Joint Confidence Regions for Communicating Uncertainty in Rankings

Jerzy Wieczorek
Colby College
jerzy.wieczorek@colby.edu
@civilstat

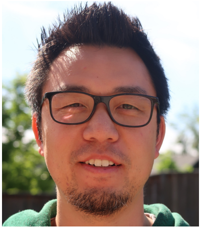
October 5, 2022
"Statistical and ML approaches for the social sciences"
University of Washington

Collaborators at US Census Bureau and beyond

Tommy Wright

Martin Klein

Adam Hall



Nathan Yau

Joel Beard

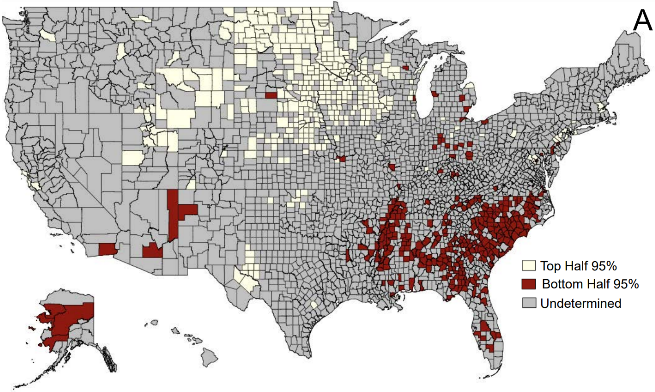
Why communicate uncertainty in statistical rankings?

Mogstad et al. 2020:

Ignoring uncertainty in rankings may lead to spurious policies.

E.g., some economists advise US counties to mimic those counties with the highest degree of upward income mobility.

But. . . which ones are they? Most counties' ranks could plausibly be in top **or** bottom half!



Can we just leave it up to users?

Jurjevich et al. 2018:
Data users often ignore SEs or MOEs, even when they are relevant to the comparisons or rankings of interest.

In a survey of 200 urban & regional planners who use ACS data,

- 70-80% use it for comparing communities with each other,
- yet 23% admitted they usually ignore MOE values,
- and only 53% agreed with “Demographic and economic estimates from the ACS are only suitable for making comparisons between places if MOEs are considered”

Advice from a respondent:

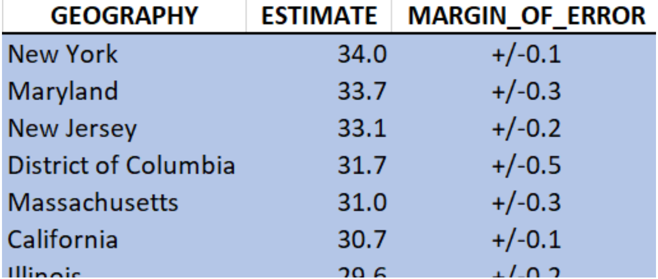
> *I might present the value as a range rather than just the number. Graphing can be helpful in this regard, as it allows the viewer to understand that the value is “somewhere” in the range but we can’t be precise enough to name it.*

Running example: mean travel time to work by US state

American Community Survey (ACS):
annual US survey, stratified by state

For a given year and topic, we have $K = 51$ independent estimates and their SEs (or 90% MOEs)

Let θ_k be the
"*Mean travel time to work of workers 16 years and over who did not work at home (minutes)*"
. . . for each k in the 50 states plus DC



GEOGRAPHY	ESTIMATE	MARGIN_OF_ERROR
New York	34.0	+/-0.1
Maryland	33.7	+/-0.3
New Jersey	33.1	+/-0.2
District of Columbia	31.7	+/-0.5
Massachusetts	31.0	+/-0.3
California	30.7	+/-0.1
Illinois	29.6	+/-0.2

How has ACS communicated uncertainty?

ACS ranking tables:
Report estimates and their 90% MOEs.
Users can choose one state, and the table will update to show which states are significantly different.

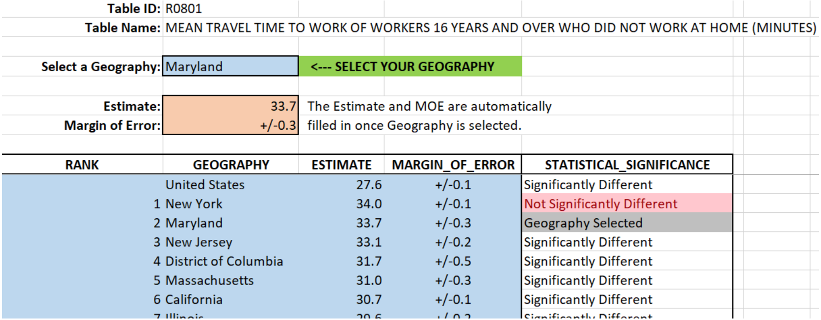


Table ID: R0801

Table Name: MEAN TRAVEL TIME TO WORK OF WORKERS 16 YEARS AND OVER WHO DID NOT WORK AT HOME (MINUTES)

Select a Geography: Maryland **<--- SELECT YOUR GEOGRAPHY**

Estimate: 33.7

Margin of Error: +/-0.3

The Estimate and MOE are automatically filled in once Geography is selected.

RANK	GEOGRAPHY	ESTIMATE	MARGIN_OF_ERROR	STATISTICAL_SIGNIFICANCE
	United States	27.6	+/-0.1	Significantly Different
1	New York	34.0	+/-0.1	Not Significantly Different
2	Maryland	33.7	+/-0.3	Geography Selected
3	New Jersey	33.1	+/-0.2	Significantly Different
4	District of Columbia	31.7	+/-0.5	Significantly Different
5	Massachusetts	31.0	+/-0.3	Significantly Different
6	California	30.7	+/-0.1	Significantly Different
7	Illinois	29.6	+/-0.2	Significantly Different

This summarizes uncertainty about state-to-state **comparisons**, but not about the **overall ranking** of all states.

Static visualizations and R packages

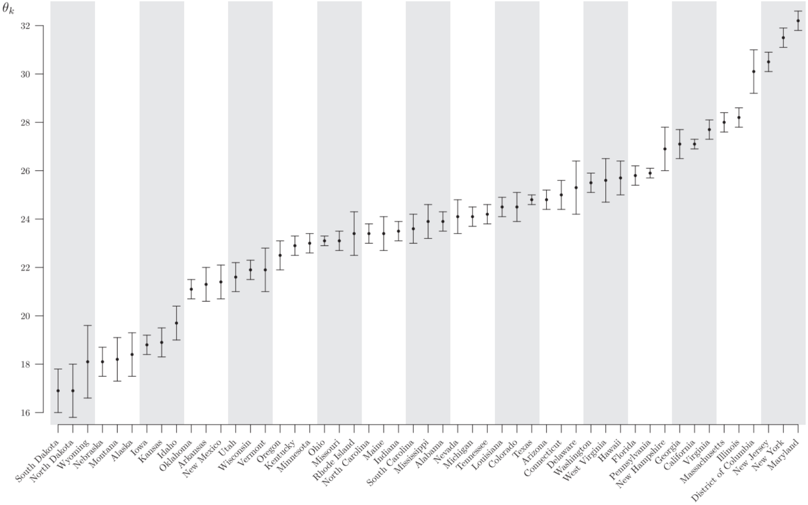
Klein et al. 2020 & {RankingProject}

Goal: one **joint confidence region** for all **ranks** r_k simultaneously

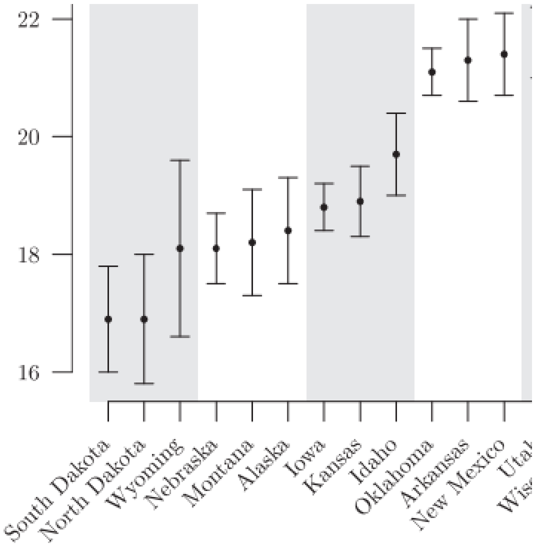
Key idea:

- Find a set of CIs for each state's **estimand** θ_k, with mult.comp.-correction for desired joint coverage across $k \in \{1, \dots, K\}$.
- For each state k, count how many other CIs-for-the-**estimand** are entirely below this one, and how many are entirely above.
- Take $\{1, \dots, K\}$ and exclude that many ranks from bottom and top, respectively. What's left is state k's CI-for-the-**rank**.

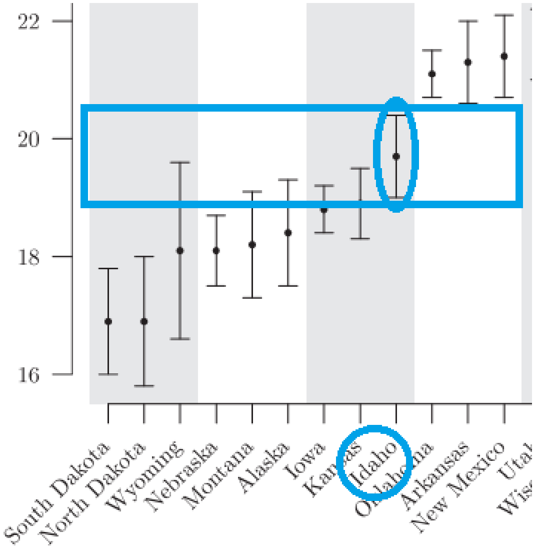
Klein et al. 2020: 90% joint conf. region for all θ_k



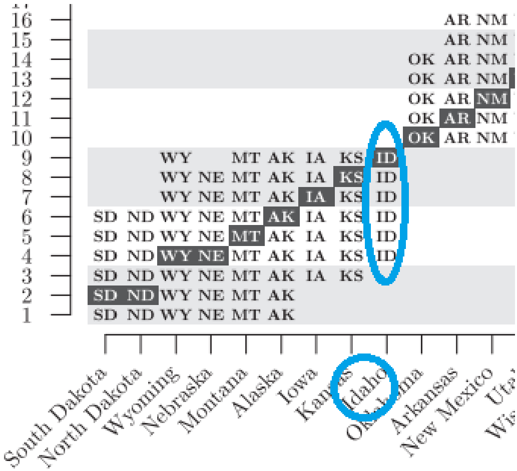
Klein et al. 2020: zoom into some of the CIs for θ_k



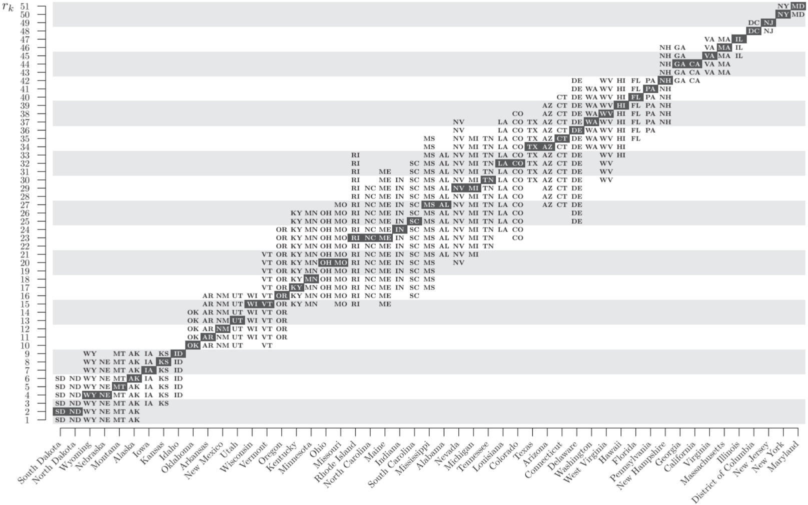
Klein et al. 2020: Idaho's CIs for θ_k is >3 and <42 others



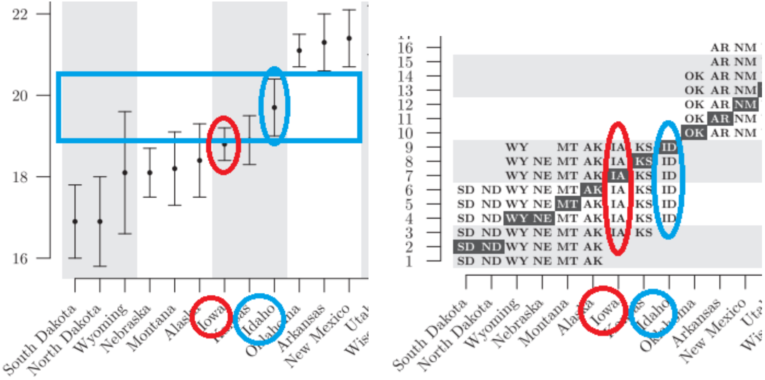
Klein et al. 2020: Idaho's CI for r_k is $\{4, \ldots, 9\}$



Klein et al. 2020: 90% joint conf. region for all r_k



Klein et al. 2020: diff. b/w wide MOE for θ_k vs for r_k



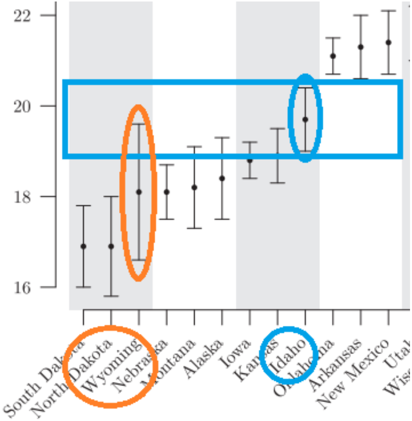
Uncertainty in ranks will depend not just on MOEs, but **also** on how close estimates are to each other.

Compared to Iowa, Idaho's CI is **wider** for θ_k, yet **narrower** for r_k.

Iowa's $\hat{\theta}_k$ was closer to other states, hence more θ_k CIs overlapped despite its smaller MOE.

Klein et al. 2020: possible misconception

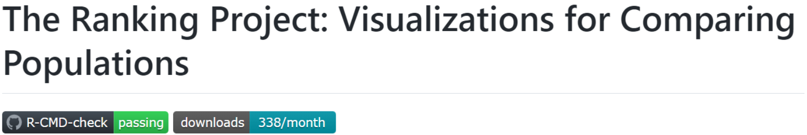
"Wyoming has estimated rank $\hat{r}_{WY} = 3$ and Idaho is not sig.diff. from Wyoming, so Idaho's rank is not sig.diff. from 3"??? **No**: Idaho CI is $>$ 3 other states, so 3 should **not** be in its conf.set.



Hence, Idaho's rank CI is 4 through 9, **not** $\{3, 5, 6, 7, 8, 9\}$.

Klein et al. 2020 & {RankingProject}

{RankingProject} R package v0.4.0 is available on CRAN:
https://cran.r-project.org/package=RankingProject
and on GitHub:
https://github.com/civilstat/RankingProject



The Ranking Project: Visualizations for Comparing Populations

See joint vignette for code to reproduce our figures.

Mogstad et al. 2020 & {csranks}

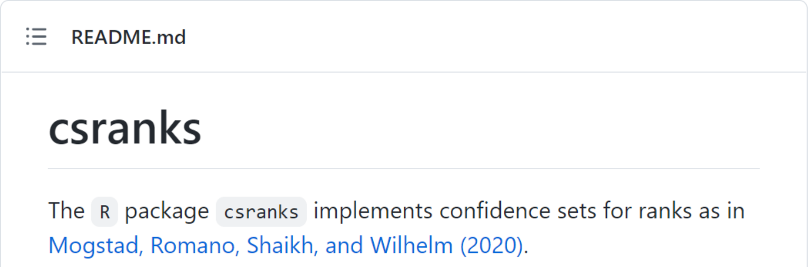
Key idea:
Refinement of Klein et al.: Instead of checking if indiv. CIs overlap, run **pairwise tests** for every difference.

This requires a **stronger mult.comp. correction** (for $\binom{K}{2}$ pairs instead of only for K states), but each indiv. comparison is **more powerful**. (Checking for CI overlap is often too conservative.)

In practice, Mogstad et al.'s confidence regions for the ranks are usually no wider and sometimes **narrower than Klein et al.'s**.

Mogstad et al. 2020 & {csranks}

{csranks} R package v0.2.0 is available on GitHub:
https://github.com/danielwilhelm/R-CS-ranks



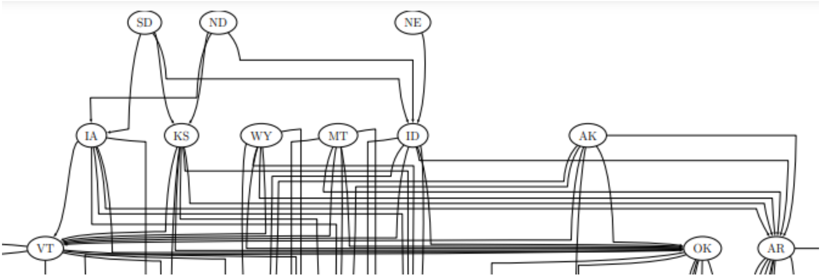
README.md

csranks

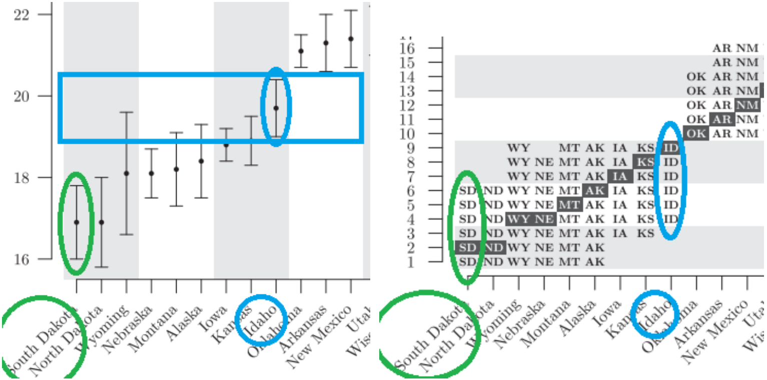
The `R` package `csranks` implements confidence sets for ranks as in Mogstad, Romano, Shaikh, and Wilhelm (2020).

Rising 2021 & {rankUncertainty}

Key idea:
Use "partial orders" and ideas from graph theory to summarize a **set of rankings compatible with the data** that is less conservative than Klein et al.



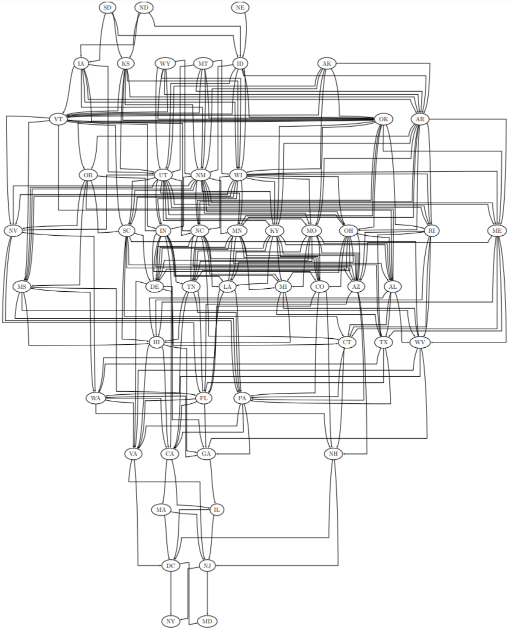
Rising 2021 & {rankUncertainty}



Recall Klein et al.: we are confident that $\theta_{ID} > \theta_{SD}$. Yet in the joint conf. region for ranks, $r_{ID} < r_{SD}$ could be plausible.

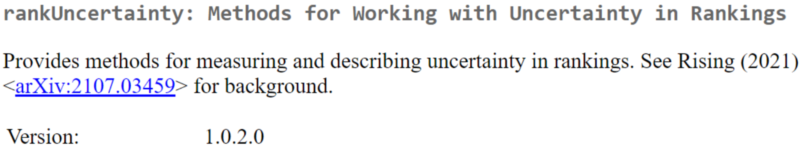
Rising's "cover graph" attempts to exclude such implausible rankings from the summary of possible rankings.

Rising 2021 & {rankUncertainty}



Rising 2021 & {rankUncertainty}

{rankUncertainty} R package v1.0.2.0 is available on CRAN:
https://cran.r-project.org/package=rankUncertainty



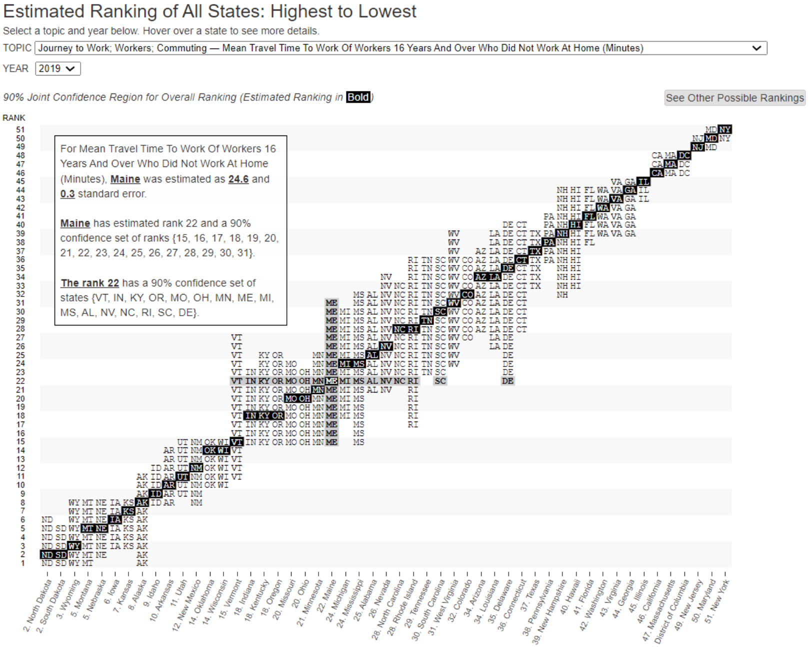
rankUncertainty: Methods for Working with Uncertainty in Rankings

Provides methods for measuring and describing uncertainty in rankings. See Rising (2021) <arXiv:2107.03459> for background.

Version:	1.0.2.0

Interactive visualizations

https://www.census.gov/csrm/rankings/



Estimated Ranking of All States: Highest to Lowest

Select a topic and year below. Hover over a state to see more details.

TOPIC: Journey to Work; Workers; Commuting — Mean Travel Time To Work Of Workers 16 Years And Over Who Did Not Work At Home (Minutes)

YEAR: 2019

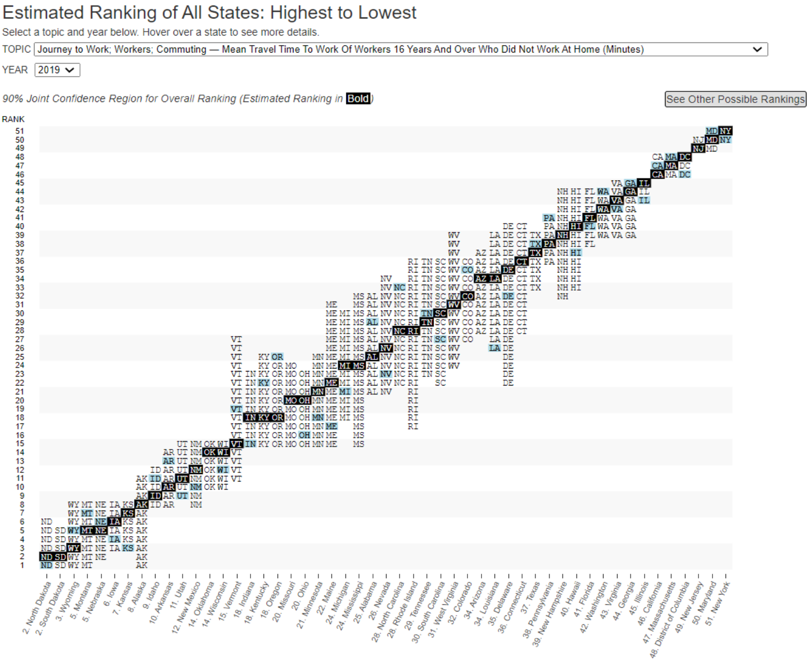
90% Joint Confidence Region for Overall Ranking (Estimated Ranking in **Bold***)*

For Mean Travel Time To Work Of Workers 16 Years And Over Who Did Not Work At Home (Minutes), **Maine** was estimated as **24.6** and **0.3** standard error.

Maine has estimated rank 22 and a 90% confidence set of ranks {15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31}.

The rank 22 has a 90% confidence set of states {VT, IN, KY, OR, MO, OH, MN, ME, MI, MS, AL, NV, NC, RI, SC, DE}.

https://www.census.gov/csrm/rankings/



Estimated Ranking of All States: Highest to Lowest

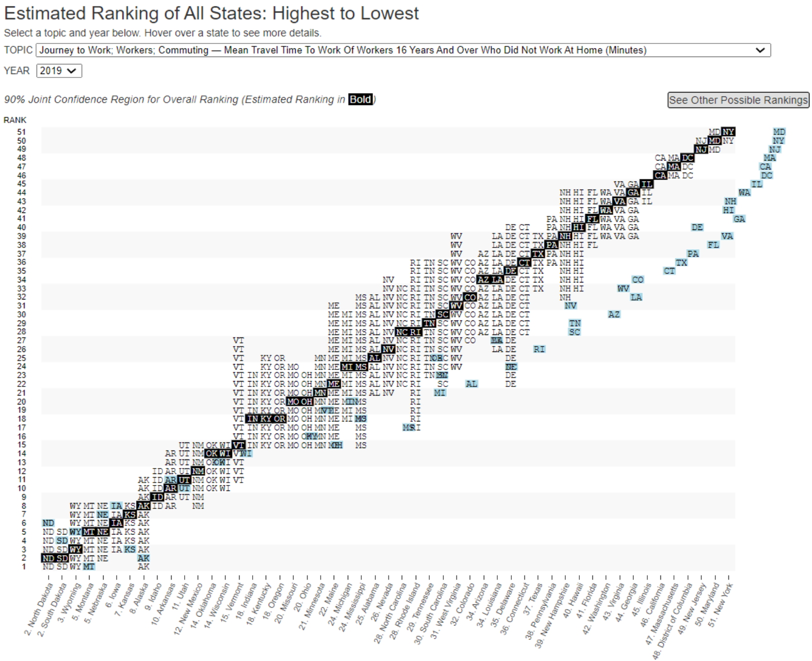
Select a topic and year below. Hover over a state to see more details.

TOPIC: Journey to Work; Workers; Commuting — Mean Travel Time To Work Of Workers 16 Years And Over Who Did Not Work At Home (Minutes)

YEAR: 2019

90% Joint Confidence Region for Overall Ranking (Estimated Ranking in **Bold***)*

https://www.census.gov/csrm/rankings/



Estimated Ranking of All States: Highest to Lowest

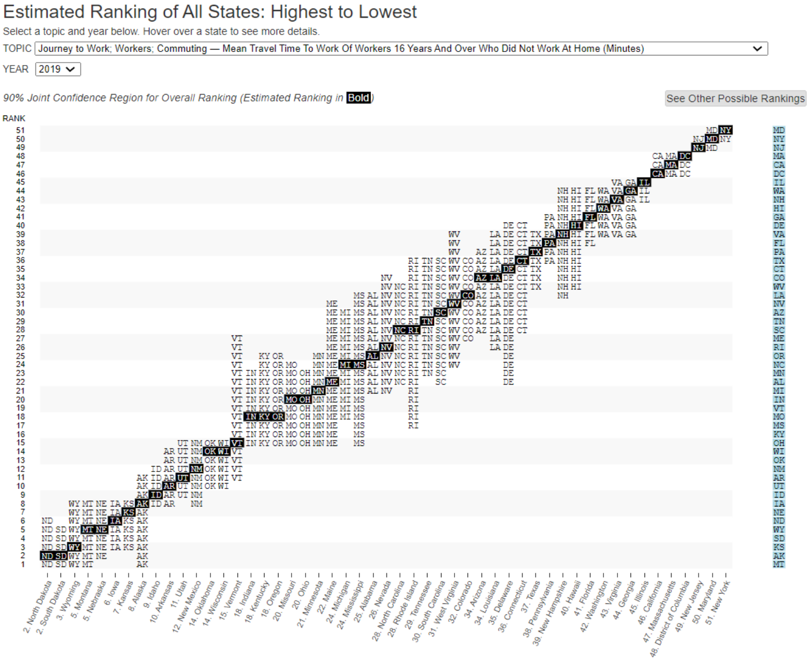
Select a topic and year below. Hover over a state to see more details.

TOPIC Journey to Work; Workers; Commuting — Mean Travel Time To Work Of Workers 16 Years And Over Who Did Not Work At Home (Minutes)

YEAR 2019

90% Joint Confidence Region for Overall Ranking (Estimated Ranking in **Bold***)*

https://www.census.gov/csrm/rankings/



Estimated Ranking of All States: Highest to Lowest

Select a topic and year below. Hover over a state to see more details.

TOPIC Journey to Work; Workers; Commuting — Mean Travel Time To Work Of Workers 16 Years And Over Who Did Not Work At Home (Minutes)

YEAR 2019

90% Joint Confidence Region for Overall Ranking (Estimated Ranking in **Bold***)*

Preliminary (informal!) user testing

What has been successful?

- Tutorial helps (though could be improved)
- Shows the variability: some places more uncertain than others
- Conveys idea that **estimated** ranking may not be **true** ranking

What challenges remain?

- Too much, too crowded... Might be easier to read on a large printout, but then no hover/pop-ups
- Concept of "confidence set" is unclear to many audiences
- How to show year-to-year comparisons?
- Allow users to group states?
  (e.g. by region: Western, Midwest, etc.)

Conclusions

- We should show the uncertainty in our rankings more often
  - Visuals can help users to decide how much to rely on the estimated ranks
  - If the estimated ranks are too uncertain to be useful, it's good to be aware!
- Instead of checking CI overlap, pairwise comparisons tend to be more powerful (even after mult.comp. corrections)
- Usability testing can help make your visuals more effective for wider audiences
- Existing visuals have room for improvement—ideas welcome!

Thank you!

Please reach out, especially about...

- other literature and visualizations for this purpose
- good test-case datasets or data-dissemination efforts
- related usability studies
- feedback on https://www.census.gov/csrm/rankings/

Contact: jerzy.wieczorek@colby.edu or @civilstat

Cited work:

- Jurjevich, Griffin, Spielman, Folch, Merrick, Nagle (2018), "Navigating statistical uncertainty: how urban and regional planners understand and work with ACS data for guiding policy," *JAPA*.
- Klein, Wright, Wieczorek (2020), "A joint confidence region for an overall ranking of populations," *JRSS-C*.
- Mogstad, Romano, Shaikh, Wilhelm (2020), "Inference for ranks with applications to mobility across neighborhoods and academic achievement across countries," *NBER* working paper 26883.
- Rising (2021), "Uncertainty in ranking," arXiv:2107.03459v3.

Supplemental slides

Mogstad et al. 2020 is less conservative: rough sketch

If states are indep. and have equal SEs...

"No CI overlap" happens if:
$\bar{X}_1 + \hat{SE} \times Z < \bar{X}_2 - \hat{SE} \times Z$
$\bar{X}_2 - \bar{X}_1 > 2\hat{SE} \times Z$

"Pairwise diff is significant" happens if:
$\hat{SE}_{diff} = \sqrt{\hat{SE}_1^2 + \hat{SE}_2^2} = \sqrt{2}\hat{SE}$
$\bar{X}_2 - \bar{X}_1 > \hat{SE}_{diff} \times Z = \sqrt{2}\hat{SE} \times Z$

Checking for CI overlap is equivalent to directly checking pairwise diffs **with a Z-score that is $\sqrt{2}$ times larger**.

For 90% confidence level and $K = 51$ and Bonferroni correction,
Klein's $Z_{K\ intervals} = 3.10$ while
Mogstad's $Z_{\binom{K}{2}\ comparisons} = 3.95 < 4.38 = \sqrt{2} \times 3.10$.
So checking overlap among K CIs is conservative, compared to checking $\binom{K}{2}$ pairwise diffs directly.

Caution: rankings $\neq$ pairwise comparisons

"...we note that a joint confidence statement about rankings is distinct from a joint set of pairwise comparisons for states. These two joint statements may seem equivalent at first glance, but in fact they can differ substantially in the presence of sampling error, especially when similar estimates may have different levels of precision...[W]e are 90% confident that Alabama's *rank* is no greater than 33 ... This does not mean that Alabama is significantly different from every *state* whose *estimated* rank is 34 or higher. Delaware, in particular, is a smaller state with a relatively large SE. Its estimated rank is 36, outside the [interval] for Alabama's rank, but Delaware's interval of (24.2, 26.4) minutes is wide enough to overlap with Alabama's interval of (23.5, 24.3) minutes. In other words, Alabama's rank is significantly different from 36, and yet Alabama's travel time to work is not significantly different from that of the state whose estimated rank happens to be 36 (Delaware). For this reason, it is important to decide whether rankings or pairwise comparisons will be of primary interest when reporting many estimates."

More quotes from Jurjevich et al. 2018

Planner 1:

> *I also found when I started bringing it in this time around, I eliminated the MOE. I took it out. Because it's just all these extra columns that I don't need.*

Planner 2:

> *Any good statistics class, software, person who just does statistics will. . . include a margin of error. . . However, we just don't use it. Nobody. . . unless you're a statistics type person presenting to statistics professors where you have to have your footnotes in there. . .*

Planner 3:

> *Depending on the use of the data—that is, if no capital or human life issues are involved—I might present the value as a range rather than just the number. Graphing can be helpful in this regard, as it allows the viewer to understand that the value is "somewhere" in the range but we can't be precise enough to name it.*

Alternatives to ranking tables

Goldstein and Spiegelhalter 1996

Goldstein, Spiegelhalter (1996), "League tables and their limitations: statistical issues in comparisons of institutional performance," *JRSS-A*.

- Do you really need to rank institutions (e.g. schools, hospitals) at all?
  - For students picking a school or patients choosing a doctor, such rankings may *seem* useful. . .
  - For goal of improving institutions, such rankings may *seem* to highlight best-practices or shine a spotlight on worst offenders. . .
  - . . . but in either case, decisions may be spurious if ranks are highly uncertain
- May be better to collaborate with institutions directly on improvements than to spur noisy "competition," and to just inform users that the data are inadequate for ranking
- At most, could privately tell each institution its *own* score or ranking relative to others, but keep the rest anonymous

Goldstein and Spiegelhalter 1996

This implies that current official support for output league tables, even adjusted, is misplaced and governments should be concerned that potential users are properly informed of their shortcomings. If such tables continue to be produced then they need an accompanying warning about their use.

. . .

An overinterpretation of a set of rankings where there are large uncertainty intervals, as in the examples that we have given, can lead both to unfairness and to inefficiency and unwarranted conclusions about changes in ranks. In particular, apparent improvements for low ranking institutions may simply be a reflection of 'regression to the mean'.

Goldstein and Spiegelhalter 1996

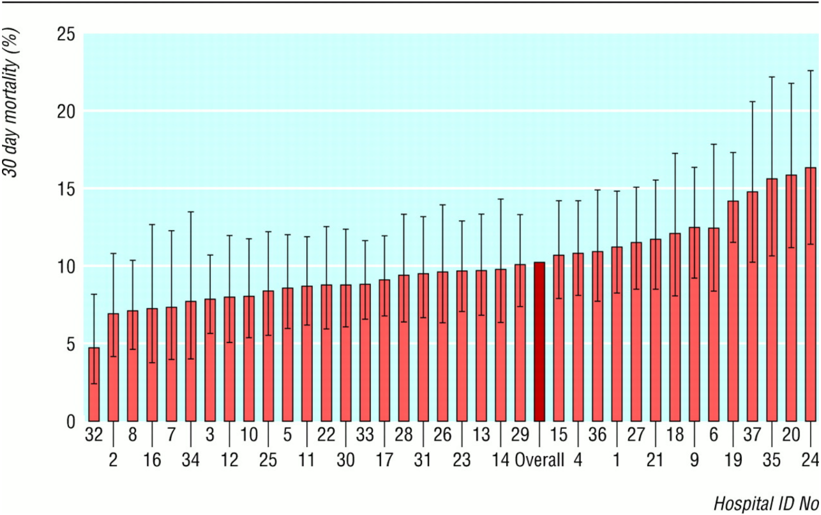
- If you must publish a ranking. . .
  - Show the uncertainty in the rankings—possibly *only* show intervals, & not point estimates, for ranks
  - Adjust for important covariates (e.g. no sense in ranking hospitals without accounting for differences in patient demographics or disease severity)
  - Be aware that modeled estimates might have lower SEs, but may also be shrunk together, which may make the conf. region for ranks *wider* after modeling than before
  - Warn users of shortcomings in the rankings: wide conf. regions, sensitivity to choice of adjustment model, etc.
  - Be aware that if rankings become public or get used to allocate resources, institutions will likely try "gaming" them
- I also strongly recommend the rich discussion published with the paper!

Adab et al. 2002

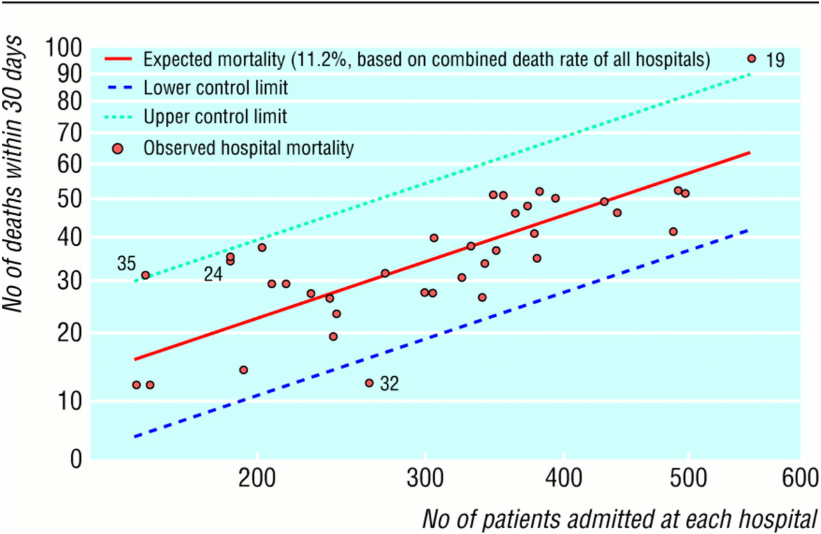
Adab, Rouse, Mohammed, Marshall (2002), “Performance league tables: the NHS deserves better,” *BMJ*.

- Performance rankings can make sense for widely varied things (like products made by different manufacturers), but it may not make sense to rank parts of a common system (like hospitals within UK’s NHS system).
- Public rankings are meant to incentivize indiv. improvements. But when most variation is due to “common cause” noise that needs to be improved at the system-level, not unit-level,
  - public rankings may be seen as unfair shaming of whoever randomly got worst score, and
  - most units may have few ways to improve their rank. . . besides “creative reporting.”
- Instead, use **control charts** to separate out “common causes” of variation (to be improved at the hospital-system level) vs. “special causes” (to be addressed at indiv. hospitals).

Adab et al. 2002: instead of ranking. . .



Adab et al. 2002: . . . use a control chart



Adab et al. 2002: . . . use a control chart

- Control chart, with size on x-axis and outcome on y-axis, using sqrt scale for both:
  - Central line shows expected deaths for that patient-pop size, assuming common mortality rate
  - Ranking table obscured the fact that small-pop hospitals *should* have more variability on % scale than large-pop hospitals; but control chart with 3-sigma tolerance band accounts for this
  - Only places outside the tolerance band need to be studied for "special causes"—and they are NOT necessarily same as places with lowest/highest ranks!
- No explicit ranking
  → no "bottom of the list" for hospitals to fear landing at
  → less incentive to game the metrics
- If control chart shows most hospitals are inside the tolerance band, it confirms that you should focus on **improving the system** rather than shaming "worst performers"