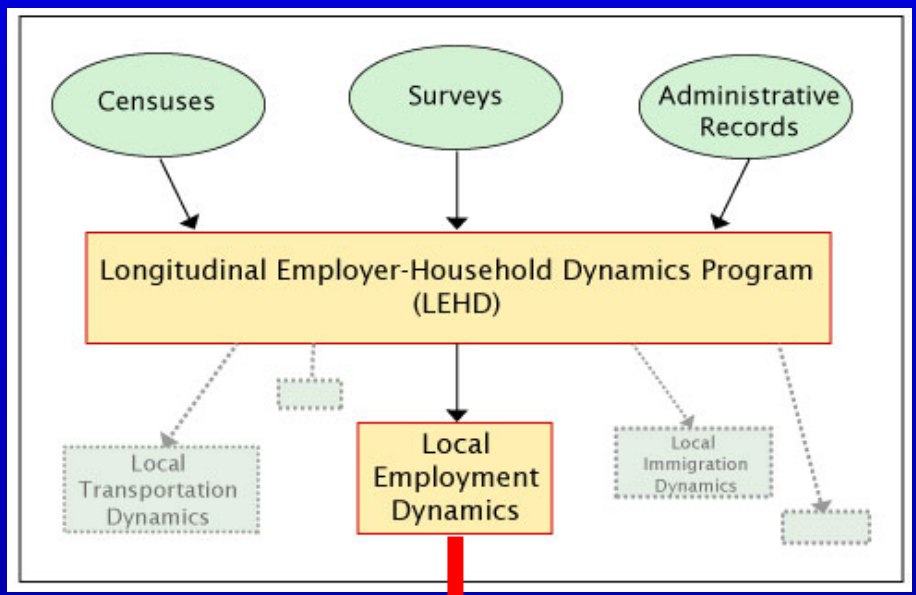
LEHD and Noise Infusion

Jeremy S. Wu
John M. Abowd
US Census Bureau
May 22, 2006

The LEHD Concept



New data and products

- Use existing data to link multiple sources
- Create new data and products
- Reduce cost and respondent burden
- Make data available while protecting confidentiality
- Enhance Census Bureau operations

What is Local Employment Dynamics?

- ✓ A voluntary partnership between the states and the Census Bureau
- ✓ States supply quarterly worker and business wage records
- ✓ Census Bureau merges the state records with other data to produce new data and products:
  - ✓ a longitudinal national frame of jobs
  - ✓ an associated data infrastructure

Three Web-based Products

- QWI Online
  - Flagship product on Quarterly Workforce Indicators
- On The Map
  - Mapping tool to show where workers work and live with companion reports
- Industry Focus
  - Top industries and local workforce characteristics

Local Employment Dynamics

Local

- **State, county, metro areas, workforce investment areas, block groups (On The Map)**

Employment

- **Demographics (age and sex)**
- **Industry**
- **Earnings**

Dynamics

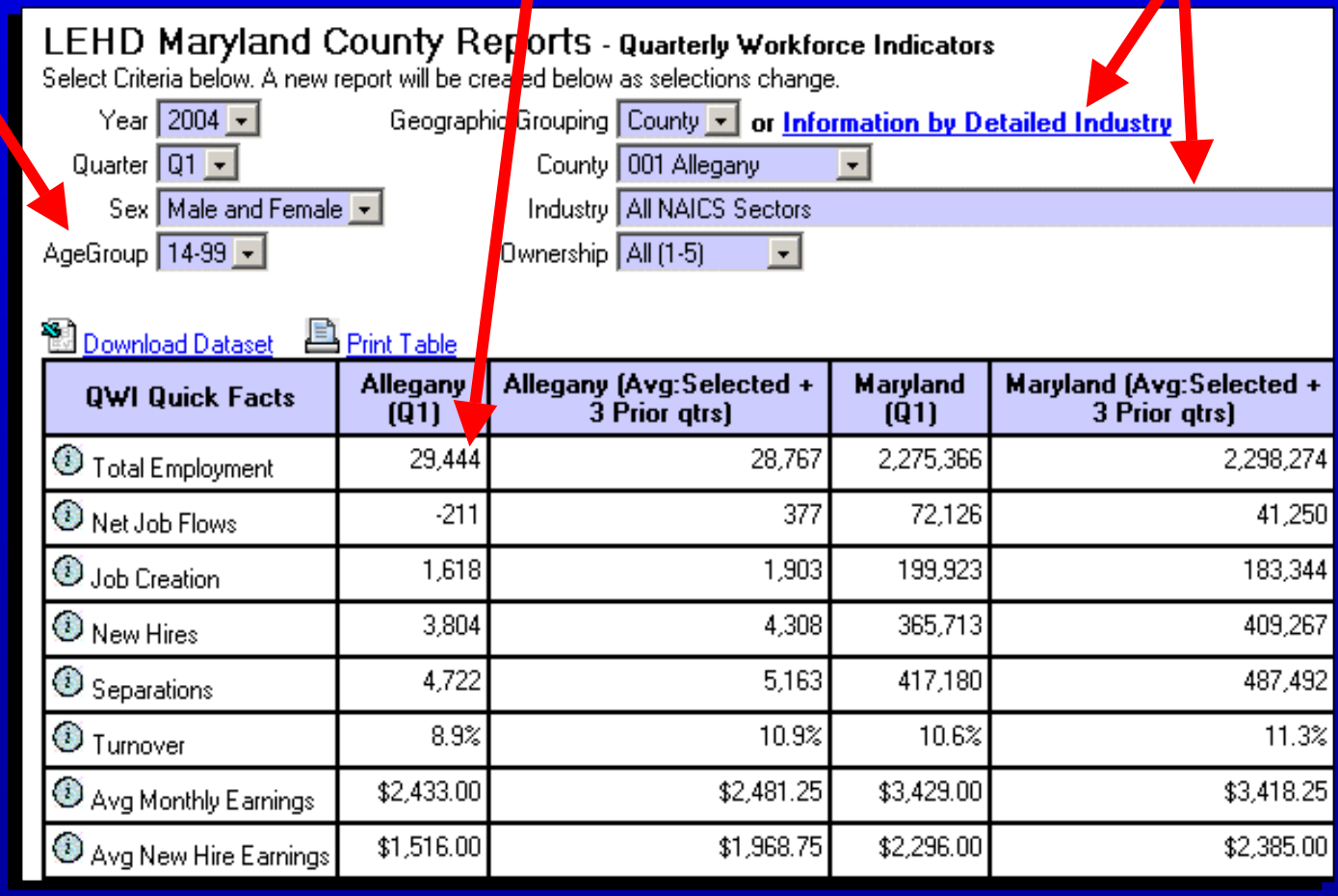
- **Quarterly** **(as far back as 1990, as recent as 9 months ago)**
- **Job gains, losses, and flows**
- **Hires, turnover, and separations**

QWI Online

8 age categories

Entries updated instantaneously

2-, 3-, and 4-digit NAICS code



LEHD Maryland County Reports - **Quarterly Workforce Indicators**

Select Criteria below. A new report will be created below as selections change.

Year: 2004 | Quarter: Q1 | Sex: Male and Female | AgeGroup: 14-99

Geographic Grouping: County or Information by Detailed Industry

County: 001 Allegany | Industry: All NAICS Sectors | Ownership: All (1-5)

Download Dataset | Print Table

QWI Quick Facts	Allegany (Q1)	Allegany (Avg:Selected + 3 Prior qtrs)	Maryland (Q1)	Maryland (Avg:Selected + 3 Prior qtrs)
Total Employment	29,444	28,767	2,275,366	2,298,274
Net Job Flows	-211	377	72,126	41,250
Job Creation	1,618	1,903	199,923	183,344
New Hires	3,804	4,308	365,713	409,267
Separations	4,722	5,163	417,180	487,492
Turnover	8.9%	10.9%	10.6%	11.3%
Avg Monthly Earnings	$2,433.00	$2,481.25	$3,429.00	$3,418.25
Avg New Hire Earnings	$1,516.00	$1,968.75	$2,296.00	$2,385.00

On The Map

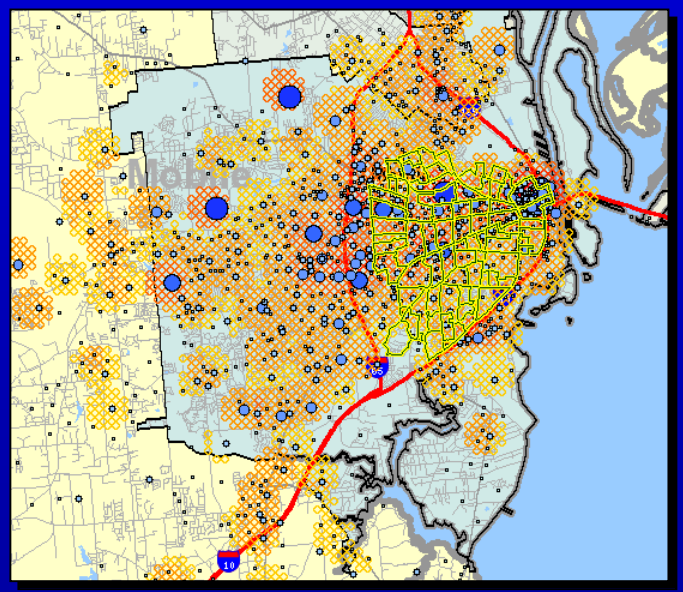
LEHD's online dynamic mapping tool

- ✓ 15 states online
- ✓ 3 states are joining

- ✓ Where do workers live?
- ✓ Where do people work?
- ✓ Companion reports on age, earnings, and industry

- ✓ User select areas
- ✓ Block is base unit for display; block group is base unit for report
- ✓ Modular geographic layers such as community colleges and zip codes

On the Map: Maps and Reports



U.S. Census Bureau *LED On The Map*

Commute Shed Report - Where Residents in the Selection Area are Employed

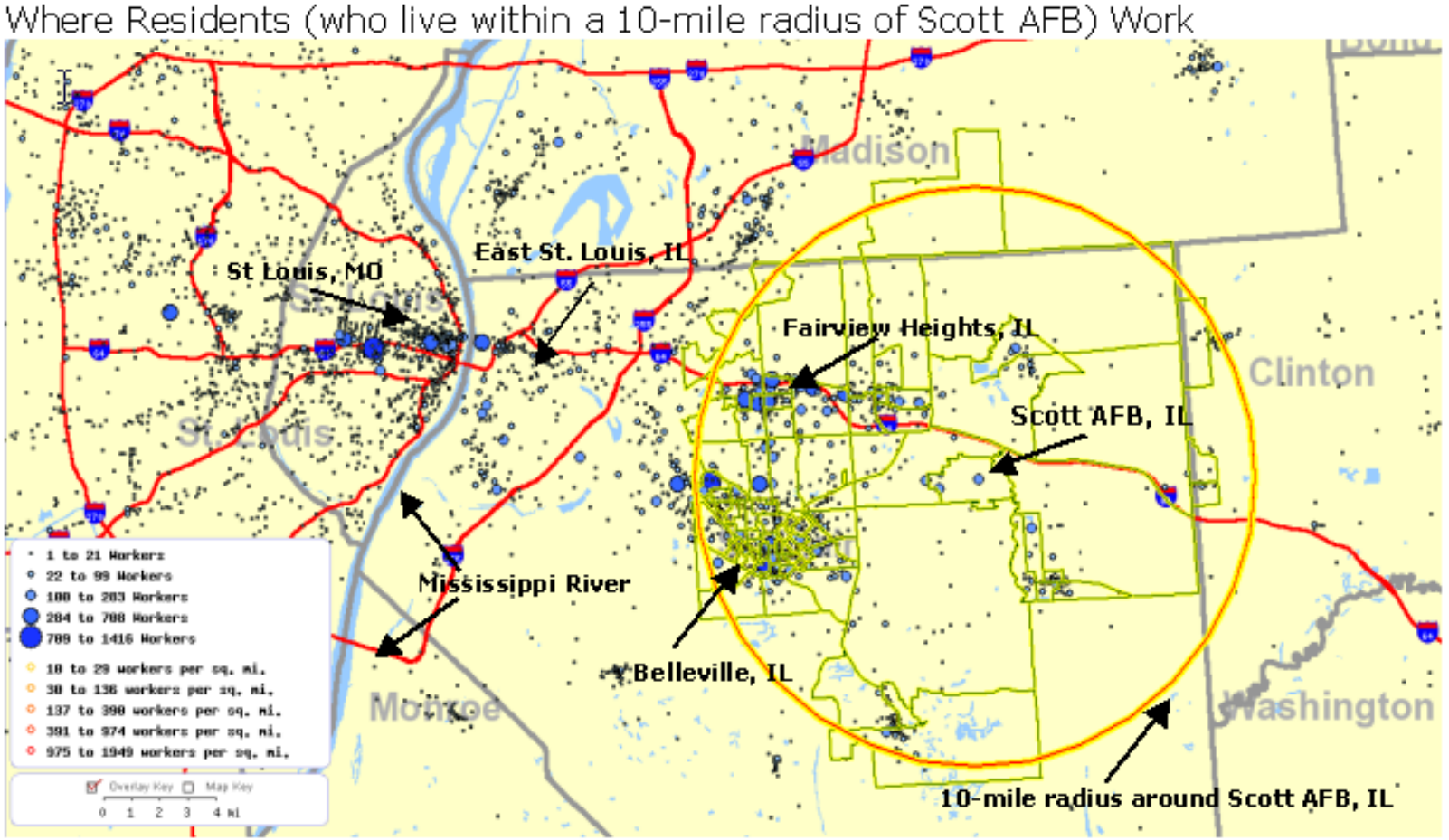
Resident Held Jobs by Category	2003		2002	
	Count	Share	Count	Share
* All Jobs	11,229	100.0%	11,707	100.0%
* All Jobs (Private Sector Only)	9,488	84.5%	10,075	86.1%
* All Primary Jobs (Worker's highest paying job)	10,359	92.3%	10,720	91.6%
* All Primary Jobs (Private Sector Only)	8,720	77.7%	9,193	78.5%

Baseline Count of Workers	2003		2002	
	Count	Share	Count	Share
All Primary Jobs (Private Sector Only)	8,720	100.0%	9,193	100.0%

Cities/Towns Where Residents are Employed	2003		2002	
	Count	Share	Count	Share
* Mobile, AL	5,753	66%	6,157	67%
* Unincorporated Areas	1,038	11.9%	1,085	11.8%
* Birmingham, AL	217	2.5%	301	3.3%
* Daphne, AL	188	2.2%	203	2.2%
* Prichard, AL	187	2.1%	238	2.6%
* All Other Locations	1,337	15.3%	1,209	13.2%

Counties Where Residents are Employed	2003		2002	
	Count	Share	Count	Share
* Mobile Co., AL	6,971	79.9%	7,365	80.1%
* Baldwin Co., AL	577	6.6%	592	6.4%
* Jefferson Co., AL	330	3.8%	409	4.4%
* Montgomery Co., AL	156	1.8%	177	1.9%
* Madison Co., AL	94	1.1%	89	1%
* All Other Locations	592	6.8%	552	6%

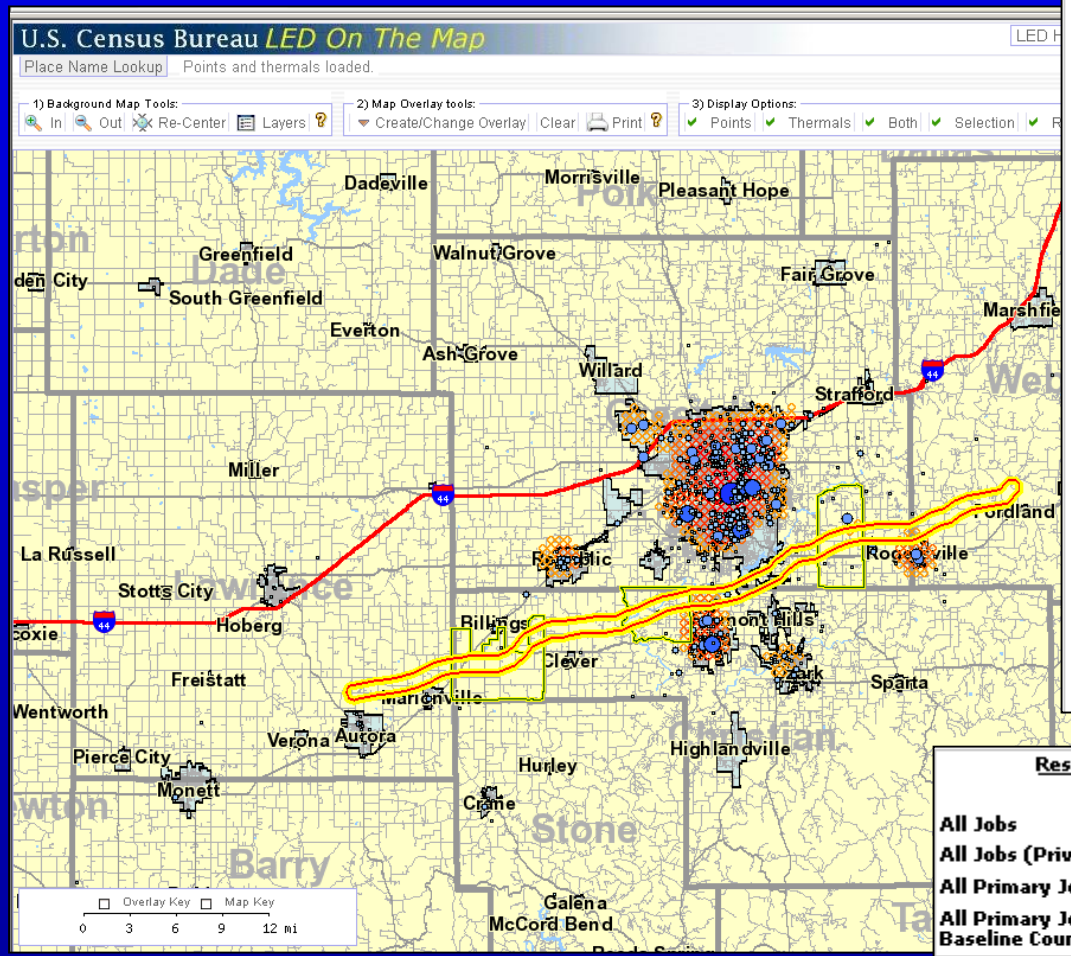
Where should Scott AFB look to recruit civilians gained by BRAC?



LED On the Map: Where People Go to Work

Springfield, MO Tornado Track

March 12, 2006



Workers by Industry Type (2-digit NAICS)	2003 Count	2003 Share
* Agriculture, Forestry, Fishing and Hunting	5	0.3%
* Mining	1	0.1%
* Utilities	3	0.2%
* Construction	134	8.6%
* Manufacturing	195	12.5%
* Wholesale Trade	132	8.5%
* Retail Trade	230	14.8%
* Transportation and Warehousing	79	5.1%
* Information	54	3.5%
* Finance and Insurance	82	5.3%
* Real Estate and Rental and Leasing	25	1.6%
* Professional, Scientific, and Technical Services	70	4.5%
* Management of Companies and Enterprises	35	2.2%
* Administration & Support, Waste Management and Remediation	76	4.9%
* Educational Services	13	0.8%
* Health Care and Social Assistance	202	13%
* Arts, Entertainment, and Recreation	23	1.5%
* Accommodation and Food Services	124	8%
* Other Services (excluding Public Administration)	74	4.8%
* Public Administration	0	0%

Resident Held Jobs by Category	2003 Count	2003 Share
All Jobs	1,943	100.0%
All Jobs (Private Sector Only)	1,644	84.6%
All Primary Jobs (Worker's highest paying job)	1,843	94.9%
All Primary Jobs (Private Sector Only) and Baseline Count of Workers	1,557	80.1%

Workers by Earnings Paid	2003 Count	2003 Share
$1,200 per month or less	452	29%
$1,201 to $3,400 per month	674	43.3%
More then $3,400 per month	431	27.7%

Where do people who live in the path of the tornado ***work***?

How do they do that? Confidentiality Protection Methods

What we are using in the applications and are moving to:

- Noise infusion
- Synthetic data

What we experimented with and rejected:

- Suppression
- Raking to related BLS estimates

Core Problem

- Disclosure avoidance system is required to protect the information about individuals and businesses that contribute to
  - Confidential unemployment insurance (UI) wage records
  - Confidential Quarterly Census of Employment and Wages (QCEW, also known as ES-202) reports
  - Confidential information from Census Bureau demographic data that have been integrated with these sources.
- Primary concern of the confidentiality protection mechanism is cells that reflect data on few a individuals or a few establishments/firms.

Definitions of Protection Provided by the QWI Noise Infusion System

- Data are considered protected when *"aggregate cell values do not closely approximate data for any one respondent in the cell"* (Cox and Zayatz, 1993, pg. 5)
- In the QWI confidentiality protection scheme, confidential micro-data are considered protected by noise infusion if
  1. Any inference regarding the magnitude of a particular respondent's data differs from the confidential quantity by at least c% even if that inference is made by a coalition of respondents with exact knowledge of their own answers, or
  2. Any inference regarding the magnitude of an item is incorrect with probability no less than y%, where c and y are confidential but generally large

Quality of the Released Data

- The confidentiality-protected data must be inference-valid for a well-defined set of analyses
- We show that
  - The theoretical properties of the disclosure avoidance system are designed to maintain analytical validity for
    - trend analysis;
    - in practice, the released data are not biased;
    - in practice, the time-series properties of the released data remain intact.

Layers of the QWI Protection System

Layer 1: Multiplicative noise-infusion at the establishment level, with three very important properties

- Every establishment-level data item is distorted by some minimum amount
- The distortion amount and direction are time-invariant: data are always distorted in the same direction (increased or decreased) by the same percentage amount in every period, for a given establishment
- When estimates are created by aggregating over establishments by geography, industry, ownership, sex and age group, the effects of the noise infusion cancel out for the vast majority of the estimates
- Estimates that would be primary suppressions or ones that are substantially different from the same estimate based on the raw confidential data are flagged as “substantially distorted to protect confidentiality.”

Layers of the QWI Protection System (II)

Layer 2: Weighting of estimates at higher levels (e.g., sub-state geography and industry detail)

- Construct weights such that state-level beginning-of-quarter employment for all private employers matches the first month in quarter employment in QCEW.
- This establishment-level weight is used for every indicator in the QWI.

Layers of the QWI Protection System (III)

Layer 3: Small-cell editing (Suppression or synthesis)

- Some aggregate estimates are based on fewer than three persons or establishments.
- Currently, these estimates are suppressed and a flag set to indicate suppression.
- In experiment version of the QWI disclosure avoidance system and in the current version of OnTheMap, these estimates are replaced with synthetic values.
  - Note: Editing is only used when the combination of noise infusion and weighting may not distort the publication data with a high enough probability to meet the criteria cited above.
  - Estimates of employment and related quantities are subject to editing.
  - Dollar measures like payroll and wages/employee are not.
  - Regardless of small-cell editing, all published estimates are still based on the noise-infused data

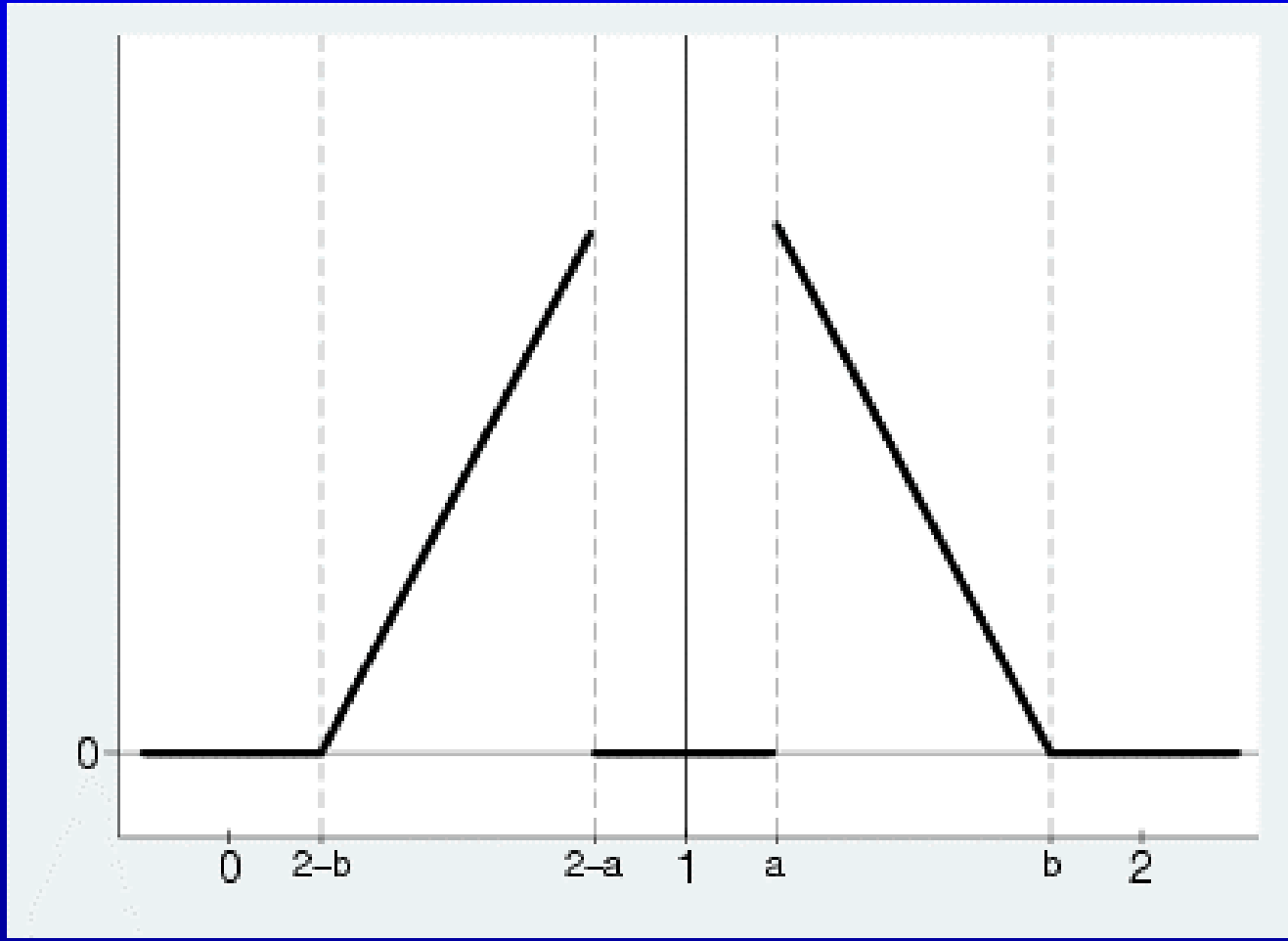
Basic Noise Factor Distribution

$$p(\delta_j) = \begin{cases} (b-\delta)/(b-a)^2, & \delta \in [a,b] \\ (b+\delta-2)/(b-a)^2, & \delta \in [2-b,2-a] \\ 0, \text{ otherwise} \end{cases}$$

$$F(\delta_j) = \begin{cases} 0, & \delta < 2-b \\ (\delta+b-2)^2 / \left[2(b-a)^2\right], & \delta \in [2-b,2-a] \\ 0.5, & \delta \in (2-a,a) \\ 0.5 + \left[(b-a)^2 - (b-\delta)^2\right], & \delta \in [a,b] \\ 1, & \delta > b \end{cases}$$

A random factor δ is created using the distribution above, where $a = 1 + c/100$ and $b = 1 + d/100$ are constants chosen such that the true value is distorted by a minimum of c percent and a maximum of d percent.

Distribution of the Noise Factor



Specific Formulas: Magnitudes

For estimates of employment or payroll (magnitude data)

$$X^{*}_{jt} = \delta_j X_{jt}$$

Specific Formulas: Ratios

For estimates of two magnitudes (payroll per employee)

$$ZY_{jt}^{*} = \delta_j \frac{Y_{jt}}{B(Y_{jt})}$$

Other Considerations

- Only two of three may be released of magnitude, base, ratio
- Non-linear functions of the micro-data (*e.g,* job creations and destructions) require special methods (developed for the QWI but not discussed here)

Item Suppression or Synthesis

- Some disclosure risk remains for employment estimates based on very few entities in a cell (fewer than three individuals or employers)
- Item suppression based on the number of either workers or the number of employers that contribute data for that item in a cell k in time period t, where a cell represents a particular combination of ownership x geography × industry × age × sex.
- Because of noise infusion, no complementary suppressions are needed
- Some denominators may be zeroes - the ratio or rate cannot be computed.
- All suppressions are replaced by synthetic values in the experimental system.

Protection: Employment

Table 2: Small Cells: B, Undistorted vs. Distorted
(a) Illinois

Undistorted count	*Distorted count* 0	1	2	3	4	5 or more
0	99.86	0.14	0.00	0.00	0.00	0.00
1	0.91	95.75	3.34	0.00	0.00	0.00
2	0.00	4.27	87.25	8.47	0.00	0.00
3	0.00	0.00	10.69	77.20	12.11	0.00
4	0.00	0.00	0.00	14.73	67.49	17.78
5 or more	0.00	0.00	0.00	0.00	1.93	98.07

Total number of cells: 14,229,968 . Both comparisons are for weighted data. For details, see text.

Protection: Employment

Table 3: Small Cells: B, Raw vs. Published
(a) Illinois

Unweighted count	*Published count*						
	Suppressed	0	1	2	3	4	5 or more
0	0.79	99.21	0.00	0.00	0.00	0.00	0.00
1	99.91	0.08	0.00	0.00	0.00	0.00	0.00
2	94.02	0.01	0.00	0.00	5.87	0.09	0.01
3	34.33	0.00	0.00	0.00	47.75	16.98	0.94
4	25.87	0.00	0.00	0.00	5.56	43.24	25.32
5 or more	15.20	0.00	0.00	0.00	0.03	0.82	83.95

Total number of cells: 14,229,968 . Raw is unweighted and undistorted. Published is after weighting, distorting, and suppression. For details,

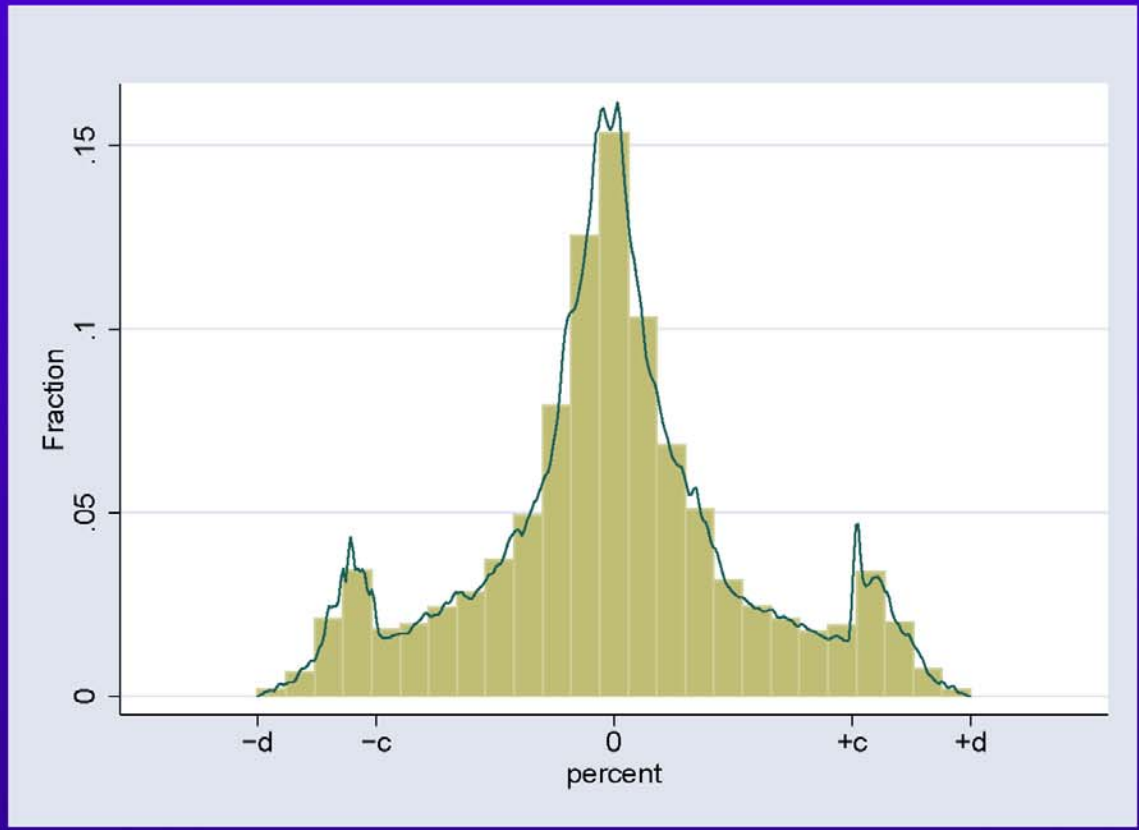
Distribution of the Error in the First Order Serial Correlation, Raw vs. Distorted Data

$$\Delta r = r - r^*$$

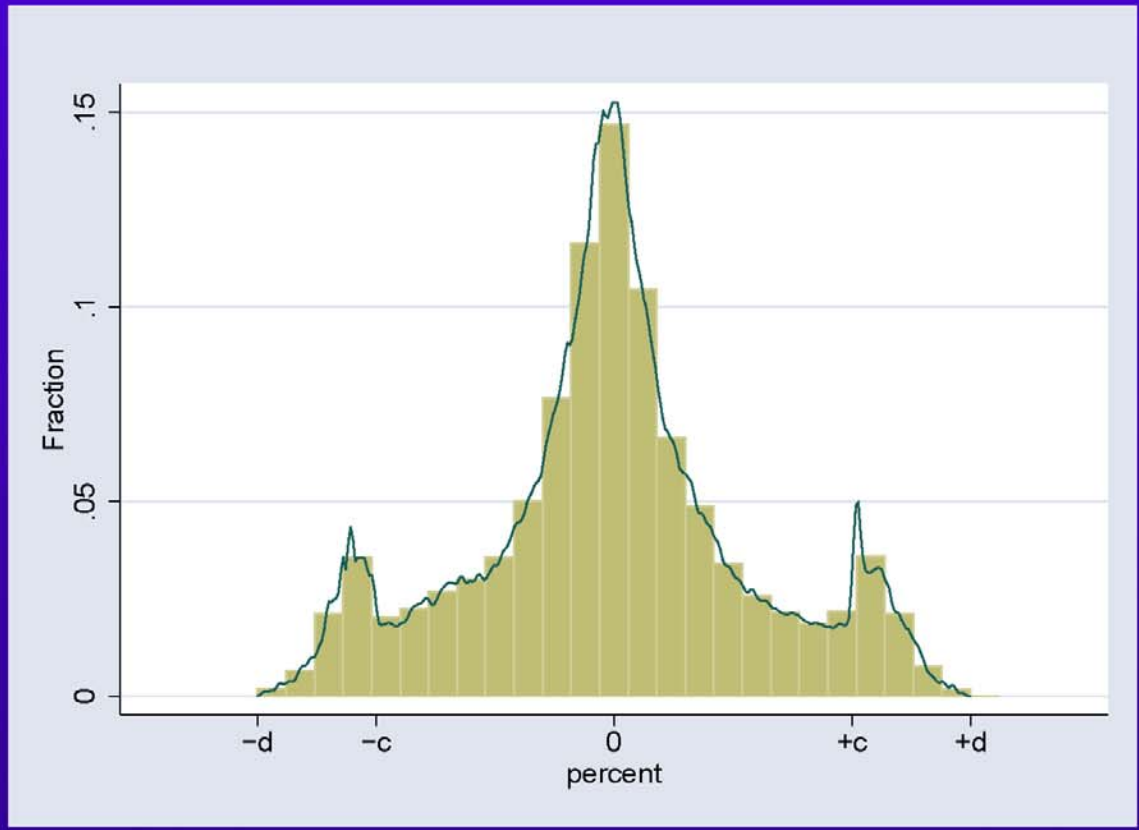
Percentile	B	A	S	F	JF
01	-0.069373	-0.049274	-0.052155	-0.066461	-0.007969
05	-0.041585	-0.031460	-0.032934	-0.039787	-0.004651
10	-0.028849	-0.022166	-0.023733	-0.027926	-0.002785
25	-0.011920	-0.009996	-0.010161	-0.011913	-0.001003
50	0.000571	0.000384	0.000768	0.000306	-0.000044
75	0.013974	0.011806	0.012891	0.012632	0.000776
90	0.030948	0.025152	0.026290	0.028299	0.002263
95	0.044233	0.033871	0.037198	0.040565	0.004375
99	0.078519	0.054415	0.060327	0.074212	0.007845

SIC-division $\times$ County, State of Illinois. r from AR(1) estimated for each cell's time series.

Cross-sectional unbiasedness: B



Cross-sectional unbiasedness: W_1



Contact Us

Program Manager
Jeremy.S.Wu@census.gov

Comments/Suggestions
dsd.local.employment.dynamics@census.gov