An Effective Approach to the Repeated Cross Sectional Design

The *American Journal of Political Science*, January 2015 issue.
Matthew Lebo, Stony Brook University
&
Christopher Weber, University of Arizona

and:

ARFIMA-MLM Package for R
by: Patrick Kraft (Stony Brook), Weber, and Lebo

February 6, 2015 The International Methods Colloquium

Panels, Pseudo-panels, and RCS Designs

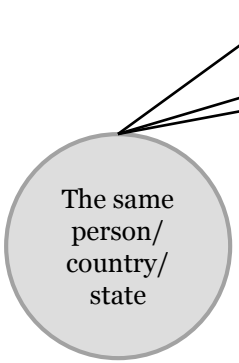
- Panels have the same observations at multiple points in time.
- Pseudo-panels do not have identical sets of cases at every point in time.
  - unbalanced panels will have some observations appearing more than once.
  - repeated cross-sectional designs (RCS) will not have any observation appearing more than once.

How prevalent are RCS data?

- Very! E.g.:
- Cumulative NES file.
- National Annenberg Election Study.
- General Social Survey.
- Stringing together archived files at ICPSR or Roper can create hundreds of consecutive Gallup Surveys, CBS/NYT polls, World Value Surveys.
- Michigan's Survey of Consumers.
- 2010-2013, 42 articles in the APSR and AJPS have RCS as the underlying data structure.

A True Panel

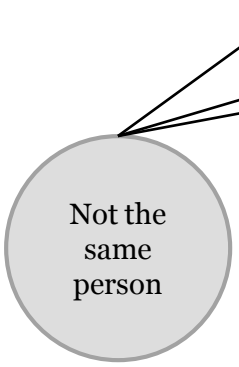
t=1	t=2	t=3	…	t=T
$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	…	$y_{1,T}$
$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	…	$y_{2,T}$
$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	…	$y_{3,T}$
$y_{4,1}$	$y_{4,2}$	$y_{4,3}$	…	$y_{4,T}$
…	…	…	…	…
$y_{n,1}$	$y_{n,2}$	$y_{n,3}$	…	$y_{n,T}$



A Repeated Cross Section Design

Individuals nested in time.

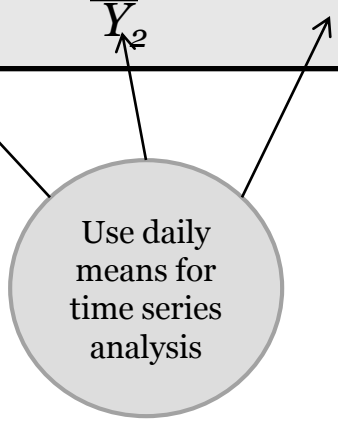
t=1	t=2	t=3	…	t=T
$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	…	$y_{1,T}$
$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	…	$y_{2,T}$
$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	…	$y_{3,T}$
$y_{4,1}$	$y_{4,2}$	$y_{4,3}$	…	$y_{4,T}$
…	…	…	…	…
$y_{n,1}$	$y_{n,2}$	$y_{n,3}$	…	$y_{n,T}$



$y_{1,1}$ indicates person 1 in wave 1 which occurs at t=1.

Option 1: Go Aggregate!

t=1	t=2	t=3	…	t=T
$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	…	$y_{1,T}$
$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	…	$y_{2,T}$
$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	…	$y_{3,T}$
$y_{4,1}$	$y_{4,2}$	$y_{4,3}$	…	$y_{4,T}$
…	…	…	…	…
$y_{n,1}$	$y_{n,2}$	$y_{n,3}$	…	$y_{n,T}$
$\overline{Y}_1$	$\overline{Y}_2$	$\overline{Y}_3$	…	$\overline{Y}_T$



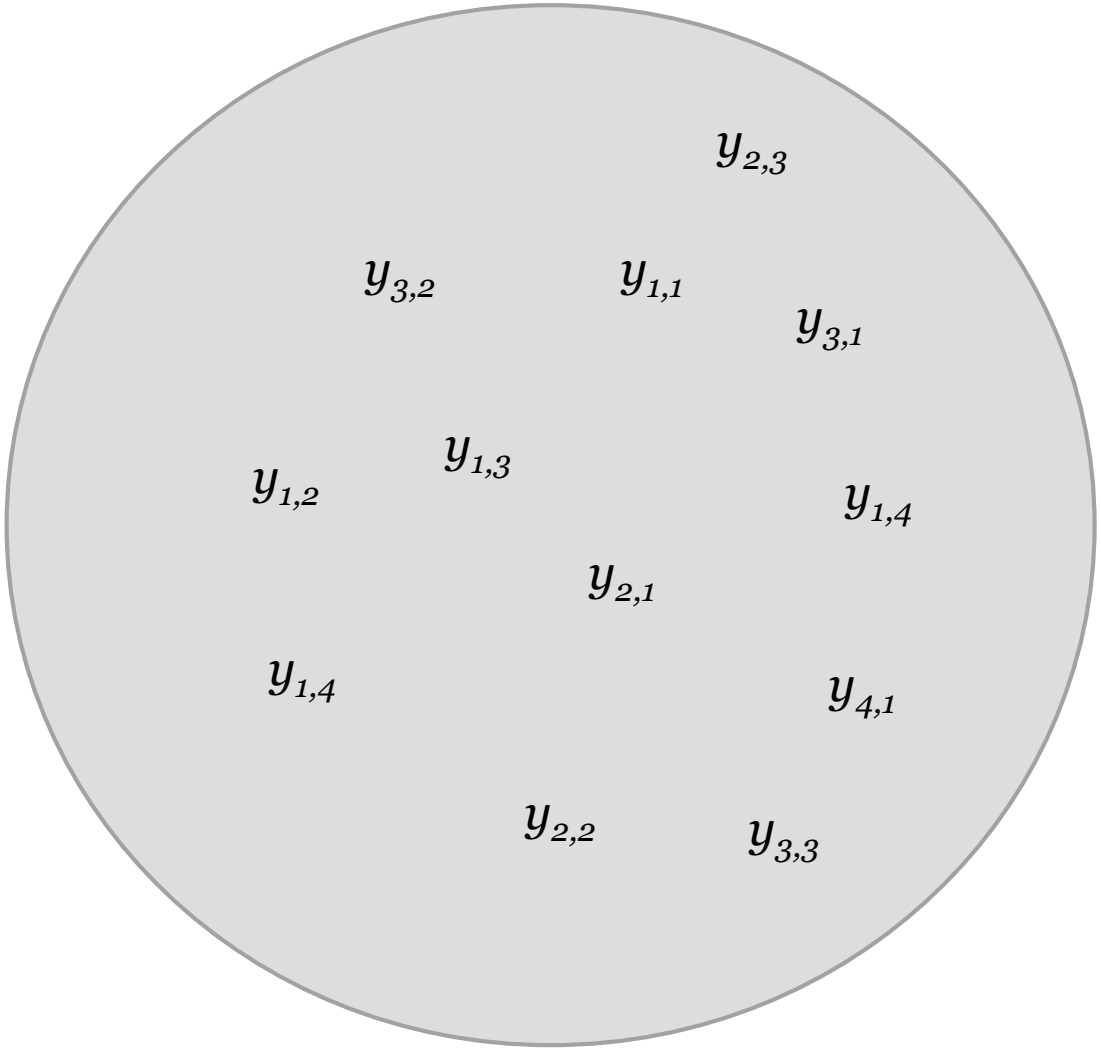
Reduces a sample of size N*T to one simply T long.

Traditional "long-t" Time Series

- Model $\overline{Y}_t$ using $\overline{X}_t$.
- Allows study of political dynamics.
- Overcome partisan bias at individual-level (Kramer 1983) and use averages in which bias cancels out.
- Allow use of time series methods like cointegration and time-varying parameters.
- Many key pieces to understanding public opinion over time.

- Examples that begin with RCS and create time series:
  - Mackuen, Erikson, and Stimson (1989; 1992). Gallup Polls.
  - Box-Steffensmeier, DeBoef and Lin (2004). CBS/NYT Polls.
  - Clarke, Stewart, Ault and Elliott (2005). Michigan's Survey of Consumers.
  - Johnston, Hagen, and Jamieson (2004). NAES.
  - Clarke and Lebo (2003). British Gallup.

Another Option: Naïve Pooling



Throw all the cases in together and ignore the time component.

E.g. Romer (2006); Moy, Xenos, and Hess (2006); Stroud (2008).

Confined to cross-sectional hypotheses – no dynamics.

Autocorrelation in a True Panel

t=1	t=2	t=3	…	t=T
$\varepsilon_{1,1}$	$\varepsilon_{1,2}$	$\varepsilon_{1,3}$	…	$\varepsilon_{1,T}$
$\varepsilon_{2,1}$	$\varepsilon_{2,2}$	$\varepsilon_{2,3}$	…	$\varepsilon_{2,T}$
$\varepsilon_{3,1}$	$\varepsilon_{3,2}$	$\varepsilon_{3,3}$	…	$\varepsilon_{3,T}$
$\varepsilon_{4,1}$	$\varepsilon_{4,2}$	$\varepsilon_{4,3}$	…	$\varepsilon_{4,T}$
…	…	…	…	…
$\varepsilon_{n,1}$	$\varepsilon_{n,2}$	$\varepsilon_{n,3}$	…	$\varepsilon_{n,T}$

Correlated due to factors specific to time-point *t*
$\text{corr}(\varepsilon_{i,t}, \varepsilon_{j,t}) \neq 0$
Thus, solutions like fixed effects and PCSE.

Correlated due to factors specific to individual *i*
$\text{corr}(\varepsilon_{i,t}, \varepsilon_{i,t+1}) \neq 0$
Thus, solutions like lagged dependent variables and differencing.

What changes with RCS?

- Importantly, the range of solutions.
  - can't use a lagged dependent variable since $y_{i,t-1}$ doesn't appear in the data set.
  - can't difference the dependent variable for the same reason.
  - And, even if you could do either of the above, the methods may be insufficient to account for between wave memory.
  - Panel Corrected Standard Errors premised on a true panel and doesn't solve bias if it exists.
  - Fixed effects don't solve autocorrelation in either direction.
- Does the problem of autocorrelation go away since each observation appears only once?
  - NO!

Autocorrelation in a Repeated Cross Section Design

t=1	t=2	t=3	…	t=T
$\varepsilon_{1,1}$	$\varepsilon_{1,2}$	$\varepsilon_{1,3}$	…	$\varepsilon_{1,T}$
$\varepsilon_{2,1}$	$\varepsilon_{2,2}$	$\varepsilon_{2,3}$	…	$\varepsilon_{2,T}$
$\varepsilon_{3,1}$	$\varepsilon_{3,2}$	$\varepsilon_{3,3}$	…	$\varepsilon_{3,T}$
$\varepsilon_{4,1}$	$\varepsilon_{4,2}$	$\varepsilon_{4,3}$	…	$\varepsilon_{4,T}$
…	…	…	…	…
$\varepsilon_{n,1}$	$\varepsilon_{n,2}$	$\varepsilon_{n,3}$	…	$\varepsilon_{n,T}$

Still correlated due to factors specific to time-point *t*

But, is $\varepsilon_{1,1}$ still too correlated with $\varepsilon_{1,2}$ even when they aren't the same individuals?

Absolutely!

Re-phrase that last question:

- Do we expect correlations between $\bar{Y}_t$ and $\bar{Y}_{t+1}$?
- Sure. Dozens of important papers establish the need to account for memory in time series measured at the aggregate level
  - Box-Jenkins techniques introduced to political science in the 1970s suggest a (p,q) ARMA model.
  - Clarke and Stewart (1994) introduce non-stationarity to the study of aggregated public opinion time series. Suggest $(p,0/1,q)$ ARIMA models.
  - Box-Steffensmeier and Smith (1996) introduce fractional integration and (p,d,q) ARFIMA models.
  - The applicability of fractional integration and ARFIMA to political data created from RCS has been shown in:
    - Lebo, Walker, and Clarke (2000): Presidential Approval, Macropartisanship, and Policy Mood.
    - Box-Steffensmeier and Tomlinson (2000): ICS and Congressional Approval.
    - Byers, Davidson, and Peele (2000): Approval and party support in many European democracies.
    - Box-Steffensmeier, DeBoef, and Lin (2004): The Gender Gap.
    - Clarke and Lebo (2003): British popularity, vote intentions, PM approval.
    - Box-Steffensmeier and DeBoef (2001): Micro-ideology.
    - Treisman (2011): Leader support in Russia.

And…

- If $\bar{Y}_t$ and $\bar{Y}_{t+1}$ are correlated, then $\varepsilon_{i,t}$ is still correlated with $\varepsilon_{j,t+1}$ more than $\varepsilon_{i,t}$ is correlated with $\varepsilon_{k,t+2}$.
- This is true since observations in each time point are dispersed around a mean correlated with the mean of the adjacent time-point.
- Put another way, $\bar{Y}_t = \mathrm{E}(y_{i,t})$ and $\bar{Y}_{t+1} = \mathrm{E}(y_{i,t+1})$.

If corr($\bar{Y}_t$ and $\bar{Y}_{t+1}$)≠0
then corr(E($y_{i,t}$), E($y_{i,t+1}$)) ≠0 and
corr(E($\varepsilon_{i,t}$), E($\varepsilon_{i,t+1}$)) ≠0.

How do we deal with these two types of autocorrelation?

- Available methods don't provide a solution.
  - Cannot difference, cannot use lagged dependent variable.
  - PCSEs do not solve the problem.
  - Fixed effects, random effects, special effects, cannot get rid of the autocorrelation.

Also, (an old question) how do we choose a level of analysis?

- Do we cut out a wealth of information and study aggregate time series?
  - Many defenders of this: Kramer (1983); MES (1989).
  - This is one way to solve autocorrelation problems – we know how to deal with it at the aggregate level.
- Or, do we ignore dynamics?
  - Throw everyone together and use cross-sectional techniques.
  - Or use PCSTS methods that allow clustering of data without estimation of parameters at the aggregate level.
- Let's do both aggregate- and individual-level.

Autoregressive Fractionally Integrated Moving Average Multi-Level-Model on Repeated Cross Sectional Data

- Or: ARFIMA-MLM
- Think about level-1 units (e.g., people) situated in level-2 structures (e.g., days/months/years).
- MLMs have been used in PCSTS (Beck and Katz 2007; Beck 2007; Shor, Bafumi, Keele and Park 2007). But these solutions either ignore autocorrelation or attempt to fix it with a lagged dependent variable (which we don't have in RCS at the individual-level).
- The MLM relies on the assumption that errors are both spatially and temporally independent. So have to deal with autocorrelation *first*.
- Our solution works for PCSTS but is especially useful for RCS and has less competition there than it does in the PCSTS toolkit.

Key Aspects

- Individual observations are embedded within multiple, sequential time-points.
- Retrieve estimates at the individual-level *and* at the aggregate level.
- Allows use of variables that vary only *within* cross-sections and some that vary between cross-sections (e.g., unemployment rate).
- Box-Jenkins and fractional differencing techniques can control for autocorrelation at level-2. (Box and Jenkins 1976; Box-Steffensmeier and Smith 1996, 1998; Lebo, Walker and Clarke 2000; Clarke and Lebo 2003).
- Introduce *Double Filtering* to clean up two kinds of autocorrelation.

Double Filtering - The Math

- Begin with level-2 (aggregate):

$$(1-L)^d \overline{Y}_t = \frac{(1-\theta_q L^q)}{(1-\phi_p L^p)} \varepsilon_t \qquad (1)$$

The ARFIMA equation for long-t time series.

p autoregressive parameters

q moving average parameters

difference d times.

Estimate to find correct values of (p,d,q) to create white noise residuals (ε_t).

Simplifying (1)

- Where d=1, model simplifies to: $\Delta\bar{Y}_t = \frac{(1-\theta_q L^q)}{(1-\phi_p L^p)}\varepsilon_t$. (ARIMA)

Where d=0, model simplifies to: $\bar{Y}_t = \frac{(1-\theta_q L^q)}{(1-\phi_p L^p)}\varepsilon_t$. (ARMA)

Choose one based on stationarity tests and direct estimation of d.

More cross-sections will allow better estimate of d.

We should prefer ARFIMA since DGP of series are likely fractionally integrated, even if we don't have data to prove it.

Where t is short, getting d is difficult and ARMA and ARIMA are just approximations of ARFIMA.

First filter: make a noise model for level-2

$$\overline{Y}_t^* = (1-L)^d \overline{Y}_t \times \frac{(1-\phi_p L^p)}{(1-\theta_q L^q)}$$

$\overline{Y}_t^*$ is just the residuals from $\overline{Y}_t$ regressed on its noise model – a series that is both stationary in the long-run and free from autocorrelation due to short-run autoregressive and moving average processes.

Next, do this for X_ts and for any variables, Z_t, that vary over time but not within time-points.

With $\overline{Y}_t^*$, $\overline{X}_t^*$, and Z_t^*, level-2 is cleansed of autocorrelation.

Second filter: individual-level deviations from daily aggregates' *noise model*.

If there is autocorrelation at level-2, then centering around daily means would not solve autocorrelation at level-1.

That is, if $\overline{Y}_t$ and $\overline{Y}_{t-1}$ are correlated, then centering level-1 observations y_{it} and y_{it-1} around them will still leave serially correlated errors in those level-1 units.

Instead, we center the level-1 observations around the value of the noise model at time t.

E.g., center y_{it} around $\overline{Y}^*_t$.

- It is important to note that using day-level variables (e.g., $\overline{X_t}$), and lagged day-level variables (e.g.,$\bar{X}_{t-1}$), may not be enough to properly control for autocorrelation.

So long as the noise models at level-2 are appropriate, this second filter will create level-1 versions of the data that are also free of autocorrelation.

That is,

$$y_{it}^{**} = y_{it} - \overline{Y}_t^* \qquad (5)$$

and

$$x_{it}^{**} = x_{it} - \overline{X}_t^* \qquad (6)$$

creates x_{it}^{**} and y_{it}^{**}, individual-level observations that vary both *within* and *between* days but are, via double filtering, cleansed of autocorrelation and safe to be included in a multi-level regression.

Now we can estimate the MLM

- Two level equation.
- Level 2 equation can include covariates that vary only between days and those that vary within and between.
  - $\overline{Y}_t^* = \alpha_2 + \beta_2 \overline{X}_t^* + \gamma Z_t^* + u_{2t}$ (7)
- The level-1 equation provides the model of within variation:
  - $Y_{it}^{**} = \alpha_1 + \beta_1 X_{it}^{**} + u_{1it}$ (8)

Or, estimate (7) and (8) together.

- $y_{it}^{**} = \alpha_1 + \beta_1 x_{it}^{**} + u_{1it} + \beta_2 \overline{X}_t^{*} + \gamma Z_t^{*} + u_{2t}$
- This gives us aggregate effects with β_2 and individual effects with β_1.

Points of flexibility

- Depending upon the length of T, one might estimate an ARMA, ARIMA, or ARFIMA model to create an appropriate noise model. E.g., an AR(1) might be best for cumulative NES.
- If (0,0,0) the model reduces to mean centering.
- Can include time varying coefficients and inclusion of level-1 data, W_{it}, not in all waves: $Y_{it}^{**} = \alpha_{1t} + \beta_{1t}X_{it}^{**} + \delta_t W_{it} + u_1$.
- Applicable to PCSTS, but more tools there.
  - PCSE, Differencing, Lags, etc.

Monte Carlo Analyses

- We know the added value of estimating cross-sectional and dynamic parameters together.
- But has double filtering solved the two directions of autocorrelation?
- We expect that the greater the time dependence in level-2, the greater the degree of bias in coefficients.
- If observations at time *t* are more correlated with one another than with observations at *t*+*s*, this is a problem of clustering in the data; the errors will not be independent and the standard errors will be incorrect.

Monte Carlo Setup

- We generate level-2 time series with varying levels of memory (serial correlation).
- Each aggregate value gives us a mean for a distribution from which to draw individual-level data.
- We do this for Xs and Ys.
- We also generate data with no serial correlation to serve as a baseline for comparison.

Monte Carlo Comparisons

- We test the statistical properties of six approaches:
  - OLS (naïve) pooling all the data
  - OLS with day-level lag (OLS-LDV)
  - OLS single filtering (OLS-ARFIMA) without level-2
  - Multi-level model with time-varying intercepts (MLM)
  - MLM with day-level lag (MLM-LDV)
  - Our double filtering method (MLM-ARFIMA)

Simulation Expectations I - OLS

- Naïve pooling should lead to bias and inefficiency since it ignores clustering.
- A lagged dependent variable, $\overline{Y}_{t-1}$, will lead to unbiased and efficient estimates if it's enough to clean up the autocorrelation at level-2. (But there is a ton of literature showing that it never is!)
- ARFIMA methods should minimize bias and inefficiency.

Simulation Expectations II - MLM

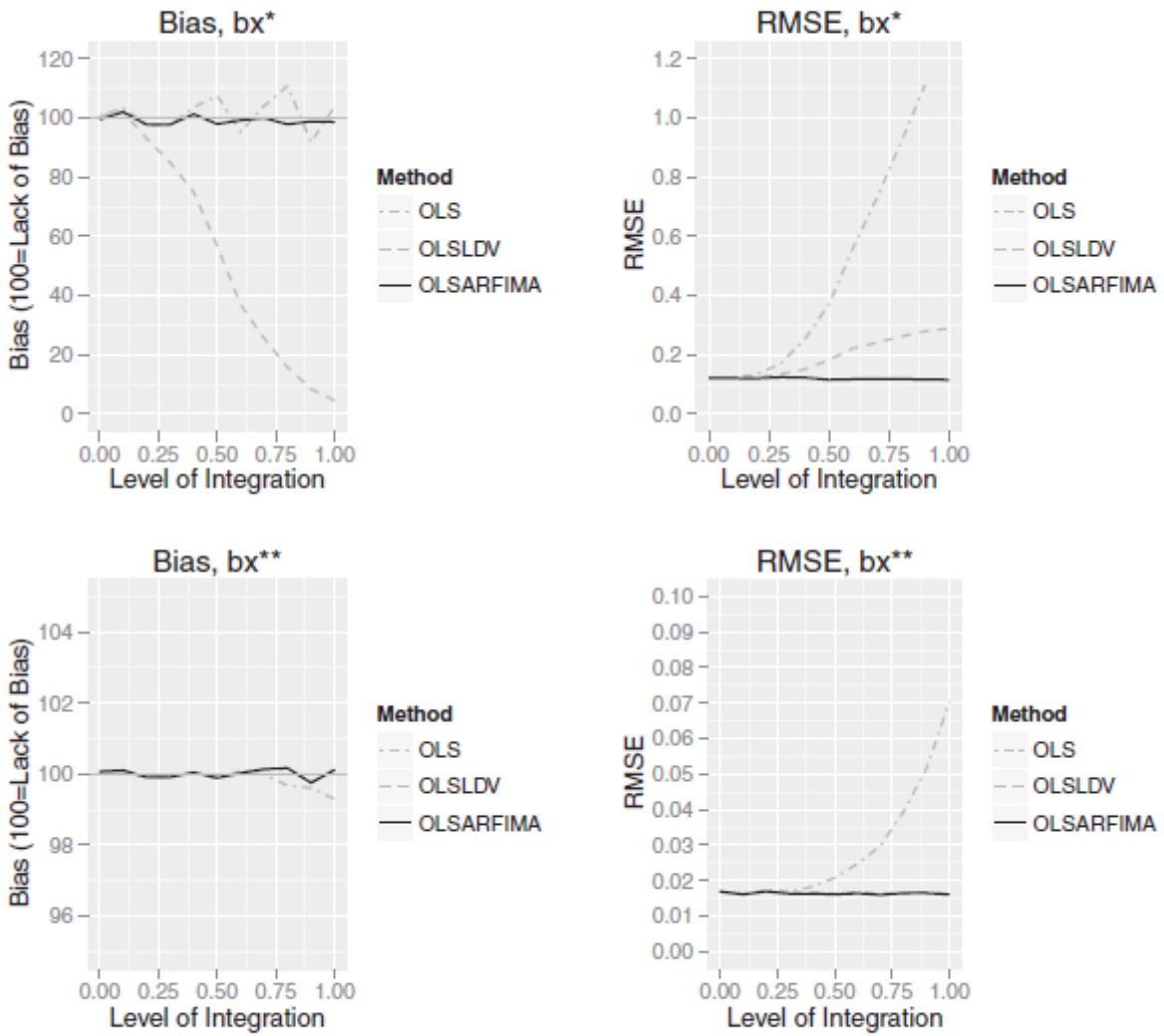
- The MLM approaches should be an improvement over OLS by accounting for the clustering in the data.
- However, an assumption of the MLM is that level-2 errors will be independently distributed, which is violated insofar as ARFIMA properties are unaccounted for at level-2.
- Simple MLM, will produce biased and inefficient estimates as *d* increases.
- Similarly, MLM-LDV – the multilevel model with a level-2 lagged dependent variable -- will produce estimates that are biased downward as *d* increases. This occurs as the level-1 units are not filtered at all.
- MLM-ARFIMA should fix everything, we hope.

Simulation Results

- We calculate comparative bias as $(\bar{\theta} - \theta)/\theta$ where $\bar{\theta}$ and θ are the estimated and true values, respectively.
- RMSE=$\sqrt{\frac{\sum(\bar{\theta}-\theta)^2}{n}}$, where n is the number of replications in each cell (i.e., 1,000). A small RMSE is preferred over a large RMSE, as it indicates less variation around the true population value.
- The standard errors calibrated to the size of the sample variation, which is "optimism" or "overconfidence" (Beck and Katz 1995; Shore et al 2007).
- $Optimism = 100 \times \sqrt{\frac{\sum_{l=1}^{1000}(\beta_l - \bar{\beta})^2}{\sum_{l=1}^{1000} SE\beta_l}}$

Over 100 means SE are smaller than they should be – we are too confident.

FIGURE 1 Bias and RMSE for OLS Coefficients



Note: For the OLS and OLS-LDV models, this is the coefficient for $\tilde{X}_t$. For the ARFIMA-OLS models, it is the coefficient for $\tilde{X}_t^*$. Lines in the bottom panels are all present but overlap.

TABLE 1 Optimism Index for Six Modeling Approaches

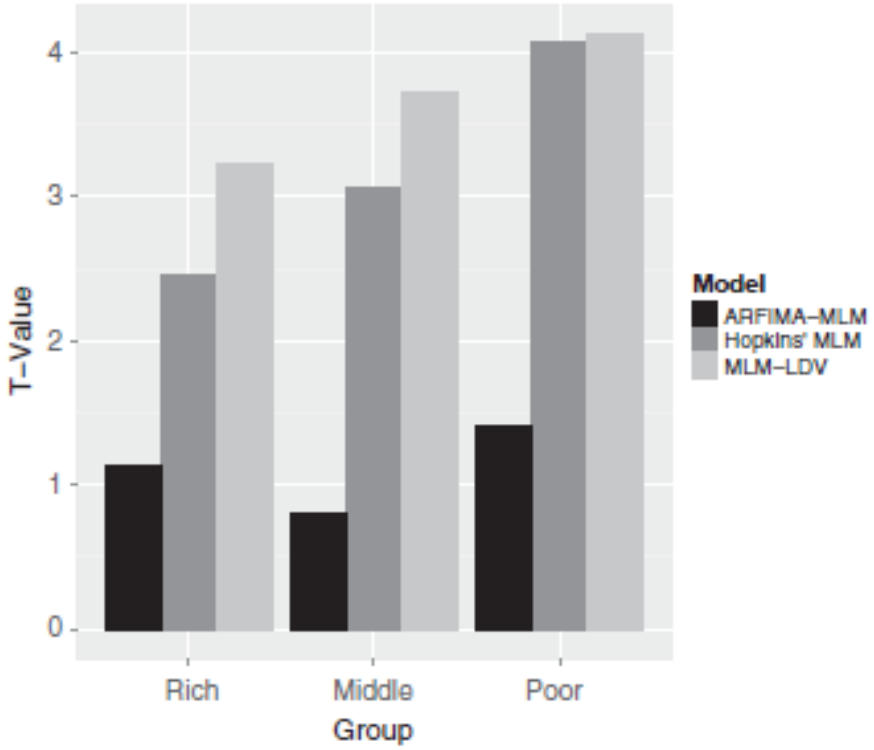
	Between-Day Effects ($b_{\bar{X}*}$)					
d	OLS	OLS-LDV	ARFIMA-OLS	MLM	MLM-LDV	ARFIMA-MLM
0	739	1896	737	104	267	104
0.1	783	1865	734	110	262	103
0.2	857	1630	726	118	228	92
0.3	1083	1436	753	145	200	106
0.4	1663	1244	746	214	172	105
0.5	2566	1056	700	312	146	98
0.6	4002	1603	713	456	221	100
0.7	5231	2433	713	565	336	100
0.8	6677	3747	714	698	521	101
0.9	8372	5895	710	856	824	100
1.0	8983	7770	699	907	1088	98

Note: For the OLS, OLS-LDV, MLM, and MLM-LDV models, these are based on the standard errors of coefficients for $\bar{X}_t$. For the ARFIMA-OLS and ARFIMA-MLM models, they are based on the standard error of the coefficient for $\bar{X}_t^*$.

Example 1: Hopkins (2012) *Whose Economy: Perceptions of National Economic Performance During Unequal Growth*

- 215,000 respondents nested in 388 months of Michigan Survey of Consumers data sets.
- In the aggregate, the series have been shown to have long memory. We find d=0.81 for poor respondents.
- "…Americans at all income levels weigh income growth at the low end in their responses" p.68
  - a level-2 conclusion, but problems at level-2.
- The key effect does not hold up once auto-correlation is accounted for.

FIGURE 3 The Effect of Income Growth for 20th Percentile of Income Using Three Modeling Approaches



Note: Each bar represents the *t*-statistics for the respective group's regression coefficient in three modeling approaches (darkest is ARFIMA-MLM, medium is MLM, and lightest is MLM-LDV).

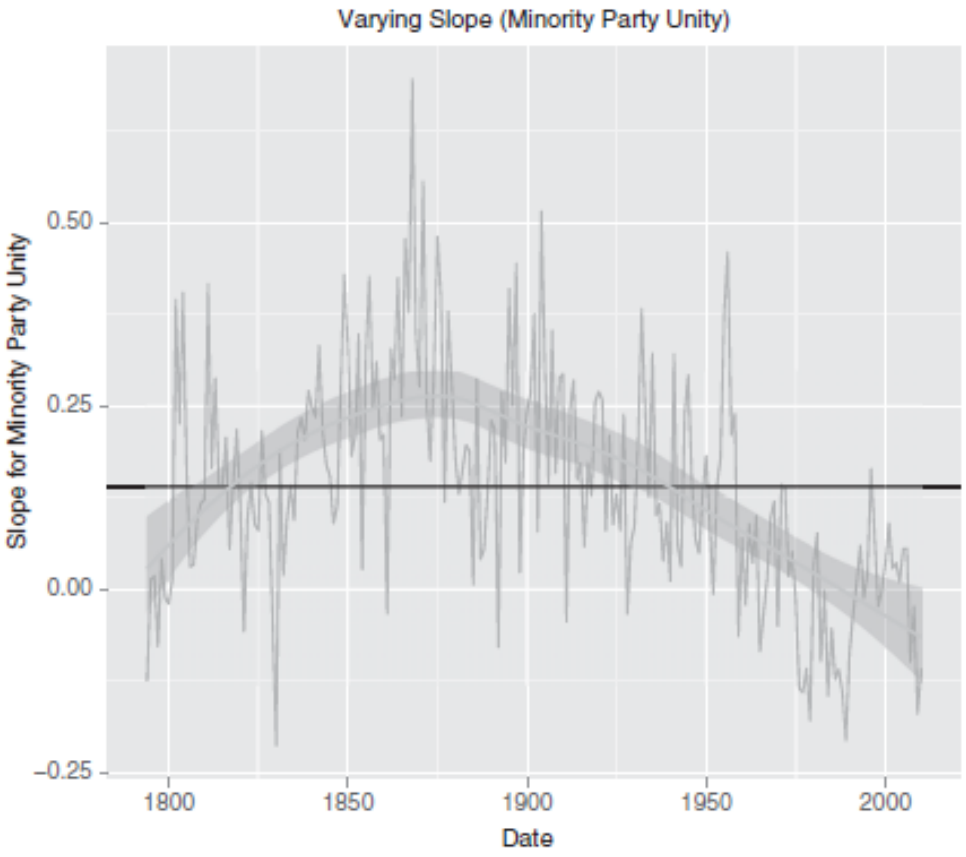
Example 2: Lebo, McGlynn, and Koger (2007) *Strategic Party Government*

- Original study is time series of 210 years.
- But there are 29,734 roll-call votes nested in those years.
- Again, data have been shown to be fractionally integrated and we have some variables that only vary at level-2.
- Time varying hypotheses possible.
- In our polarized era there is little variation in unity across votes.
- Not much within-Congress variation or interaction.

TABLE 2 An ARFIMA-MLM Model of Majority Party Voting Unity in the House of Representatives, 1789–2006 (Party Votes Only)

	β (SE)	*t*
Level-1 (within years) N = 29,599		
Intercept	1.390 (0.66)	2.11
Minority unity	0.128 (0.01)	23.15
Policy vote	–6.988 (0.32)	–21.86
Long session	–1.237 (1.37)	–0.90
First session	4.715 (1.33)	3.56
Level-2 (between years) T = 217		
Minority unity	0.557 (0.07)	7.67
Majority cohesion–NOMINATE 1	–44.461 (28.22)	–1.58
Majority cohesion–NOMINATE 2	–42.09 (16.20)	–2.60
Ideological distance–NOMINATE 1	–12.05 (14.68)	–0.82
Ideological distance–NOMINATE 2	1.936 (5.76)	0.34
Majority size	–0.504 (0.10)	–5.28
Error Correction Mechanism (t–1)	–0.226 (0.07)	–3.61
Number of votes/years	29,599/217	

FIGURE 4 The Time-Varying Effect of Minority Unity on Majority Unity, 1794–2006



Note: Coefficient and smoother for the roll-call-level effect of Minority Party Unity on Majority Party Unity.

Example 3: Kenski, Hardy, and Hall-Jamieson (2010) *The 2008 National Annenberg Study*

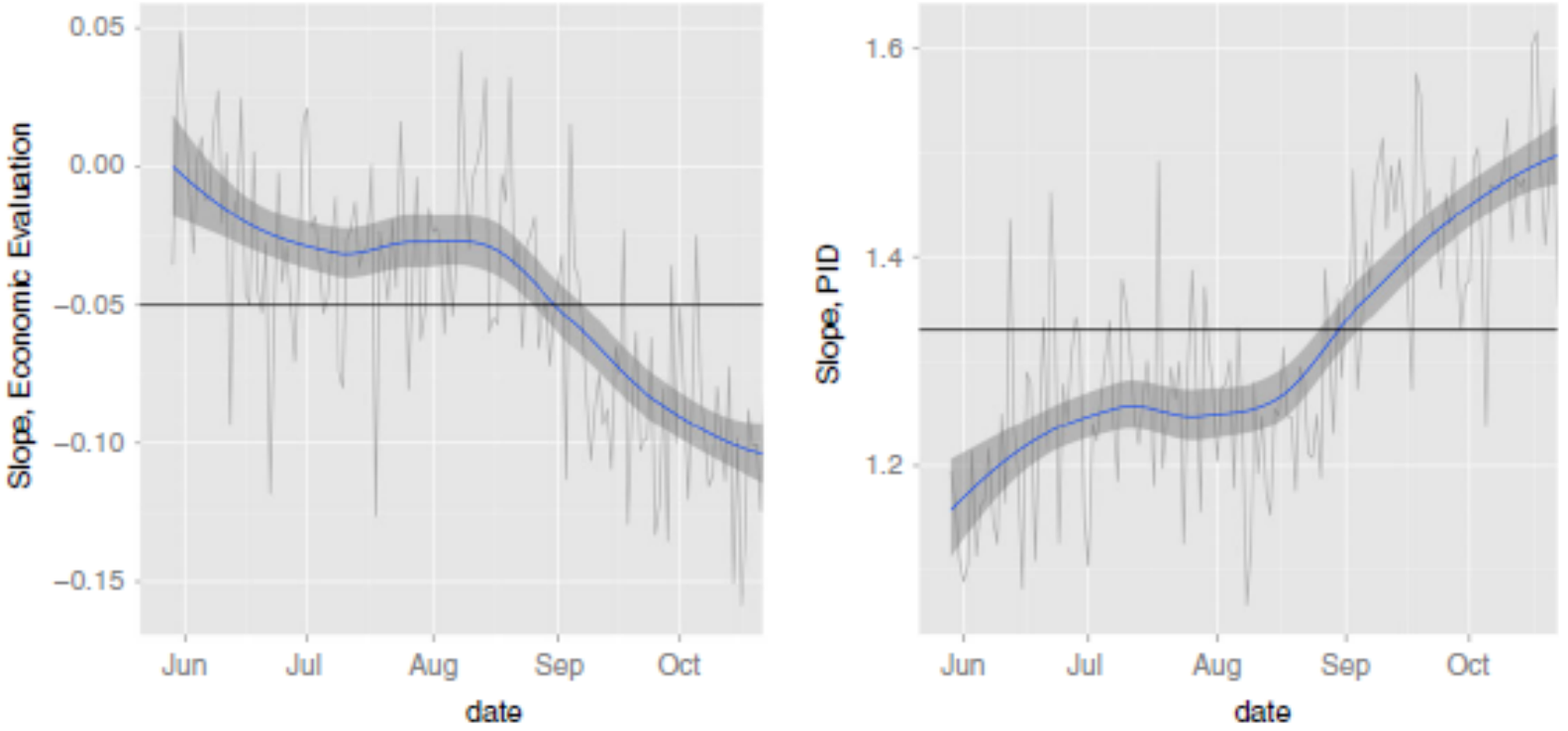
- Rolling cross-sectional design with some studies pooled and some time series.
- We model campaign effects of Evaluation of Obama – Evaluation of McCain.
- Our MLM has within and between effects.
- Economic effects are complicated and our MLM shows a lot of nuance.
- Time varying hypotheses of economic evaluations and party identification.
- Indeed, campaigns activiate and reinforce partisanship.

TABLE 3 An ARFIMA-MLM Model of Campaign Effects

	Within Day		Between Day	
	B (SE)	t	*B* (SE)	t
Fundamentals				
Intercept	–0.09 (0.11)	–0.77		
Party identification	0.13 (0.01)	10.09	–0.74 (0.63)	–1.16
Ideology (conservative)	–0.06 (0.02)	–3.28	–1.01 (0.78)	–1.29
Vote Bush in 2004	–0.00 (0.05)	–0.10	—	
Approve Bush	–0.36 (0.05)	–7.03	2.65 (2.44)	1.09
National economy	–0.11 (0.05)	–2.41	0.14 (2.02)	0.07
Personal economy	–0.02 (0.05)	–0.48	–4.67 (2.33)	–2.00
Sociodemographics				
Gender (female)	–0.03 (0.03)	–0.79		
Age (in years)	0.004 (0.001)	2.86		
Black	0.15 (0.07)	2.19		
Hispanic	0.16 (0.17)	0.97		
Education	–0.005 (0.007)	–0.70		
Income (in thousands)	0.0003(0.0004)	0.71		
Media				
Number of days saw campaign info (TV)	0.02 (0.008)	2.31	0.46 (0.30)	1.34
Number of days heard about campaign (radio)	0.02 (0.006)	3.25	0.19 (0.29)	0.68
Number of days saw info (newspaper)	0.008 (0.006)	1.34	0.35 (0.26)	1.35
Number of days saw info (Internet)	0.008 (0.006)	1.34	–0.49 (0.25)	–1.95
Campaign Messages				
Elect McCain is like reelecting Bush	0.65 (0.05)	13.87	0.47 (2.20)	0.21
McCain is too old	0.37 (0.04)	8.73	0.65 (1.72)	0.38
Obama's ideology (liberal)	–0.05 (0.02)	–2.33	0.60 (1.03)	0.59
Experience (McCain-Obama)	–0.14 (0.01)	17.29	0.23 (0.37)	0.63
Judgment (McCain-Obama)	–0.24 (0.01)	–29.90	0.64 (0.40)	1.58
Patriotic (McCain-Obama)	–0.12 (0.01)	–16.92	–0.57 (0.36)	–1.60
Values (McCain-Obama)	–0.36 (0.01)	–47.61	–1.35 (0.38)	–3.52

Note: For consistency, we present the random intercept model. The intra-class correlation is small (<0.01), and the OLS-ARFIMA model separating the within- and between-day effects yields equivalent results. The model is similar to Kenski, Hardy and Hall Jamieson's (2010), with several qualifications. Rather than vote choice, the dependent variable is Evaluation of Obama–Evaluation of McCain.
Source: 2008 NAES.

FIGURE 5 Slopes for Sociotropic Economic Evaluations and Party Identification over the 2008 Campaign



Limitations of ARFIMA-MLM

- For now, applies to a continuous dependent variables. We are working on the model for dichotomous variables at the individual-level.
- Length of T should be at least 40 to begin using ARFIMA instead of ARMA or ARIMA.
  - Paper says 50 but some estimators are pretty good at 40.
  - For example, cumulative NES is too short.

R Package ArfimaMLM

- Facilitates implementation of ArfimaMLM in R
- Performs procedures and analyses described previously based on a single function call
  - uses fractal, fracdiff, mle4
- 3 major functions:
  1. arfimaMLM
  2. arfimaOLS
  3. arfimaPrep

R Package ArfimaMLM - Example I

```
arfimaMLM(formula, data, timevar
          , d = "Hurst", arma = NULL
          , ecmformula = NULL, decm = "Hurst"
          , drop = 5, report.data = TRUE, ...)

arfimaMLM(y.ydif ~ x1.xdif + x1.fd + x2 + z1.fd + z2.fd
          + (1 | time)
          , data = data, timevar = "time", ...)
```

R Package ArfimaMLM - Example II

```
###################################
Fractional Differencing Parameters:

    Method   Estimate
y    Hurst 0.41713434
z1   Hurst 0.01679005
z2   Hurst 0.10887414
ecm    GPH 0.55248930


##################
Summary OLS Model:

Call:
lm(formula = formula, data = new$data.merged)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0890 -0.9588  0.0402  1.0017  5.7285

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  0.0613607  0.0067106    9.144   <2e-16 ***
x1.xdif      0.2033355  0.0033485   60.724   <2e-16 ***
x2          -0.0499045  0.0001677 -297.662   <2e-16 ***
z1.fd        0.2076489  0.0071888   28.885   <2e-16 ***
z2.fd       -0.0846972  0.0065389  -12.953   <2e-16 ***
```

R Package ArfimaMLM

- CRAN:
  - http://cran.r-project.org/web/packages/ArfimaMLM/
- GitHub (dev version + further documentation):
  - https://github.com/pwkraft/ArfimaMLM

Thank you. Questions?