Transparent and Reproducible Research in Agricultural Official Statistics

Andreea L. Erciulescu

National Institute of Statistical Sciences
USDA National Agricultural Statistics Service

Government Advances in Statistical Programming! Workshop
October 24, 2018

Disclaimer

The Findings and Conclusions in This Preliminary Presentation Have Not Been Formally Disseminated by the U.S. Department of Agriculture and Should Not Be Construed to Represent Any Agency Determination or Policy.

USDA NASS Official Statistics; –500 Reports Annually

Agricultural Labor Survey

- Time: biannual official statistics for four quarters
  - May (April and January) and November (October and July)

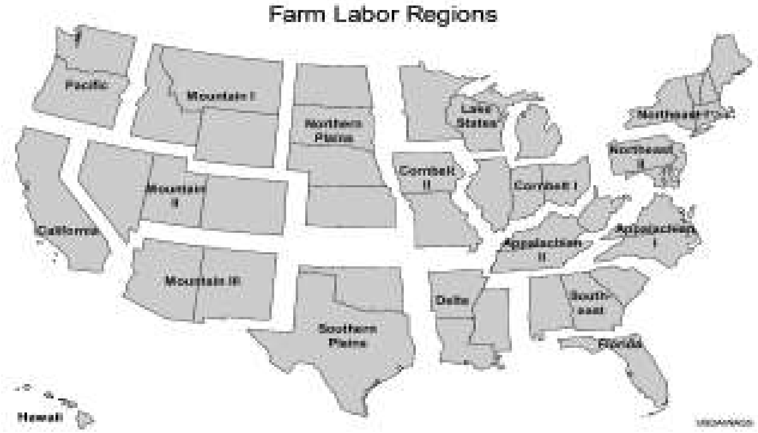
- Quantities: number of workers, hours, **wage rate**
  - point estimates only (no measures of uncertainty currently*)

- Domains: **region****, **worker-type**, farm-type, economic class, +
  - large number of fine domains (cross tabulations)

*quality measures are published for some *survey* estimates
**group of neighboring states, within the nation

Acknowledgements: NASS RDD, SD, MD

Farm Labor Report



Farm Labor Official Statistics - Example

Table 1: Wage Rates by Type of Worker - Region and United States: April 2018

Region	**Type of worker**			
	Field	*Livestock*	*Field and livestock*	*All*
Northeast I	13.44	13.03	13.25	14.46
Northeast II	13.37	12.68	13.10	13.89
Appalachian I	11.87	11.80	11.85	12.69
Appalachian II	11.56	11.64	11.60	12.69
Southeast	10.35	11.03	10.55	11.23
Florida	11.20	12.20	11.25	11.89
Lake	12.28	12.41	12.35	13.02
Cornbelt I	12.49	13.20	12.75	13.71
Cornbelt II	12.74	13.43	13.05	13.64
Delta	11.30	10.91	11.15	11.62
Northern Plains	14.52	13.47	14.00	14.70
Southern Plains	11.40	12.13	11.75	12.26
Mountain I	13.16	13.02	13.10	13.84
Mountain II	12.08	13.59	13.05	14.13
Mountain III	11.77	12.09	11.90	13.13
Pacific	14.16	14.10	14.15	14.97
California	13.45	14.15	13.58	15.10
Hawaii	14.55	15.70	14.77	16.71
United States	12.72	12.78	12.74	13.72

Modeling Agricultural Wage Rates

Strategy

- hierarchical Bayes subarea-level model, Erciulescu et al. (2018)
- model number of workers and hours (rate) $\xrightarrow{\text{multiply}}$ number of hours
- model wage rate $\xrightarrow[\text{derived number of hours}]{\text{aggregate using}}$ larger domains
- one year, one quarter, one questionnaire and one variable
- statistical software developments in R, Rjags and R2jags

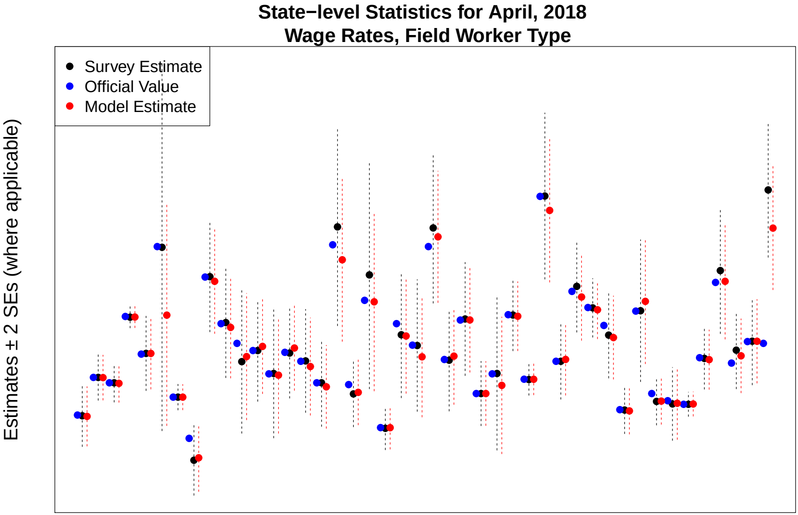
Input

- subdomain-level survey summaries: state × worker-type
  - point estimates, uncertainty measures, realized sample sizes
- past year, same quarter, official values: state × worker-type
  - point estimates

Output

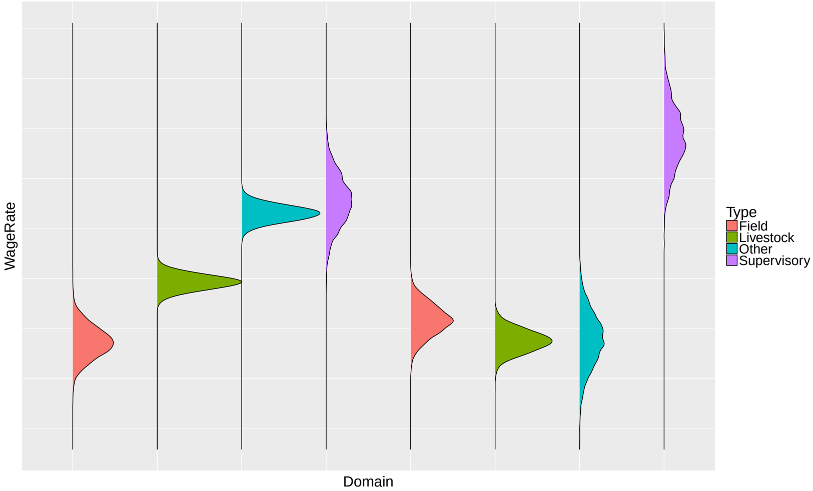
- subdomain-level and domain-level estimates
  - geography: states, regions, nation
  - type of workers: field, livestock, supervisory, other, and combinations
  - point estimates, distributions

Results: One Year, One Quarter, One Variable

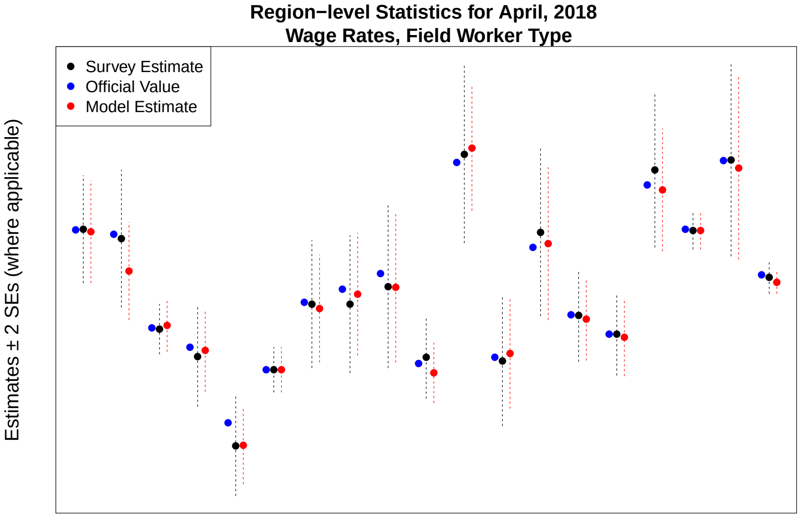


Results: One Year, One Quarter, One Variable, One Region

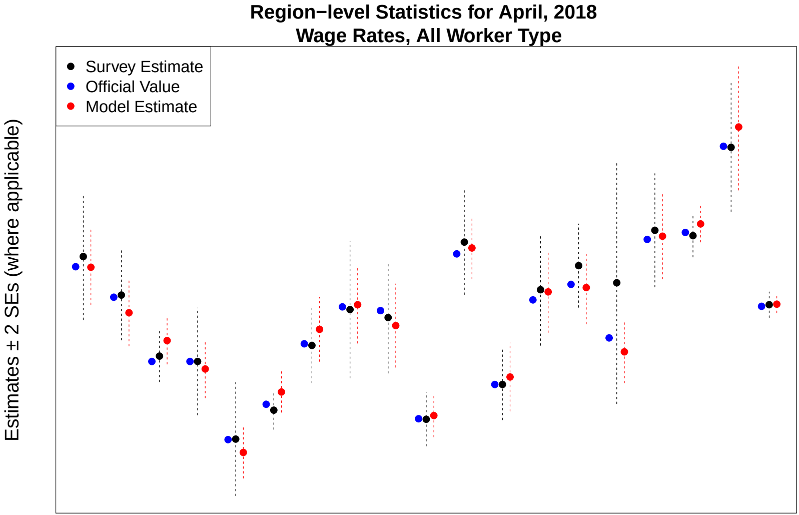
Posterior Distributions of Wage Rates
April 2018, 1 Region, 2 States, 4 Worker Types



Results: One Year, One Quarter, One Variable



Results: One Year, One Quarter, One Variable



Summary on Producing Reproducible Statistics

Transparency

- sound methodology: explicit model at disaggregated levels and implicit model at aggregated levels
- clear use and contribution of available information
- distribution of estimates
- quantifiable changes

Consistency

- agreement with current estimation process
- harmony among nested levels, types, categories, definitions
- same method for all states, regions and quantities; across time
- comparable model estimates (ME), survey estimates, official values

Efficiency

- all states and regions handled at once, not individually
- increase precision and decrease relative variability; model vs survey
- fast computation time

Additional Needs

Survey attributes

- different years, quarters, variables of interest
- different questionnaires (2018 experiment, V1 and V2)

Updates in survey summary

- from other colleagues/divisions, within the agency

Model validation

- $(1-\alpha)\%$ CIs
- relative differences, i.e. (ME - official statistic)/($c\times$ME SE)
- posterior (predictive) distributions

Comparison study

- nonoverlapping $(1-\alpha)\%$ CIs (model and survey)
- posterior distribution of differences, i.e. ME V1 vs ME V2

Literate Statistical Programming - Donald E. Knuth (1984)

Pros

- one document
  - data, code, documentation, order
- track record
  - work progress, ongoing changes, no need to save output
- live code
  - results automatically updated to reflect external changes

Cons

- assume data structure
- assume reasonable amount of data
- if lots of code, then slower process (may be avoided using cache)

R Knitr: combine (R) code and (*LaTeX*) text

Some resources

- homepage, development repository
- options page, examples
- stackoverflow

Example code chunk options

- echo: show
- warning/message/error: show/stop
- eval: evaluate
- cache: cache results

Example functions

- knit(): knit input document and write output (RStudio recognizes .Rnw input extension and .tex output extension)
- purl(): extract R code from an input document

Simple Examples

documentclass{article}

begin{document}

Will provide simple code chunks and inline text in the following slides.

«»= (R code) @, i.e. «exchunk,echo=TRUE,eval=FALSE»=

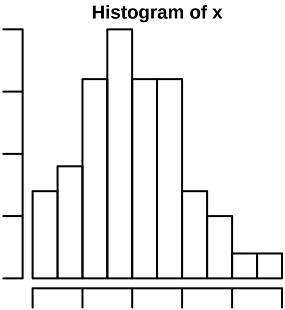
end{document}

Simple Examples 2

```
set.seed(2018)
x <- rnorm(100,0,1)
mx <- mean(x)
```

Sexpr{mx}: 0.0211919

```
set.seed(2018)
x <- rnorm(100,0,1)
par(mar = c(1,1,0.3,0.3) + 0.1)
hist(x, cex.main=0.6,cex.lab=0.6, cex.axis=0.6, xlab=’’)
```



Simple Examples 3

```
set.seed(2018)
matx <- matrix(rnorm(20,0,1),nrow=4)
d <- formatC(matx,3, format='f')
```

```
begin{table}[ht]
centering
begin{tabular}{rrrrrr}
hline
& 1 & 2 & 3 & 4 & 5
hline
1 & Sexpr{d[1,1]} & Sexpr{d[1,2]} & Sexpr{d[1,3]} & Sexpr{d[1,4]} & Sexpr{d[1,5]}
2 & Sexpr{d[2,1]} & Sexpr{d[2,2]} & Sexpr{d[2,3]} & Sexpr{d[2,4]} & Sexpr{d[2,5]}
3 & Sexpr{d[3,1]} & Sexpr{d[3,2]} & Sexpr{d[3,3]} & Sexpr{d[3,4]} & Sexpr{d[3,5]}
4 & Sexpr{d[4,1]} & Sexpr{d[4,2]} & Sexpr{d[4,3]} & Sexpr{d[4,4]} & Sexpr{d[4,5]}
hline
end{tabular}
end{table}
```

	1	2	3	4	5
1	-0.423	1.735	-0.611	0.712	-1.827
2	-1.550	-0.265	0.637	-0.446	0.015
3	-0.064	2.099	-0.643	0.249	-1.684
4	0.271	0.863	-1.030	-1.074	0.204

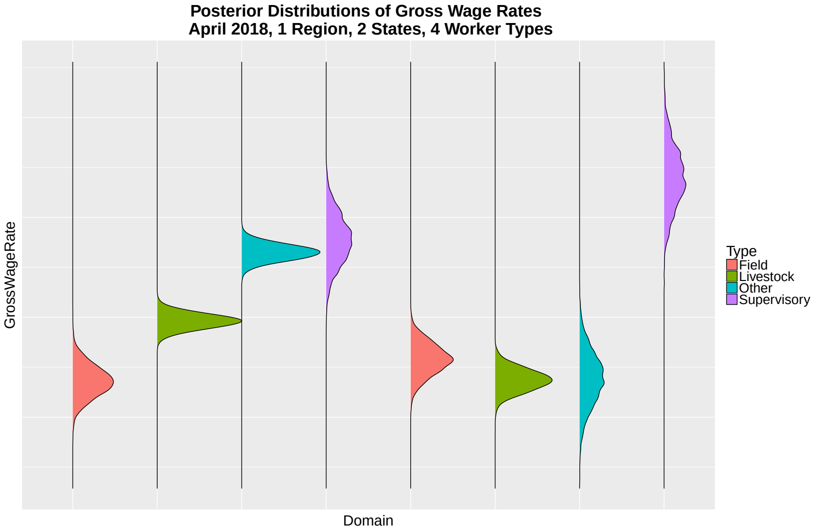
Farm Labor Report using Model-based* Estimates

*** Due to disclosure limitations, official estimates are used for illustration**

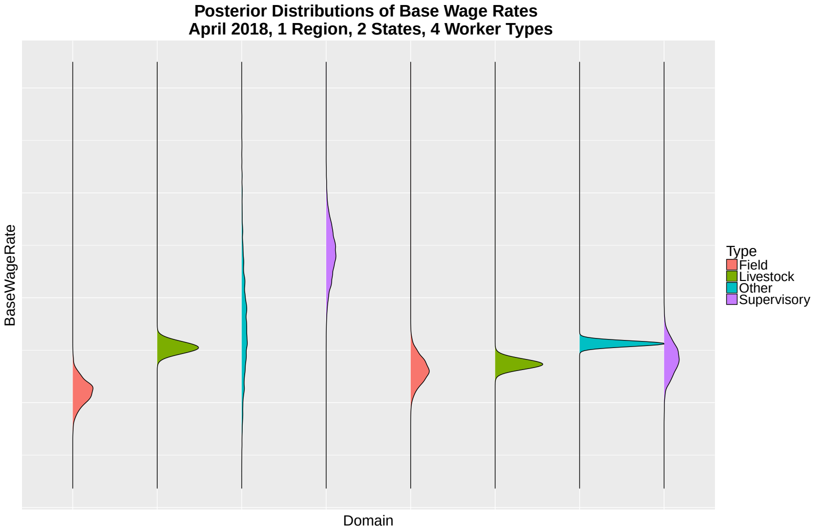
Table 2: Wage Rates by Type of Worker - Region and United States: April 2018

Region	**Type of worker**			
	Field	*Livestock*	*Field and livestock*	*All*
NortheastI	13.44 (1.68)	13.03 (1.56)	13.25 (0.41)	14.46 (0.58)
NortheastII	13.37 (2.09)	12.68 (1.48)	13.10 (0.37)	13.89 (0.42)
AppalachianI	11.87 (1.42)	11.80 (1.57)	11.85 (0.40)	12.69 (0.24)
AppalachianII	11.56 (1.73)	11.64 (2.20)	11.60 (0.44)	12.69 (0.51)
Southeast	10.35 (1.46)	11.03 (1.65)	10.55 (0.31)	11.23 (0.53)
Florida	11.20 (0.97)	12.20 (1.48)	11.25 (0.25)	11.89 (0.19)
Lake	12.28 (1.23)	12.41 (1.30)	12.35 (0.25)	13.02 (0.35)
CornbeltI	12.49 (1.57)	13.20 (2.01)	12.75 (0.64)	13.71 (0.64)
CornbeltII	12.74 (1.59)	13.43 (2.10)	13.05 (0.51)	13.64 (0.53)
Delta	11.30 (0.96)	10.91 (2.60)	11.15 (0.18)	11.62 (0.24)
NorthernPlains	14.52 (1.64)	13.47 (1.59)	14.00 (0.43)	14.70 (0.49)
SouthernPlains	11.40 (1.46)	12.13 (1.88)	11.75 (0.35)	12.26 (0.33)
MountainI	13.16 (2.37)	13.02 (2.19)	13.10 (0.55)	13.84 (0.51)
MountainII	12.08 (1.68)	13.59 (2.99)	13.05 (0.38)	14.13 (0.38)
MountainIII	11.77 (1.74)	12.09 (1.52)	11.90 (0.33)	13.13 (1.04)
Pacific	14.16 (1.74)	14.10 (1.64)	14.15 (0.31)	14.97 (0.52)
California	13.45 (0.58)	14.15 (0.67)	13.58 (0.15)	15.10 (0.21)
Hawaii	14.55 (1.37)	15.70 (1.92)	14.77 (0.81)	16.71 (0.60)
UnitedStates	12.72 (0.36)	12.78 (0.40)	12.74 (0.09)	13.72 (0.12)

Results: One Year, One Quarter, One Variable, One Region



Results: One Year, One Quarter, One Variable, One Region



Summary

Reproducible research framework: producing and reporting

List of R packages used:

Dynamic Reporting: knitr
Text Mining and Manipulation: tm, stringr
Data Manipulation: dplyr
Model Fit and Estimation: rjags, R2jags
Data Visualization: ggplot2, ggridges

Selected References

Erciulescu A.L., Cruze N., Nandram B. (2018) “Model-Based County-Level Crop Estimates Incorporating Auxiliary Sources of Information,” *Journal of the Royal Statistical Society, Series A*, DOI 10/1111/rssa.12390.

Knuth D. (1984) “Literate Programming,” *The Computer Journal*, 27, 2, 97 - 111.

Reist B., Wilson T., Ball S., Young L. (2018) “Preliminary Findings for April 2018 Agricultural Labor base Wage Question Experiment,” *USDA NASS RDD Research Report*, RDD-17-xx.

USDA NASS (2018) “Farm Labor Survey,” *https://www.nass.usda.gov/Surveys/GuidetoNASSSurveys/FarmLabor/*.

Xie, Y. (2005-2018) “knitr,” *https://yihui.name/knitr/*.

Thank you!

aerciulescu@niss.org
andreea.erciulescu@nass.usda.gov