Expanding measurement capabilities
- Improving documentation

#### The Issue of Benefits is IMPORTANT

- The National RDC program is justified based on the Benefits it provides to the Census Bureau
- The RDC program is NOT justified in terms of benefits to the research community

Accordingly, Projects are Justified in Terms of Census Benefit

# Role of the TXCRDC Director and Staff

### **During Year 1**

Implement the Center (finalize and execute contracts with US Census Bureau, implement the secure computing lab to federal requirements, recruit RDC Administrator, etc.)

Promote awareness, share information, encourage proposal development

#### **Down the Road**

**Support researchers and active projects** 

**Expand and nurture the consortium** 

Maintain and upgrade the secure lab

**Explore options for implementing "Nodes" or "Branches"** 

# **Notes on Access and Usage for Approved Projects**

#### **Costs for Researchers at Consortium Institutions**

- Access will be at no cost or at very low cost in comparison to RDC charges for non-consortium users
- Scheduling will be prioritized, if needed, on several criteria.
   Research involving external funding support will receive highest priority. Not likely to be an issue in the near term.

#### **Costs for Non-Consortium Researchers**

- Researcher access will be on a per project cost schedule in keeping with practices in the national RDC network
- Individual arrangements may be negotiated.
- Collaborating with researchers at consortium institutions will bring more favorable terms.

The Intent is for Consortium Membership to be a Cost Effective way for Universities/Agencies to Giver Researchers Access

# **Important Practical Issues to Consider**

### **Obtaining Census Project Approval**

- Census Bureau approval of projects is required
- The criteria for Census Bureau approval is not on scientific merit (although that is expected). Approval is based on clear "benefits to the Census".

### The RDC Computing Environment – Software and Hardware

- Computing is conducted in the secure lab
- The data are on Census Bureau servers
- Workstation hardware and software follow Census Bureau specifications

### **Getting Research Products Out is Non-Trivial**

- There is rigorous mandatory review on confidentiality criteria;
   NOTHING leaves the secure lab without approval
- It is crucial to plan ahead for getting project approval and for getting project output (e.g., regression results) out of the lab

# **Near Term RDC Logistics**

# **Implementing the Full RDC**

Acquire space and set up the "non-secure" areas of the center; recruit staff, phones, mailing address, etc.

Build out website and other information resources; develop user group contacts and information

**Negotiate and implement the TAMU-Census Bureau contract** 

**Hire and Train the US Census Employee** 

Build the secure facility to federal specifications (including making necessary renovations)

Install the hardware and software for the secure lab

Sad But True; It Really Does Take 12-15 Months?

Implementing the RDC is a State-Federal collaboration involving hiring a federal (US Census Bureau) employee under federal guidelines for secure facilities

# Things Coming Soon of Interest to Researchers

### **News updates from TXCRDC**

- Responding to information requests
- Organizing user communities
- Presentations to departments and working groups
- Meeting with individual researchers

# **Workshops on RDC Promotion**

- The basics of conducting research in RDCs Ron Jarmin
- Sessions on research in different fields
- Sessions on methods and software

**Seed Grant Program for Researchers at Consortium Institutions** 

**Build the Consortium to Assure RDC Stability/Viability** 

# **Short- and Long-Term Projected Time Lines**

#### Fall 2011

**Implement non-secure portion of Center** 

Promote awareness, disseminate information, encourage proposal development, seed grant program

**Spring – Summer 2012** 

RDC Workshops, Proposal Development, Implement Secure Computing Lab, Hire Census Bureau Employee

Fall 2012

**Secure Computing Lab "Opens for Business"** 

Fall 2014 and Fall 2016

2014 - NSF 3-year funding ends

2014 - UT-Austin and Baylor commitments reviewed

2016 - TAMU and TAMUS commitment reviewed

### I Want In! What Do I Do and When?

Contact the TXCRDC to learn about next steps

**Attend TXCRDC Workshops** 

**Plan for Working on Two Proposals** 

**Census Bureau Project Approval** 

- Absolutely crucial, do not take it for granted
- Involves non-standard review criteria

**Project proposals to external funding agencies** 

- Needed to support research assistants
- Needed for priority access to workstations
- The Center

#### When?

**As Soon As Possible** 

Census Bureau project approval needed for Virtual RCD access

# **How Will the TXCRDC Help?**

**Researching information relevant to researcher interests** 

- Data availability; available software
- Project budgets and timeline
- Contact information

**Guidance on Proposal Development Process** 

Navigating the Census Project Approval Process

**Seed Grant Program for Proposal Development** 

**Assistance and Training for using the Secure Lab** 

**Scheduling On-Site Project and Workgroup Space** 

# **The Independent Census Review Process**

# There are Up to Three Stages of Census Project Review

# **Development Stage Review at the TXCRDC**

Researchers consult with TXCRDC staff to learn about the review process and the relevant criteria to receive preliminary feedback on proposals.

#### **Formal Census Review**

Typically this takes about 90 days and may involve an iterative process.

### **Other Agency Review**

- If non-census data are involved, additional agencies will be involved as appropriate.
- Relevant for IRS, Social Security, NCHS, AHRQ
- May involve additional fees and protocols

# **Tips for a Good RDC Experience**

### Be Realistic. Plan Ahead, and Plan Carefully

- Start the process early do not underestimate the longer time line for projects
- Use standard data sets if time-constrained

# Think Strategically About RDC Projects

- When feasible, write broad proposals geared towards multiple papers over long project life span
- Use proposal development as research time
- Use the Virtual RDC to be efficient in developing analyses

#### **Learn About the Research Environment**

Attend RDC workshops; Read on-line documentation and advice; Consult with RDC researchers; Examine CES research annuals and Working Papers

**Remember** the RDC System is justified by Benefits to the Census

# **Census Programs Available to RDC Researchers**

**CES Mentorship Program Ph.D. Students** 

**Census Bureau Dissertation Fellowship (\$50K)** 

http://www.census.gov/srd/www/DissertationFellowshipTopics.pdf

**ASA/NSF/Census Bureau Fellowship Program (\$50K)** 

http://www.census.gov/srd/www/fellweb.html

# The Cornell Virtual RDC

The NSF Funded Cornell Virtual RDC is a potentially valuable resource to researchers

- The Virtual RDC provides researchers convenient access to "synthetic data" in a replicated RDC computing environment.
- The synthetic data are facsimiles of the non-public data sets researchers plan to use in their projects.

#### **Benefits**

- Internet based; does not require travel to a physical RDC.
- Facilitates proposals and projects by potential RDC users.
- Helps train new users in the RDC operating system environment and the data and software available in the "real" RDC

Service is Prioritized to Researchers with Census-Approved Projects

Further Information: <a href="http://www.vrdc.cornell.edu/news/">http://www.vrdc.cornell.edu/news/</a>

# **Key Links and Sources of Information**

```
Information about Data, Data Sets, and New Developments
 Center for Economic Studies (CES, US Census)
   http://www.census.gov/ces/
  National Center for Health Statistics (NCHS)
   http://www.cdc.gov/rdc/
 Agency for Healthcare Research and Quality (AHRQ)
   http://www.meps.ahrq.gov/mepsweb/data_stats/
   onsite datacenter.jsp
Developing Projects
  TXCRDC Director — Mark Fossett (<u>m-fossett@tamu.edu</u>)
Key Folks at the Center for Economic Studies (CES) & NCHS
  Ron Jarmin (Ron.S.Jarmin@census.gov)
 Lucia Foster (<u>Lucia.S.Foster@census.gov</u>)
```

Peter Meyer (prm7@cdc.gov)

# **Overview and Conclusions**

# The TXCRDC is an Important Opportunity

- Available data will support a variety of research agendas
- The data in RDCs will be increasingly important for being at the forefront of research in many fields
- The resource can attract top researchers to TAMU

### The Location and the Consortium Model are Major Benefits

- Currently Texas researchers must travel long distances to use an RDC and incur substantial time and travel costs
- Consortium membership provides cost advantages; "Pay as you go" RDC access typically costs \$15K-\$20K annually per project

#### **Benefits More than Just Senior Researchers**

Assistant and associate-level researchers can advance careers;
 Graduate students, recent graduates, and postdocs can benefit

### **Take Full Advantage of this Opportunity!**

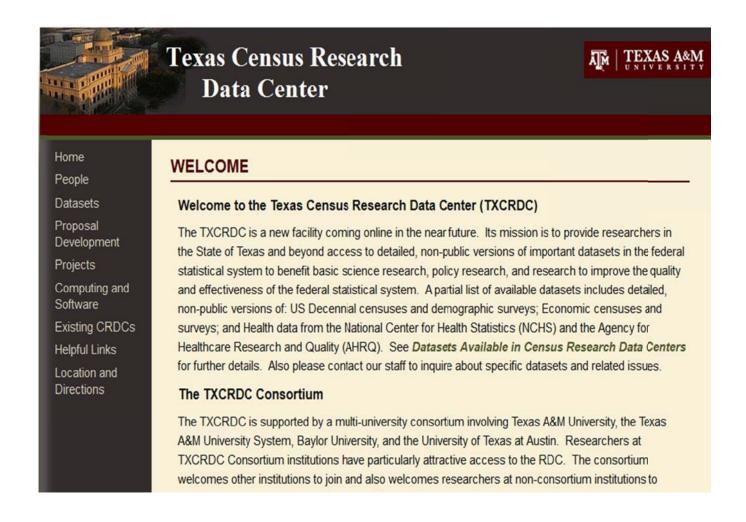
# **TXCRDC Contact Information**

Mark Fossett, Director
Texas Census Research Data Center
TAMU 2403
101 Donald L. Houston Building
200 Discovery Drive
College Station, Texas 77843-2403
txcrdc@tamu.edu
979-845-5618

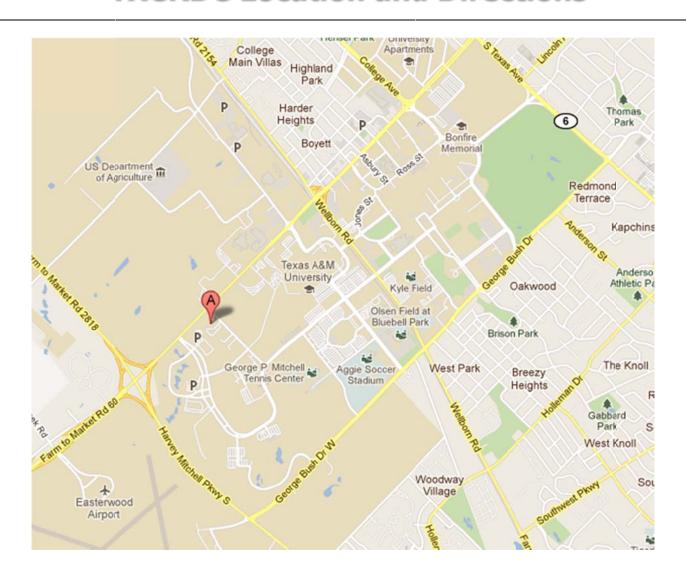
TXCRDC website: <a href="http://txcrdc.tamu.edu">http://txcrdc.tamu.edu</a>

# **TXCRDC Web Site**

TXCRDC website: <a href="http://txcrdc.tamu.edu">http://txcrdc.tamu.edu</a>



# **TXCRDC Location and Directions**



### TXCRDC - Donald L. Houston Bldg, 200 Discovery Dr.



### **TXCRDC – Houston Building Main Entrance**



## **TXCRDC – Convenient Parking (Free!)**



## **TXCRDC – Houston Bldg Main Entrance**



### **TXCRDC – Houston Bldg Entry Area**



### TXCRDC – Welcome to the TXCRDC (Room 101)



#### **End**

Mark Fossett
Director, Texas Census Research Data Center
College of Liberal Arts, Texas A&M University
College Station, Texas

Texas A&M University College Station, Texas November 15, 2011

Mark Fossett
Director, Texas Census Research Data Center
College of Liberal Arts, Texas A&M University
College Station, Texas

Baylor University
Waco, Texas
November 15, 2011

Mark Fossett
Director, Texas Census Research Data Center
College of Liberal Arts, Texas A&M University
College Station, Texas

Population Research Center
The University of Texas
Austin, Texas
November 16, 2011

Mark Fossett
Director, Texas Census Research Data Center
College of Liberal Arts, Texas A&M University
College Station, Texas

Department of Demography
College of Public Policy
The University of Texas – San Antonio
San Antonio, Texas
November 16, 2011

Mark Fossett
Director, Texas Census Research Data Center
College of Liberal Arts, Texas A&M University
College Station, Texas

Kinder Institute for Urban Research and
Department of Sociology
Rice University
Houston, Texas
November 17, 2011