Expanding the Frontier of Economic Statistics Using Big Data: A Case Study of Regional Employment

Abe Dunn (BEA), Eric English (Census, Presenting), Kyle Hood (BEA), Lowell Mason (BLS, Presenting), Brian Quistorff (BEA, Presenting)

FESAC

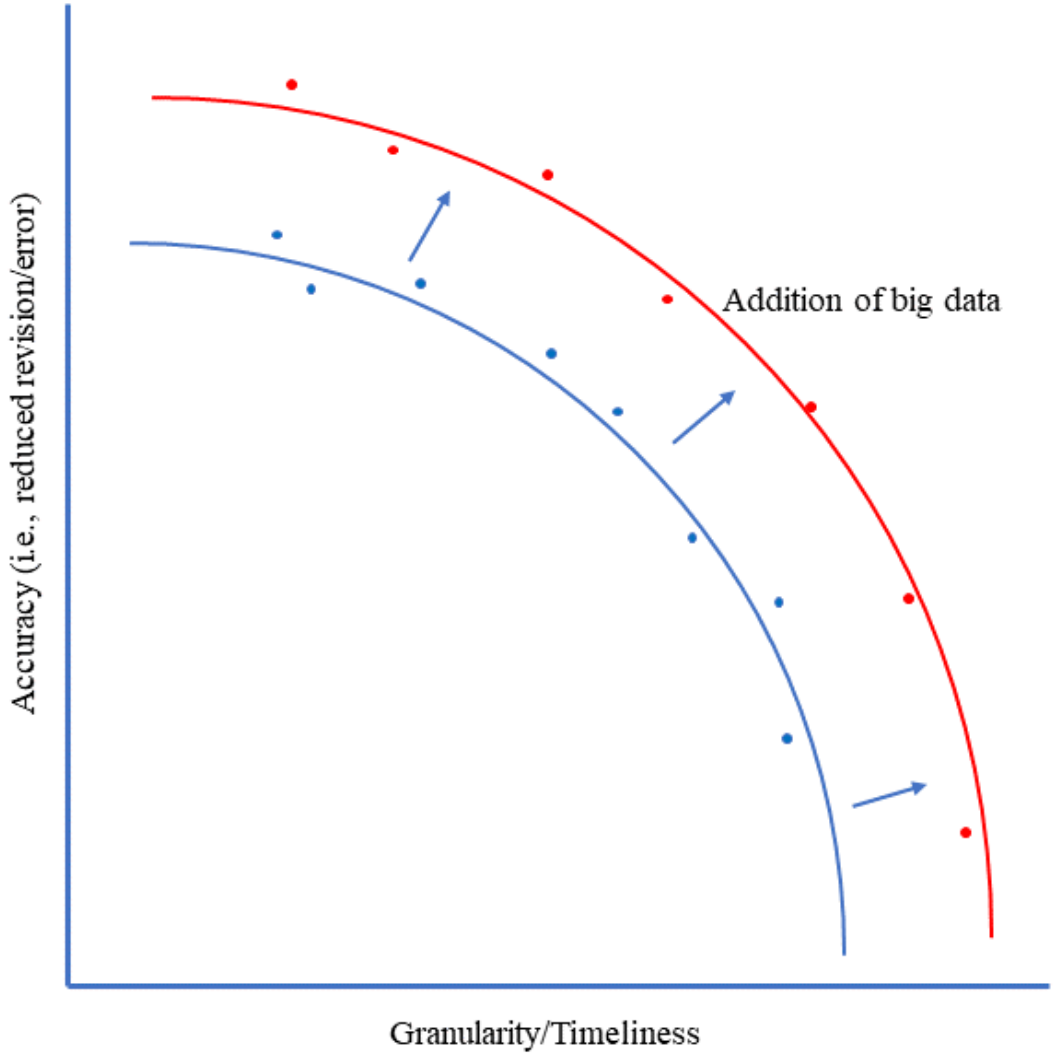
June 14, 2024

Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the U.S. Census Bureau, the Bureau of Economic Analysis, or the Bureau of Labor Statistics. This paper does not use any confidential Census Bureau data.

Motivation and background

- There's been a dramatic increase in the amount of data available
- Simultaneously, there has been a decline in survey response rates
- Can statistical agencies:
  - Leverage this data to produce timely, high-frequency and granular statistics
  - If so, are they an improvement?
- Third party data have issues
  - Non-random sample
  - Unstable sample
- How do statistical agencies currently weigh the advantages and disadvantages of official data sources and data releases?

Production possibility frontier for economic measurement



What we do

Application to county- and state-by-industry employment

1. Assess current error tolerance from official sources (CES)
2. Use alternative payroll data to produce alternative estimates (state- and county-by-industry level)
3. Evaluate the accuracy (i.e., error/revision) of these alternative estimates
4. Compare accuracy to tolerance levels from official sources to determine the value of the new estimates

What we find

- For currently targeted estimates, solely using alternative payroll data produces errors higher than tolerance levels → combine with CES
- Alternative payroll data (combined with CES) shows some improvement in the accuracy of state-level employment series
- Alternative payroll data (combined with CES) produces new county-level estimates in line with tolerance level for this amount of granularity
- We apply new estimates to examine resource allocation around the pandemic → We find improved timeliness/accuracy in identifying worst hit counties

Data

- Quarterly Census of Employment and Wages (QCEW)
  - Released quarterly, 5 months after the end of the quarter
  - County by industry level
- Current Employment Statistics (CES) survey
  - National by industry level - ~3 weeks after the 12th
  - State/MSA by industry level - ~5 weeks after the 12th
- Payroll processor
  - Paycheck level aggregated to county by industry
  - Lag/period is customizable
- All series measure employment on the 12th of the month
- Period: 01/2017 to 06/2021
- Main focus is state and county by industry series

Payroll processor data

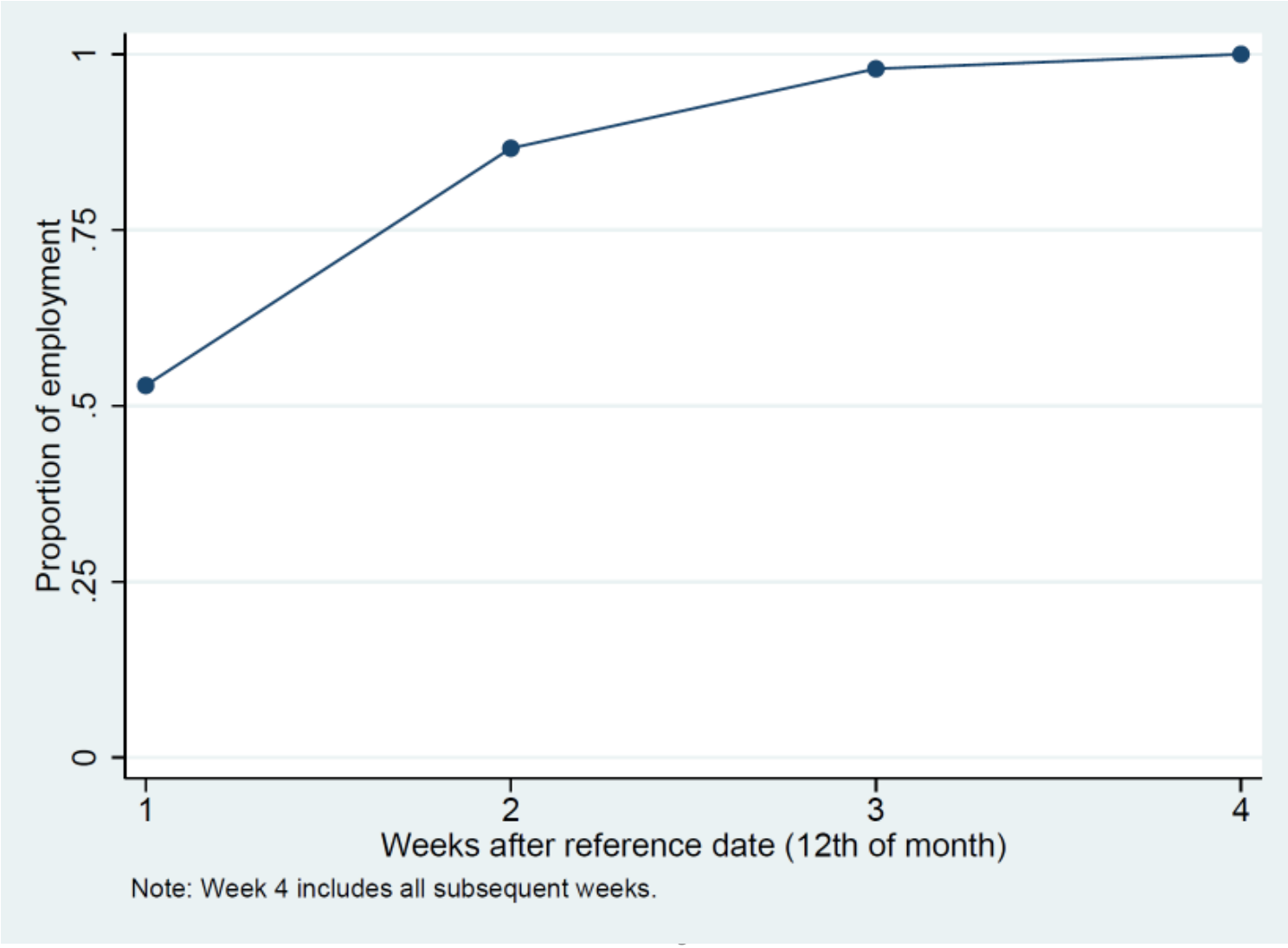
- Over 5% market share
- Mostly small- and medium-sized businesses
- Includes 2-digit NAICS
  - Coverage across all industries
- Micro data, aggregated monthly
  - Open to changes in data delivery

Payroll data processing

- All employers or continuing units?
  - Payroll company has attracted many new clients → use continuing units

- Paycheck level data
- Payroll processor gets data at the end of each pay period
  - Most employees are paid weekly, bi-weekly or monthly
- More data comes in as time passes
  - How long do we want to wait to collect the data?

Payroll frequency and payroll timing



Comparison of input data: weeks from reference period (state)

	(1)	(2)	(3)
Pay Emp Gr (cont; t=1)	0.468***		
	(37.92)		
Pay Emp Gr (cont; t=2)		0.603***	
		(43.81)	
Pay Emp Gr (cont; t=3)			0.612***
			(44.32)
Observations	41,924	41,924	41,924
$1 - R^2$ OOS (rolling)	0.509	0.389	0.374
$1 - R^2$ OOS (k-fold)	0.383	0.333	0.322
MAE OOS (rolling)	0.0130	0.0121	0.0120
MAE OOS (k-fold)	0.0114	0.0109	0.0108

Framework for evaluating statistics

1. Assess tolerance of existing statistics across types of granularity
   - Use Mean Absolute Error (MAE) for presentation
     - CES initial release vs. final QCEW
   - Types of granularity:
     - Geography: County, MSA, State or National
     - Industry: All industries, 2-digit and 3-digit NAICS

2. Produce new or improved statistics (i.e., county/state-level employment growth)

3. Calculate errors and compare to tolerance levels

Error Tolerance (CES initial vs QCEW)

Geo	NAICS digits	Mean Abs. Error	N
USA	All Indus.	0.0014	92
USA	2d	0.0032	2,024
USA	3d	0.0048	6,900
State	All Indus.	0.0030	4,876
State	2d	0.0086	86,712
State	3d	0.0111	136,207
MSA	All Indus.	0.0068	29,024
MSA	2d	0.0119	157,528
MSA	3d	0.0129	94,172
Counties	All Indus.	0.0102	179,114
Counties	2d	0.0180	3,124,127
Counties	3d	0.0214	4,844,287

Prediction equation

We wanted to compare (unmodelled) CES with models that include CES and Payroll data.

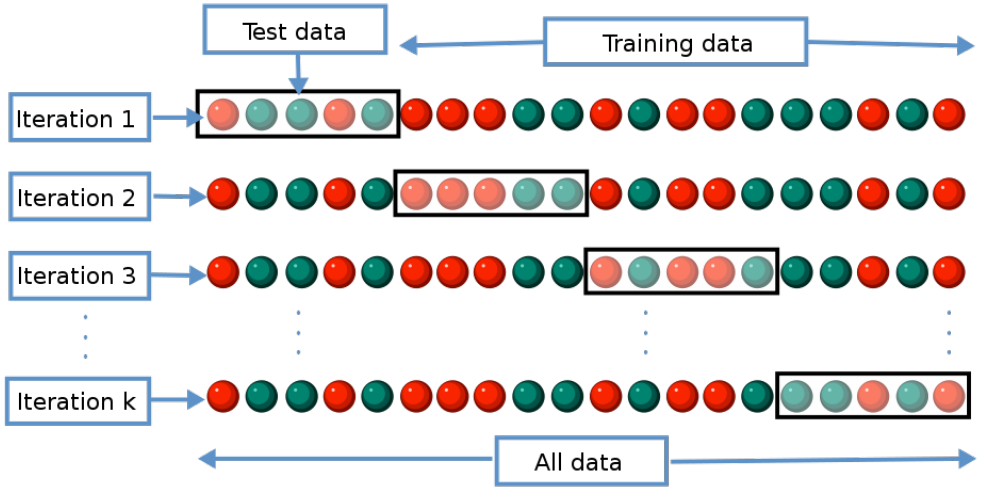
Monthly QCEW Employment Growth = *ƒ (CES, Payroll) + Error*

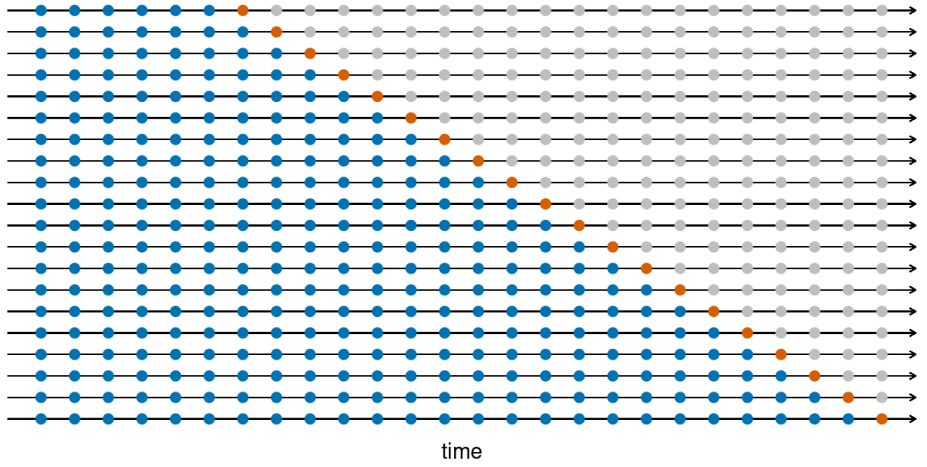
- *ƒ ()* — linear functional form
- We try various specifications (CES only, Payroll only, CES + Payroll)
- We also include some characteristics of the Payroll data
  - E.g. Payroll coverage

Evaluation criteria

Two out-of-sample approaches:

- K-fold cross validation:
  - Advantage – Uses all time-series variation in the data
  - Disadvantage – Does not match how estimation will work in practice
- Rolling one-step-ahead estimation:
  - Advantage – Closer to how estimation would work in practice
  - Disadvantage – Does not use the full time series variation

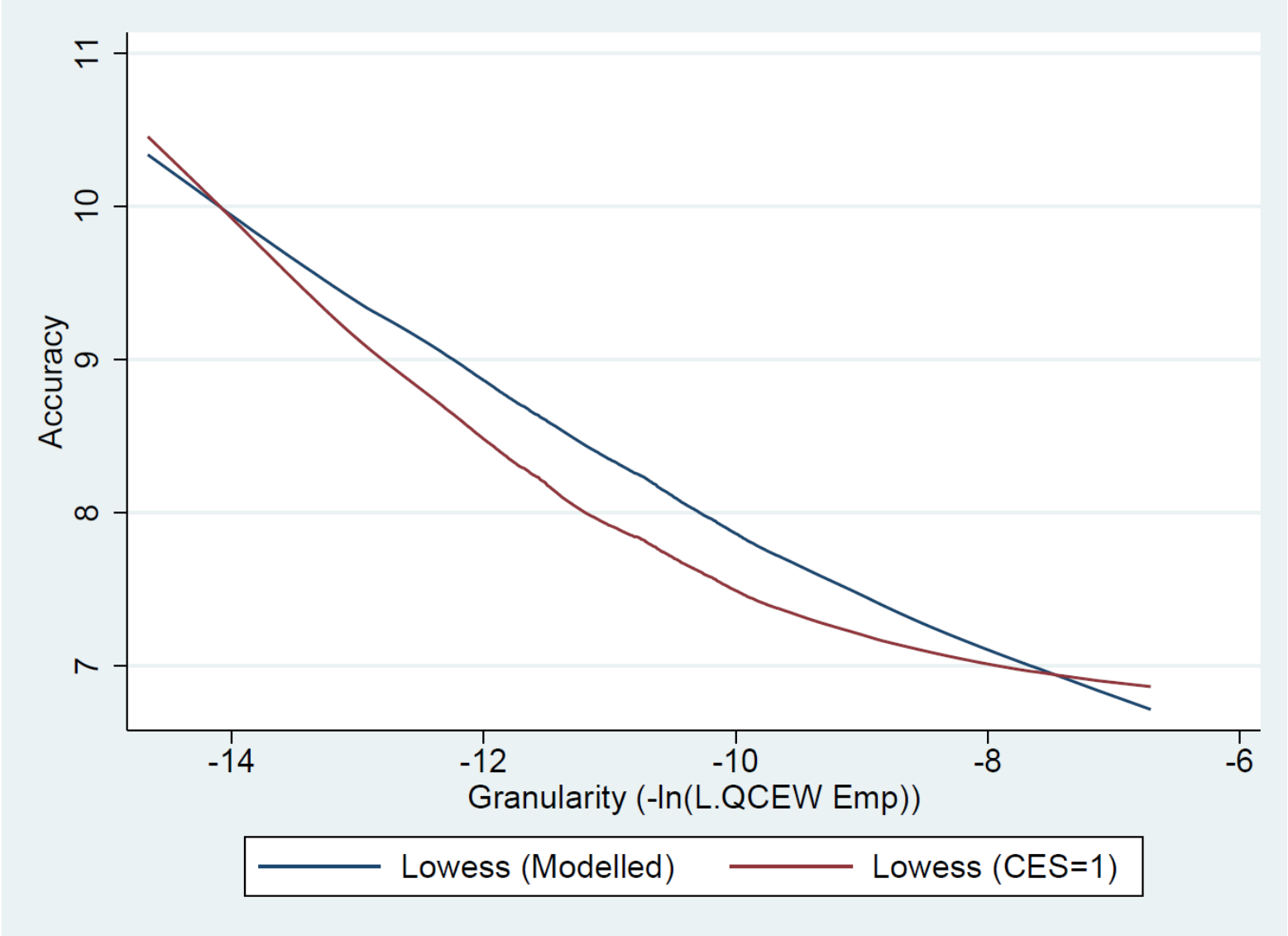




Improved state-level estimates?

	CES (1)	Pay (2)	CES + Pay (3)	CES + Pay (4)
CES Emp Gr	1		0.680***	0.607***
	(.)		(258.44)	(217.93)
Pay Emp Gr (cont; t=3)		0.612***	0.213***	0.140***
		(44.32)	(89.42)	(15.06)
Pay (cont; t3) Coverage x Emp GR				2.008***
				(27.72)
Pay Emp Gr (st-Agg; t3)				0.0369***
				(10.52)
Observations	41,924	41,924	41,924	41,924
$1 - R^2$ OOS (rolling)	0.177	0.374	0.176	0.154
$1 - R^2$ OOS (k-fold)	0.179	0.322	0.151	0.136
MAE OOS (rolling)	0.00876	0.0120	0.00810	0.00777
MAE OOS (k-fold)	0.00853	0.0108	0.00759	0.00733

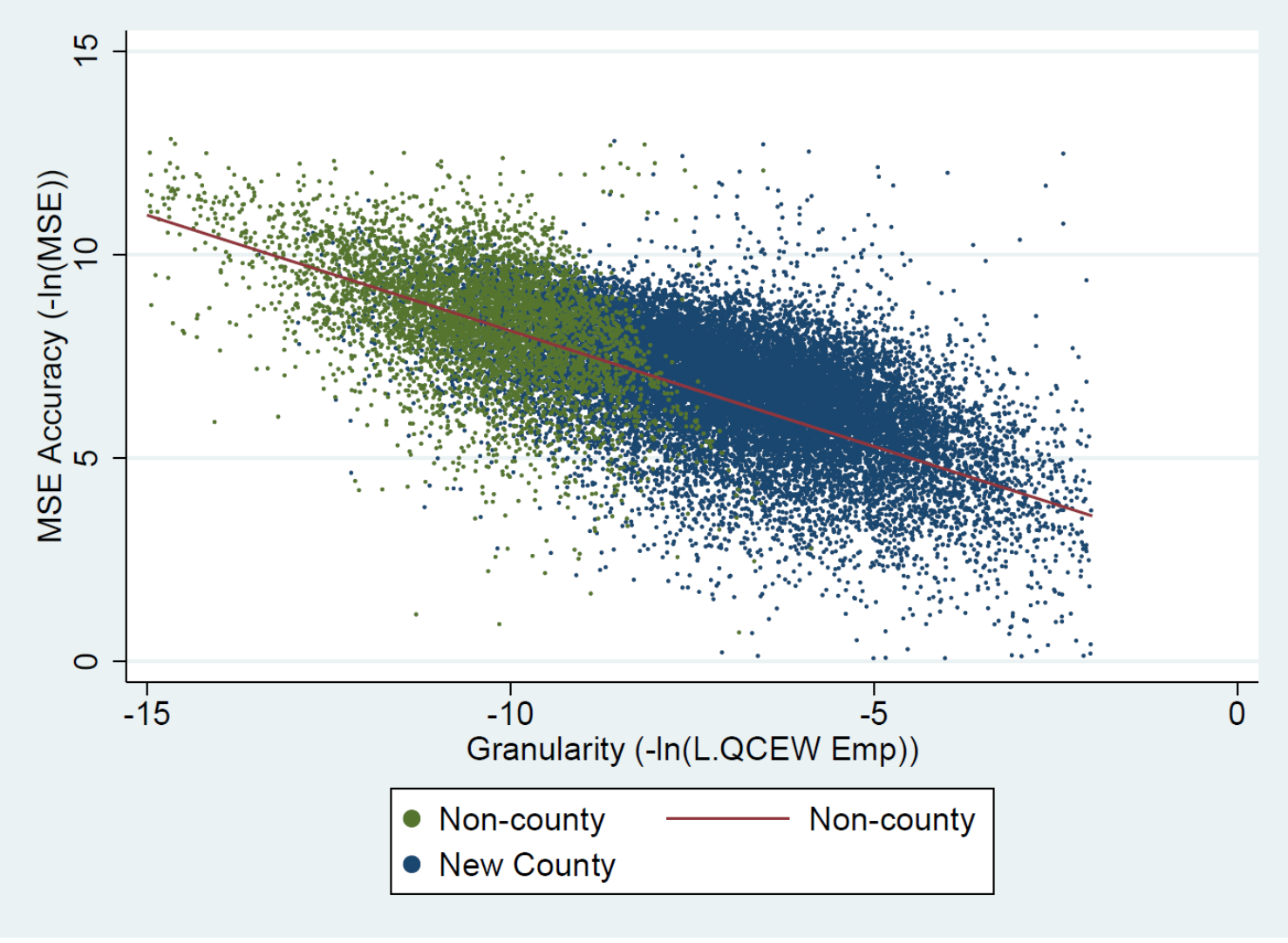
Expanding the PPF for state estimates



New county-level estimates?

	CES (1)	Pay (2)	Pay (3)	CES + Pay (4)	CES + Pay (5)
CES Emp GR (MSA or State)	1			0.769***	0.667***
	(.)			(600.29)	(485.81)
Pay Emp Gr (cnty-naics2; t3)		0.0986***	0.0256***	0.0274***	0.00749***
		(46.98)	(10.91)	(13.56)	(3.77)
Pay Emp Gr (Cnty-naics2) x Pay-Coverage (naics2) (t3)			0.463***		0.293***
			(65.41)		(48.83)
Pay Emp Gr (cnty-Agg; t3)			0.144***		0.0897***
			(143.58)		(104.94)
Pay Emp Gr (state-naics2; t3)			0.307***		0.111***
			(342.04)		(133.09)
Observations	898,139	898,139	898,139	898,139	898,139
$1 - R^2$ (rolling)	0.497	0.934	0.732	0.524	0.484
$1 - R^2$ (k-fold)	0.501	0.721	0.617	0.445	0.431
MAE (rolling)	0.0159	0.0185	0.0170	0.0149	0.0144
MAE (k-fold)	0.0157	0.0172	0.0158	0.0139	0.0136

Comparing Accuracy of Existing Statistics and New County Estimates



Application of alternative data to COVID-19 pandemic

- Suppose we wanted to target government aid that was quickly dispersed (the Paycheck Protection Program was not targeted).
- Can Payroll data help us identify the hardest hit counties?
  - Payroll + CES
    - No better in April
    - Improvements in the rest of 2020
  - Payroll alone
    - Gives meaningful signal
    - Could have been available in late March instead of early May

Counterfactual: Identifying hardest hit counties

- Identify 25% of counties with the lowest increase in employment
  - What percent of the bottom 25% overlap

Specifications	2020-04	2020-05	2020-06	2020-07	2020-08	2020-09
Baseline Random Allocation	0.25	0.25	0.25	0.25	0.25	0.25
QCEW	1	1	1	1	1	1
=CES	0.832	0.510	0.552	0.412	0.389	0.397
Main	0.827	0.541	0.581	0.438	0.424	0.415
=Pay t3	0.619	0.430	0.407	0.358	0.336	0.351
=Pay t2	0.614	0.430	0.397	0.366	0.345	0.338
=Pay t1	0.587	0.437	0.379	0.325	0.320	0.332

Conclusions

- Current estimates provide guidance for “tolerance” for new or improved estimates
- Some evidence of improved estimates at the state level (about 11 percent reduction in error)
- County-level estimates appear promising and at reasonable accuracy standard
- Demonstrate benefits of alternative data for improving policy and efficiently allocating resource

Questions for the Committee

1. Do you have any suggestions for extensions, applications, or policy analyses related to employment statistics? Or suggested applications of this methodology outside of employment statistics?

2. Considering this effort and other work across the statistical agencies, how should alternative data sources be used to improve timeliness and granularity of Federal Statistics, while maintaining standards of accuracy and reliability?

3. How should federal statistical agencies address other issues that arise with using alternative sources (e.g., transparency, coverage, and stability of data providers)?

Acknowledgements

- Opportunity Insights for their exceptional research in this area as well as their committed partnership and support of the statistical agencies throughout this process.

- Payroll data provider for the generosity of allowing the statistical agencies to work with their valuable data leading to new statistics and insights that would not be possible without their involvement.

- FESAC for your time today in hearing about this work and providing feedback. We are looking forward to your comments

Contact

Abe.Dunn@bea.gov

Eric.English@census.gov

Kyle.Hood@bea.gov

Mason.Lowell@bls.gov

Brian.Quistorff@bea.gov