Regression Composite Estimation for Current Population Survey

Tim Trudell
U.S. Census Bureau

Joint work with
Yang Cheng, formerly U.S. Census Bureau
Daniel Bonnery, University of Maryland
Partha Lahiri, University of Maryland

2019 Joint Statistical Meetings, Denver

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Outline

1. Background
2. Current Population Survey (CPS) sample design
3. CPS direct estimator
4. AK composite estimation
5. Composite regression estimation
6. Simulation study
7. Application to CPS data
8. Conclusion

CPS background

- American labor force statistics
- Oldest national survey - since 1940
- A household sample survey sponsored by Bureau of Labor Statistics and U.S. Census Bureau
- Monthly sample of 72,000 Households
- Primary source of labor force data:
  - Monthly national unemployment rate
- Labor force data for people aged 16 and over

CPS sample design

Stratified multi-stage sampling design:

- Primary Sampling Unit (PSU)
  - Consist of county or a group of counties
  - 1,987 PSUs with 506 self-representing (SR) PSUs, roughly 70% of total population (2,205 PSUs with 446 SR PSU for 2000)
  - Group non-self-representing (NSR) PSUs into stratum
- Stratified two-stage design on NSR groups
  - First stage: select one PSU per stratum by probability proportional to size method restricted to population 16+
  - Second stage: Systematic sampling on the clusters of 4 sampled housing units

CPS rotation panel design

- Repeated rotating sample design with 8 rotation panels
- 4-8-4 Sample Pattern: a housing unit selected is interviewed for 4 consecutive months, out for 8 months, and then interview for another 4 months
- Survey modes: personal visit interviews and telephone interviews based on sampling in the different panels
- Self weighting at the state level: all sampled units have equal weights. Self weighting samples often yield smaller variance, and sample statistics are more robust

CPS/SCHIP ROTATION CHART

January 2012 - March 2014

Sample and Rotation

Year/Month		A88/B88	A89/B89	A90/B90	A91/B91	A92/B92	A93/B93
2012	JAN	. . 3 4 5 6 . .	. . . . . . 7 8	1 2			
	FEB	. . . 4 5 6 7 .	 8	1 2 3			
	MAR	. . . . 5 6 7 8	. . . .	1 2 3 4			
	APR	. . . . . 6 7 8	1	. 2 3 4 5			
	MAY	. . . . . . 7 8	1 2	. . 3 4 5 6			
	JUNE	 8	1 2 3	. . . 4 5 6 7			
	JULY	. . . .	1 2 3 4	. . . . 5 6 7 8			
	AUG	. . . .	. 2 3 4 5 . . .	. . . . . 6 7 8	1		
	SEPT	. . . .	. . 3 4 5 6 . .	. . . . . . 7 8	1 2		
	OCT	. . . .	. . . 4 5 6 7 .	 8	1 2 3		
	NOV	. . . .	. . . . 5 6 7 8	. . . .	1 2 3 4		
	DEC	. . . .	. . . . . 6 7 8	1	. 2 3 4 5		
2013	JAN	. . . .	. . . . . . 7 8	1 2	. . 3 4 5 6		
	FEB	. . . .	 8	1 2 3	. . . 4 5 6 7		
	MAR	. . . .	. . . .	1 2 3 4	. . . . 5 6 7 8		
	APR	. . . .	. . . .	. 2 3 4 5 . . .	. . . . . 6 7 8	1	
	MAY	. . . .	. . . .	. . 3 4 5 6 . .	. . . . . . 7 8	1 2	
	JUNE	. . . .	. . . .	. . . 4 5 6 7 .	 8	1 2 3	
	JULY	. . . .	. . . .	. . . . 5 6 7 8	. . . .	1 2 3 4	
	AUG	. . . .	. . . .	. . . . . 6 7 8	1	. 2 3 4 5	
	SEPT	. . . .	. . . .	. . . . . . 7 8	1 2	. . 3 4 5 6	
	OCT	. . . .	. . . .	 8	1 2 3	. . . 4 5 6 7	
	NOV	. . . .	. . . .	. . . .	1 2 3 4	. . . . 5 6 7 8	
	DEC	. . . .	. . . .	. . . .	. 2 3 4 5 . . .	. . . . . 6 7 8	1
2014	JAN	. . . .	. . . .	. . . .	. . 3 4 5 6 . .	. . . . . . 7 8	1 2
	FEB	. . . .	. . . .	. . . .	. . . 4 5 6 7 .	 8	1 2 3
	MAR	. . . .	. . . .	. . . .	. . . . 5 6 7 8	. . . .	1 2 3 4

Notations

- t: time indicator (month)
- U_t: the population at the time t, of size N_t
- $S_{t,i}$: the ith rotation group for month t
- $S_t = U_{i=1}^{8} S_{t,i}$
- $w_{t,k}$: the weight after adjustment of individual k at month t
- y_t : a vector of study variables $\left(y_{t,k}\right)_{k \epsilon U_t}$
- $Y_t = \sum_{k \epsilon U_t} y_{t,k}$: the unknown total to be estimated at month t

Direct estimator

The weights $w_{t,k}$ are obtained after different adjustment procedures:

- 2 procedures at the household level
- 5 procedures at the individual level

The direct estimator of Y_t is

$$\hat{Y}_t = \frac{1}{8}\sum_{i=1}^{8}\hat{Y}_{t,i}$$

where $\hat{Y}_{t,i} = \sum_{k \epsilon s_{t,i}} w_{t,k}\, y_{t,k}$, the estimator of Y_t based on data in the ith rotation panel at month t, $i = 1, \dots, 8$.

Composite estimator before 1985

Let $\Delta_t = Y_t - Y_{t-1}$ be the month-to-month change. Then

$$Y_t = Y_{t-1} + \Delta_t$$

which suggests that we may construct more efficient estimator by incorporating historical information properly. Because of the 4-8-4 sample rotation scheme, six of eight rotation panels in the sample for month *t-1* remains in sample for month t. The month-to-month change can be estimated as

$$\hat{\Delta}_t = \frac{1}{6}\sum_{i \in s}\left(\hat{Y}_{t,i} - \hat{Y}_{t-1,i-1}\right)$$

where $S = \{2,3,4,6,7,8\}$. The composite estimator before 1985 is defined as

$$\hat{Y}_t^C = (1-K)\hat{Y}_t + K(\hat{Y}_{t-1}^C + \hat{\Delta}_t), 0 \leq K \leq 1$$

AK composite estimator

- In 1985, a different composite estimator was introduced by adding another term to the previously described composite estimator $\hat{Y}_t^C$, a term that is the estimator of the net difference between the incoming and continuing parts of the current month's sample:

$$\hat{\beta}_t = \frac{1}{8}\left(\sum_{i \notin S} \hat{Y}_{t,i} - \frac{1}{3}\sum_{i \epsilon s} \hat{Y}_{t,i}\right)$$

- Then the AK composite estimator is defined as

$$\hat{Y}_t^{AK} = (1-K)\hat{Y}_t + K\left(\hat{Y}_{t-1}^{AK} + \hat{\Delta}_t\right) + A\hat{\beta}_t, 0 \le K \le 1.$$

Note: constants A and K are determined empirically.

Composite regression estimation

- Singh, A. and Merkouris, P. (1995)
  Composite estimation by modified regression for repeated surveys
  ASA Proceedings Survey Research Methods Section, page 400-425

- Fuller, W. A. and Rao, J. (2001)
  A regression composite estimator with application to the Canadian labor force survey
  Survey Methodology, Vol. 27, No. 1, page 45-51

Composite regression estimation (cont.)

- $x_{t,k}$: a vector of auxiliary variables for unit *k* at time *t*
- Definition of $z_{t,k}$ and $\hat{t}^{C}_{y,t}$:

$$z_{t,k} = \begin{cases} \alpha\left(\tau^{-1}\left[y_{t-1,k} - y_{t,k}\right] + y_{t,k}\right) + (1-\alpha)y_{t-1,k} & if\ k\epsilon S_t \cap S_{t-1} \\ \alpha y_{t,k} + (1-\alpha)\ N_{t-1}^{-1}\ \hat{t}^{C}_{y,t-1} & if\ k \in S_t \backslash S_{t-1} \end{cases}$$

and

$$\hat{t}^{C}_{y,t} = \sum_{k \in S_t} w^{C}_{t,k} y_{t,k}$$

- τ is a fixed number, which closes to $\left(\sum_{k\in S_t} w_{t,k}\right)^{-1} \sum_{k\epsilon S_t \cap S_{t-1}} w_{t,k}$

Composite regression estimation (cont.)

- $w^C_{t,k}$ minimizes $\sum_k \left(w^C_{t,k} - w_{t,k}\right)^2 / w_{t,k}$ under the constraints:

$$\sum_{k \epsilon s_t} w^C_{t,k}\, x_{t,k} = X_t$$

and

$$\sum_{k \epsilon s_t} w^C_{t,k}\, z_{t,k} = \hat{t}^C_{y,t}$$

Then, the regression composite estimator of Y_t is

$$\hat{t}^C_{y,t} = \sum_{k \in S_t} w^C_{t,k} y_{t,k}$$

Simulation study

- Create population of size 100,000
- Rotation group size: 100
- Employment status constrained at the population level to match CPS estimates
- Change of status affects the smallest number of people
- Number of replications of the sample selection process: 1,000

Figure : Level estimates

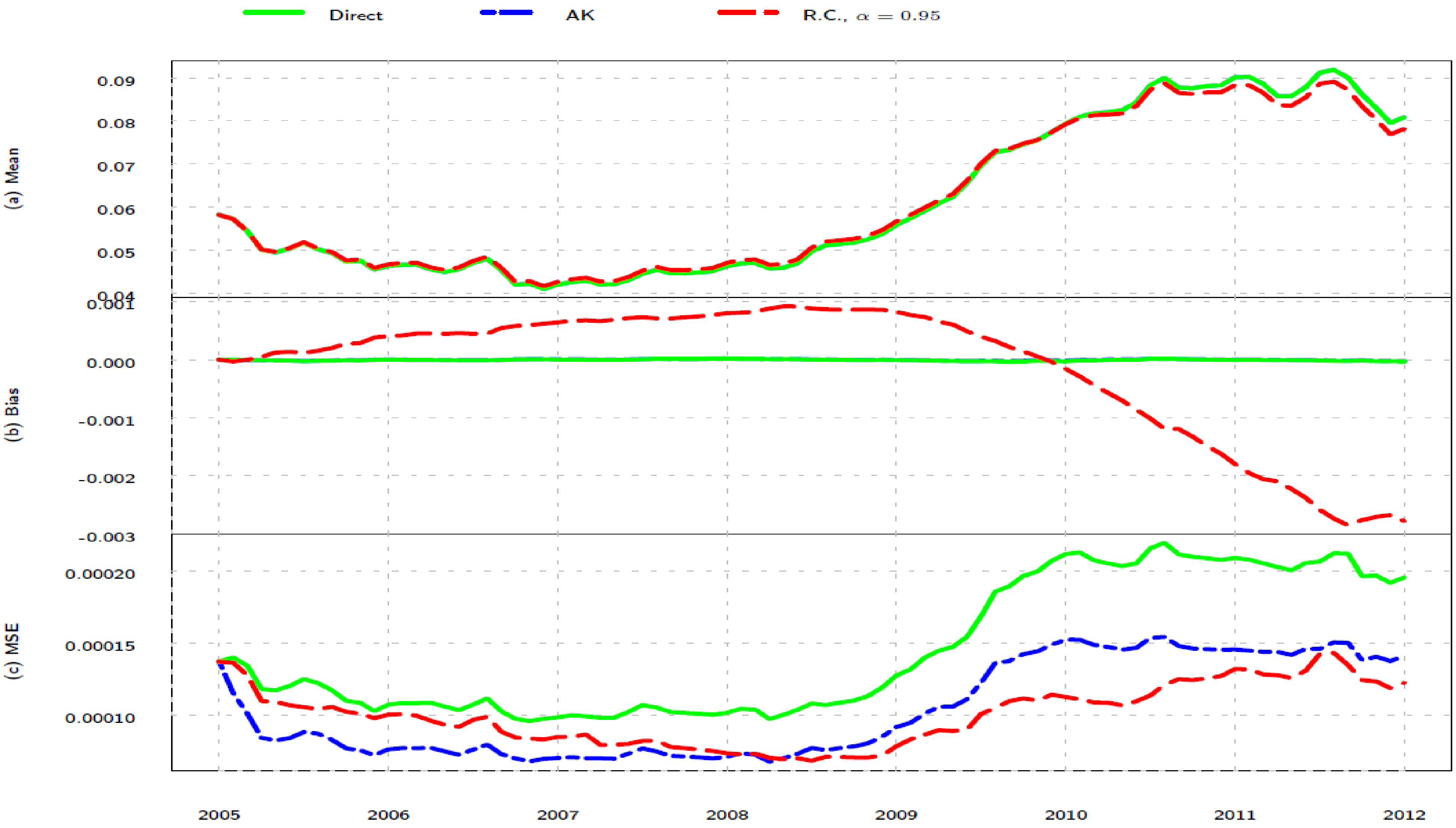


Figure : Month-to-month change estimates

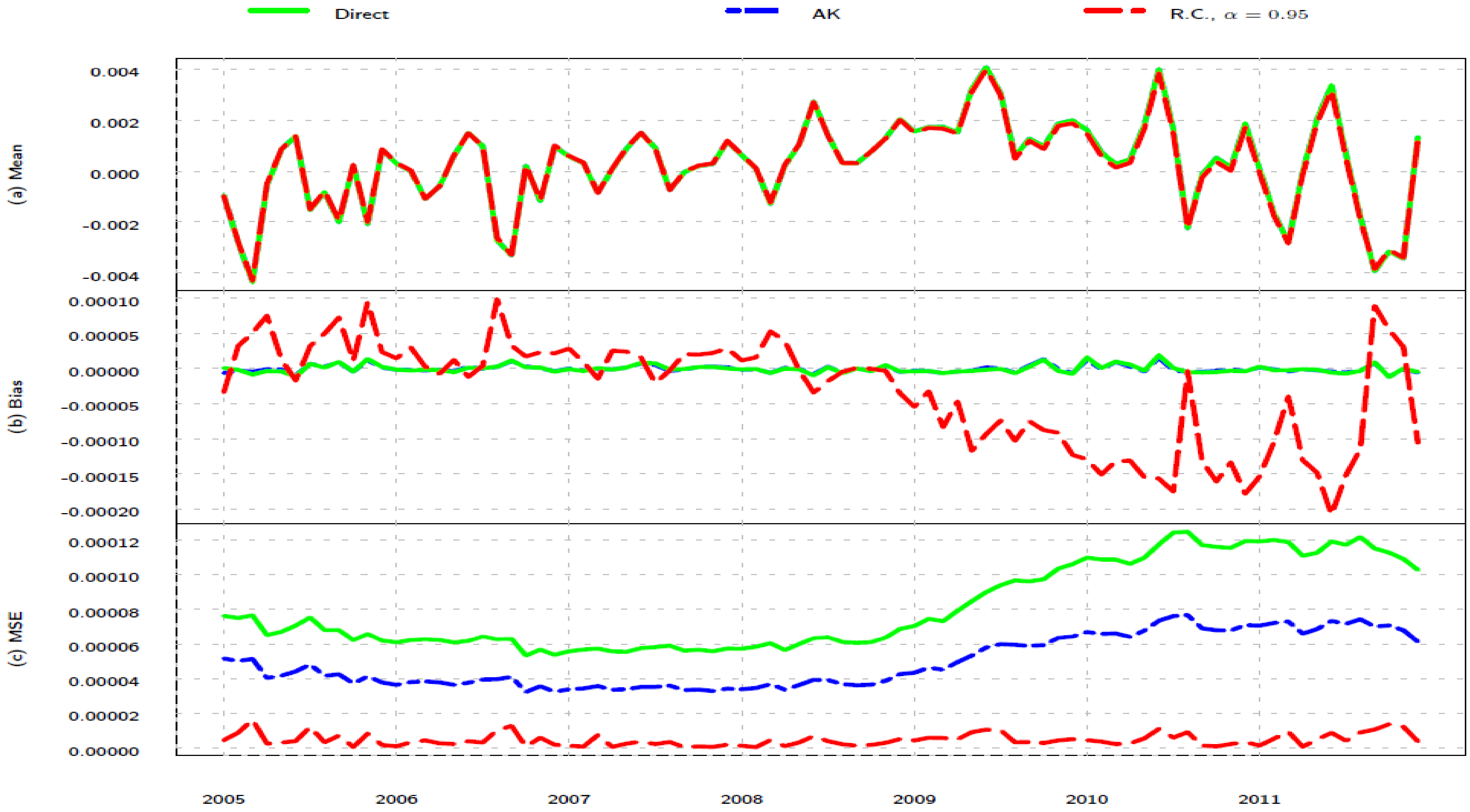


Figure : Level estimates

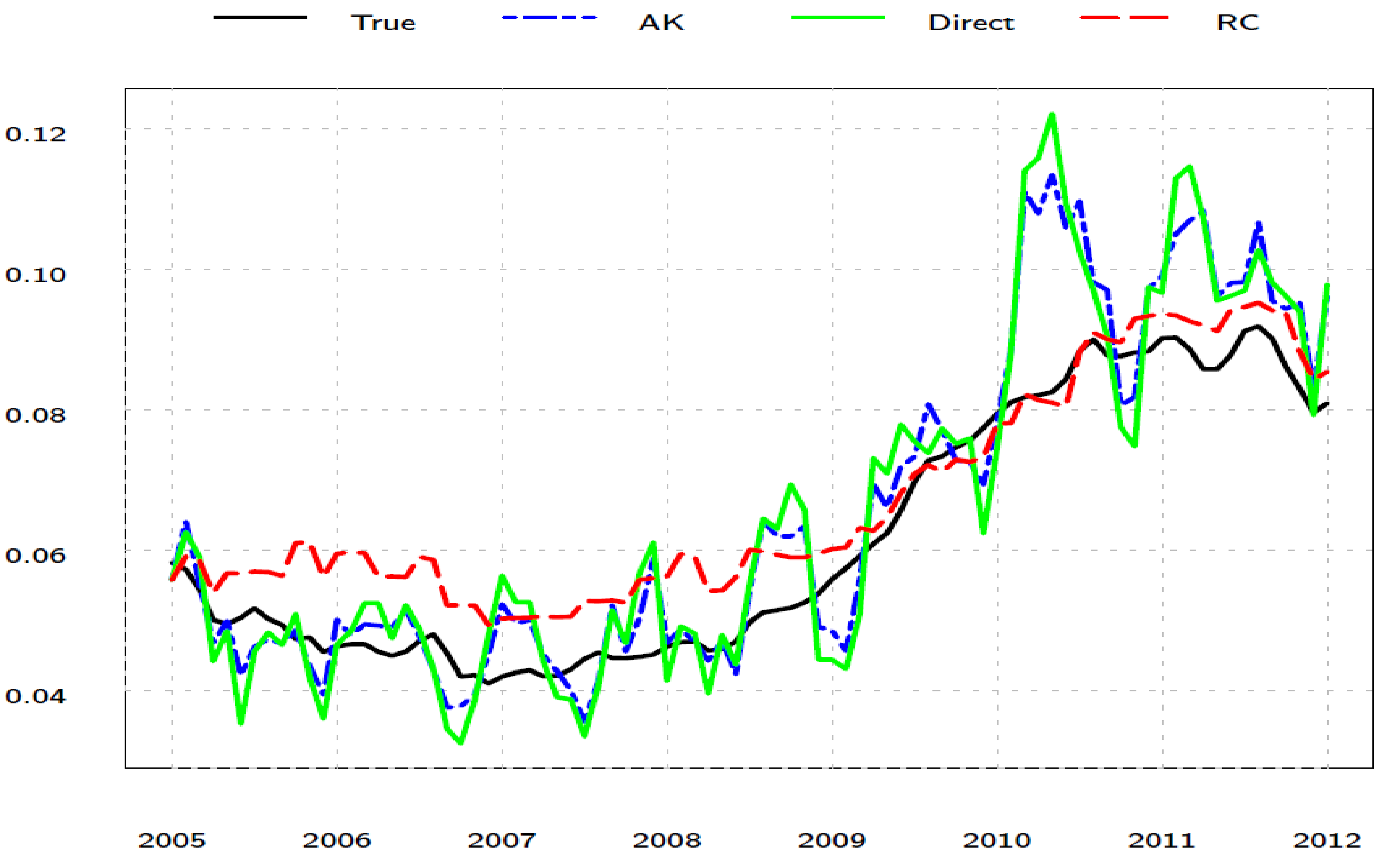


Figure : Level estimates

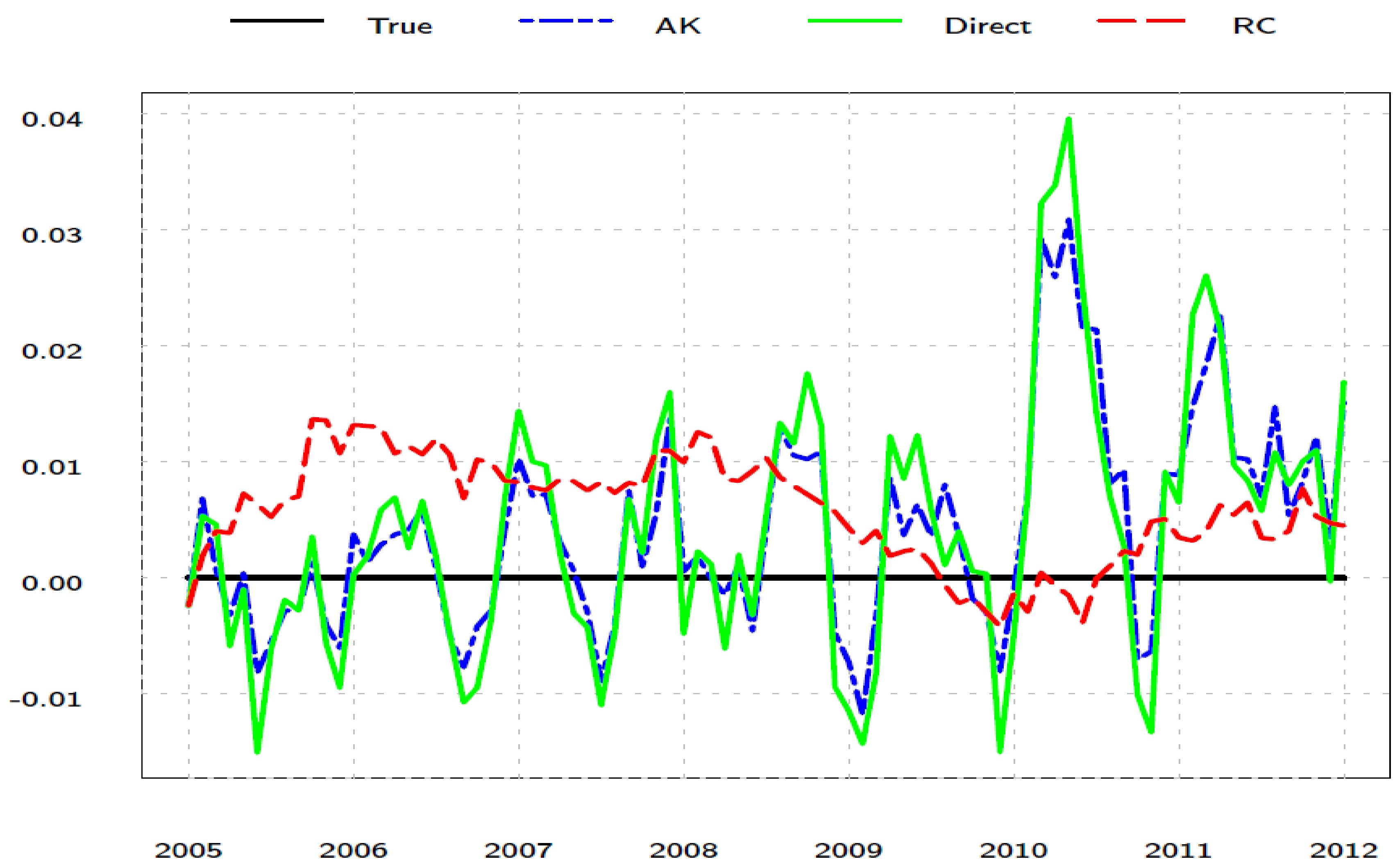
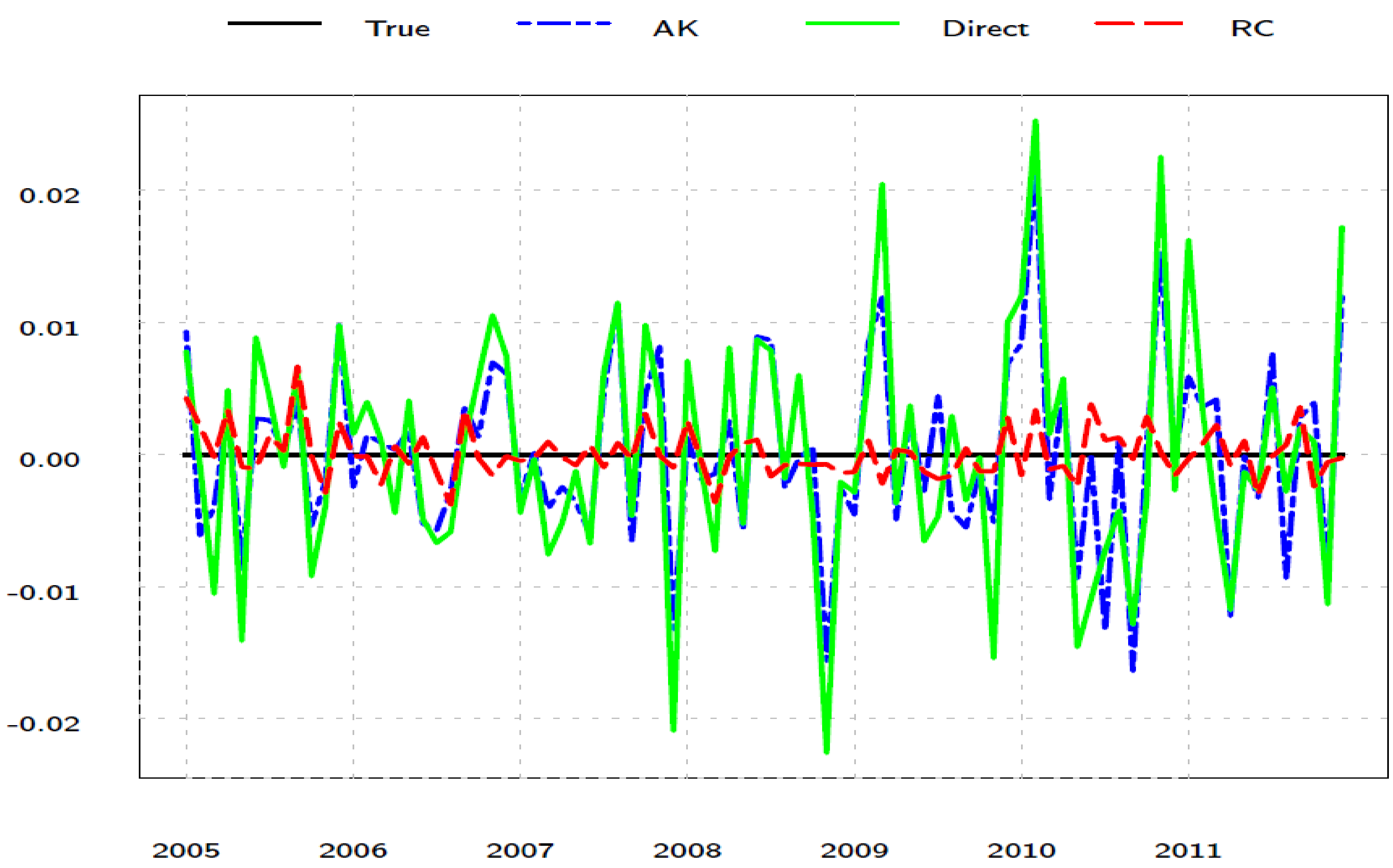


Figure : Month-to-month change estimates



Application to CPS data

Figure : Level estimates: difference with direct estimates

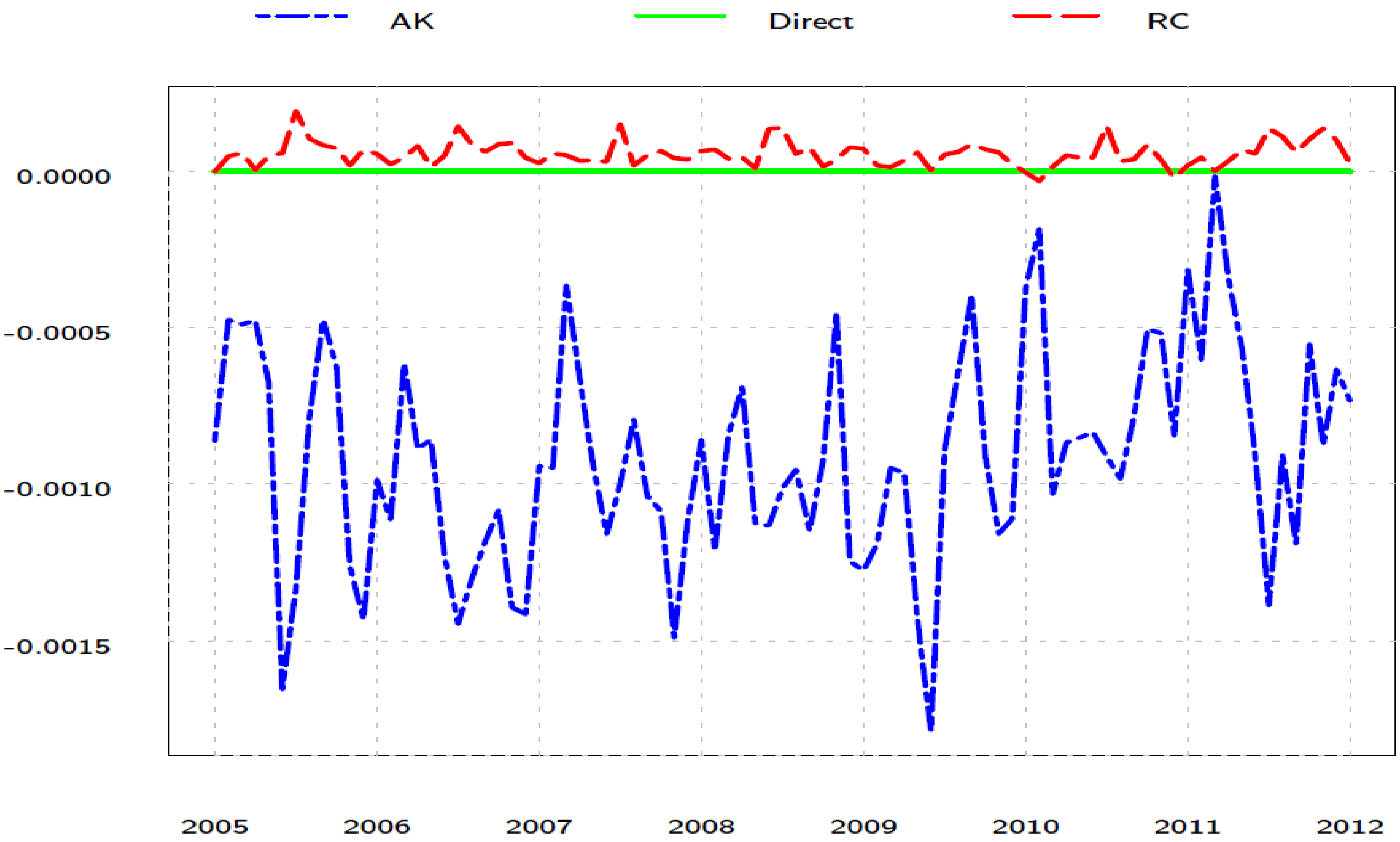


Figure : Month-to-month: difference with direct estimates

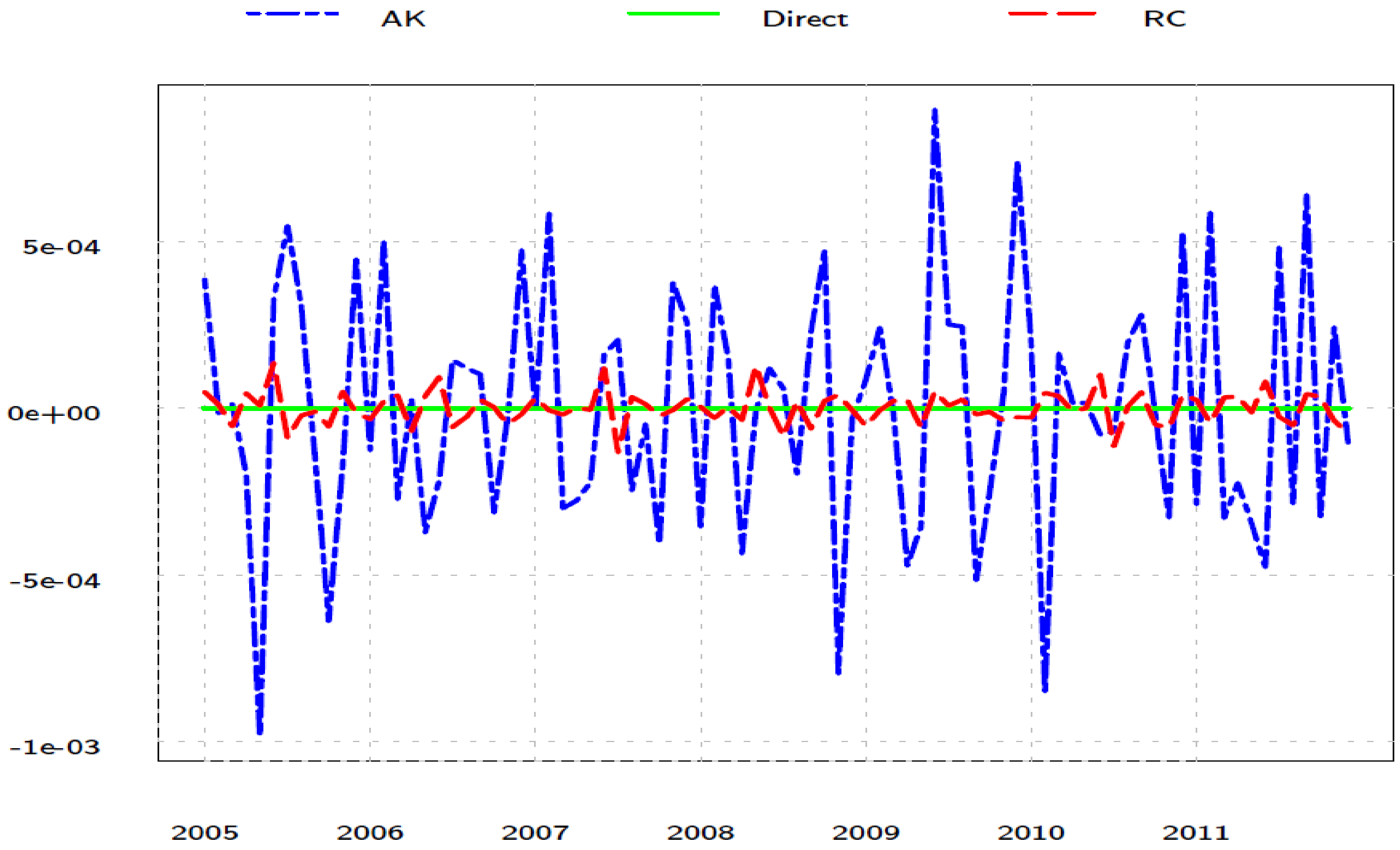


Figure : Weights ratio dispersion

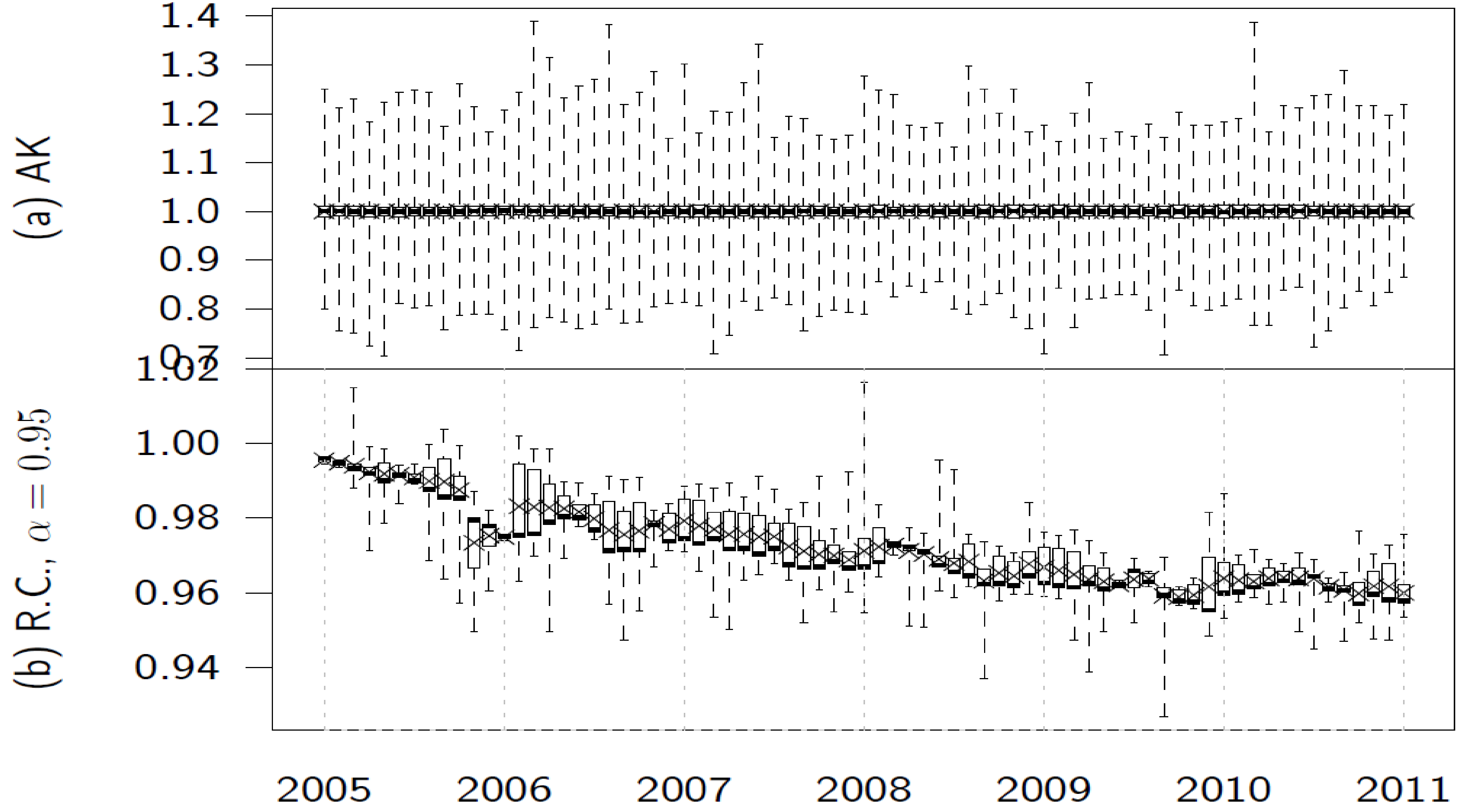
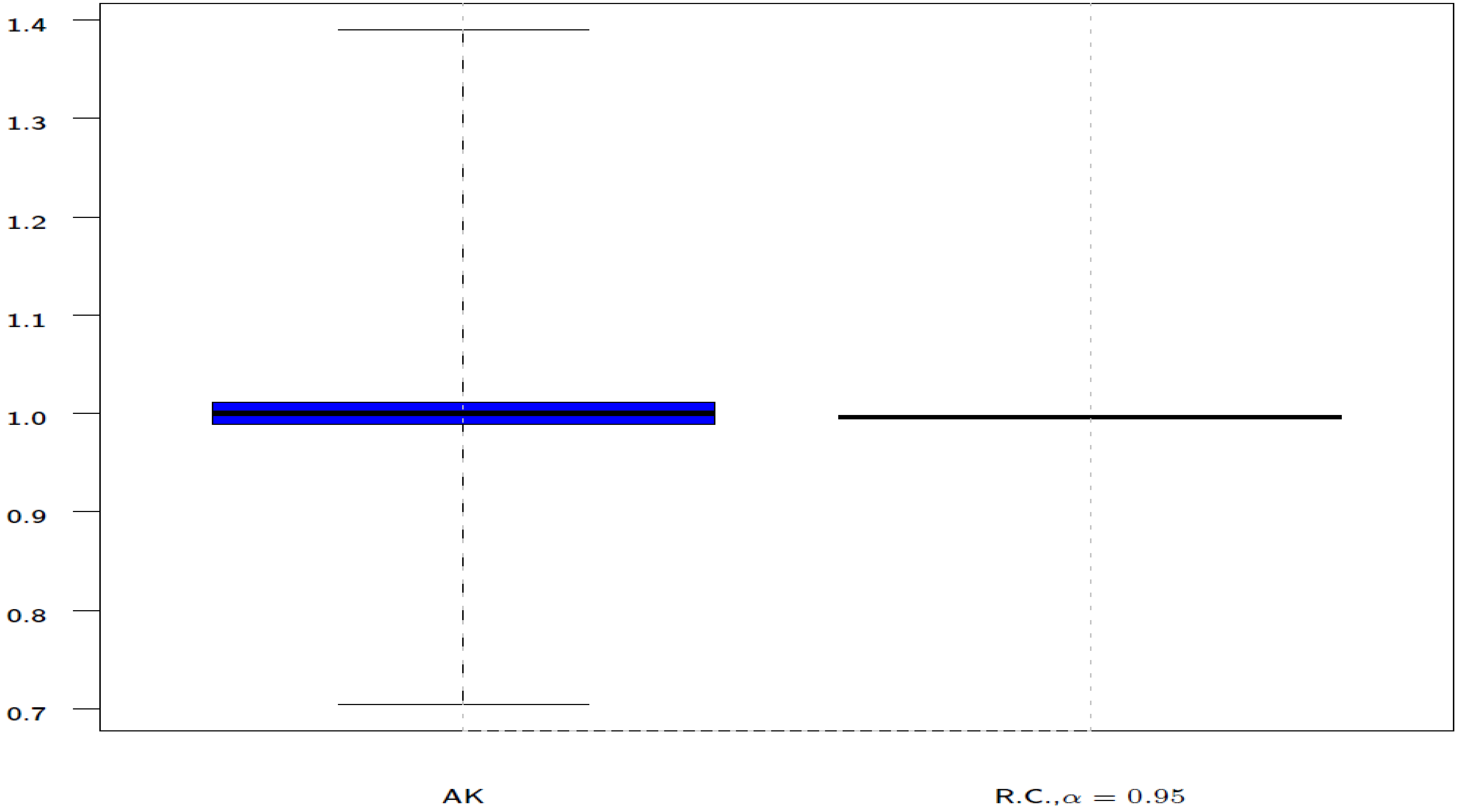


Figure : Weights ratio dispersion



Conclusion

- Regression Composite performs better with respect to month-to-month change estimation on simulated population
- Regression Composite estimate is closer to the direct estimate
- Weights ratio dispersion after Regression Composite adjustment is smaller

Thank you! Questions?

Contact information:

Tim Trudell: tim.trudell@census.gov