Supplementary Appendix: How to Measure Legislative District Compactness If You Only Know it When You See it[*]

Aaron Kaufman[†] Gary King[‡] Mayya Komisarchik[§]

September 25, 2018

[*] The current version of this supplementary appendix is available at j.mp/Compactness.

[†] PhD Candidate, Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; AaronrKaufman.com; aaronkaufman@fas.harvard.edu, (818) 263-5583.

[‡] Albert J. Weatherhead III University Professor, Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; GaryKing.org, King@Harvard.edu, (617) 500-7570.

[§] PhD Candidate, Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; scholar.harvard.edu/mkomisarchik; mkomisarchik@fas.harvard.edu, (720) 220-9328.

Appendix A Geometric Features of Legislative Districts

We define many useful existing compactness measures, and other geometric features of legislative districts we introduce. We use all of these quantities in Section 3.2. We begin with basic notation used in many of the measures and then define the measures.

Notation Denote a generic legislative district as D, and define it as a non-self-intersecting closed polygon with n vertices, each labeled (x_i, y_i) and numbered i in clockwise order (for $i = 1, \dots, n$). We choose an arbitrary starting vertex for label $i = 1$ and (using clock or modular algebra) define $i = n + 1 = 1$. The length of the line segment from vertex i to $i + 1$ is then $L_i = ||(x_i, y_i), (x_{i+1}, y_{i+1})||$ where $||(a, b), (c, d)|| = \sqrt{(a - c)^2 + (b - d)^2}$. Denote the set of all horizontal vertex coordinates as $X = \{x_i : i = 1, \dots, n\}$, vertical vertex coordinates as $Y = \{y_i : i = 1, \dots, n\}$, and line lengths as $L = \{L_i : i = 1, \dots, n\}$.

Then the area of D is $A(D) = \frac{1}{2}\sum_{i=1}^{n}(x_i y_{i+1} - x_{i+1} y_i)$ and perimeter is $P(D) = \sum_{i=1}^{n} L_i$. Occasionally, as in the case of islands, D is composed of multiple polygons. In these cases, $A(D)$ and $P(D)$ are the sums of the areas and perimeters of all the polygons in D, and all subsequent notation refers to all vertices in all polygons taken together.

Denote the district centroid as $C(D)$, defined by a vertex with coordinates $C(D)_x = \frac{1}{6A(D)}\sum_{i=0}^{n-1}(x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i)$ and $C(D)_y = \frac{1}{6A(D)}\sum_{i=0}^{n-1}(y_i + y_{i+1})(x_i y_{i+1} - x_{i+1} y_i)$, and radii $r_i = ||[C(D)_x, C(D)_y], (x_i, y_i)||$. Then denote as Circle$(D)$ the minimum bounding circle (Nielsen and Nock, 2008) and as Hull(D) the minimum bounding convex hull (King and Zeng, 2006; Kong, Everett, and Toussaint, 1990). Finally, for set S with cardinality $\#S$, denote the mean over i of function $g(i)$ as $\text{mean}_{i\in S}[g(i)] = \frac{1}{\#S}\sum_{i=1}^{\#S} g(i)$, the variance as $\text{var}_{i\in S}[g(i)] = \text{mean}_{i\in S}\left[\{g(i) - \text{mean}_{j\in S}[g(j)]\}^2\right]$, and the mean absolute deviation as $\text{mad}[g(i)] = \frac{1}{\#S}\sum_{i=1}^{\#S}|g(i) - \text{mean}[g(i)]|$.

Measures The perimeter of the minimum bounding circle is $\textsc{pc} = P(\text{Circle}(D))$ and minimum bounding convex hull is $\textsc{pch} = P(\text{Hull}(D))$. The area of each is the $\textsc{ac} = A(\text{Circle}(D))$ and $\textsc{ach} = A(\text{Hull}(D))$. The number of polygons is $\textsc{Parts}$ and vertices, or sides, is Sides $= n$ (Timmerman, 100 N.Y.S. 57, 51 Misc. Rep. 192 (N.Y. Sup. 1906)). We then have Reock $= A(D)/A(\text{Circle}(D))$; Grofman $= P(D)/\sqrt{(A(D))}$; Hull ratio $= A(D)/A(\text{Hull}(D))$; Schwartzberg $= P(D)/(2\pi\sqrt{A(D)/\pi})$ and the mathematically related PolsbyPopper $= 4\pi A(D)/P(D)^2$; the variation in the coordinates of the x-axis, xvar $= \text{var}_{i\in X}[x_i]$, y-axis, yvar $= \text{var}_{i\in Y}[y_i]$, and the ratio of the two $|1 -$ xvar/yvar$|$; the average, AvgLL $= P(D)/n = \text{mean}_{i\in L}\, L_i$, and variance, VarLL $= \text{var}[L_i]$, of the polygon line segment lengths; jaggedness, the average line length divided by the perimeter; Length-Width Ratio $= [\max_i(x_i) - \min_i(x_i)]/[\max_i(y_i) - \min_i(y_i)]$; (our simplified expression of modified) Boyce-Clark $= 1 - \frac{1}{2\,\text{mean}_i[r_i]}\,\text{mad}_i[r_i]$ (MacEachren, 1985, p.56); Points $= n$ for the district polygon defined by the official US Census TIGER shapefiles; using the Harris Corner Detector algorithm (Harris and Stephens, 1988), we also have the number of significant "corners" (i.e., vertices), Corners, and the variance in the x-coordinate xVarCorners and y-coordinate yVarCorners of each corner. As well, we measure cornerRatio, equal to $|1-$xVarCorners/yVarCorners. The Equal-Land-Area Circle, defines noncompactness as a threshold occurring when a circle with origin at $C(D)$ and area $A(D)$, i.e. with radius $\sqrt{A(D)/\pi}$, captures less than half the area of D (Angel and Parent, 2011, p.93). Finally, we have Y-Symmetry, the area of district D overlapping with the reflection of D around a vertical line going through $C(D)$, divided by $A(D)$, and X-Symmetry, which is the same except for reflection of D around a horizontal line going through $C(D)$.

We calculate these features from the US Census TIGER shapefiles, and derive all measures from the coordinate sets which define districts according to that file. This is as fine-grained a resolution as possible; though many of these measures change depending on the resolution used, we argue that using the highest-resolution data possible reduces measurement error and subsequently improves modeling performance.

Appendix B How Paired Comparisons Fails

The method of paired comparisons has been touted for more than a century and a half for its two key advantages. First, this approach puts fewer demands on survey respondents than asking respondents to do a full ranking. That is, to produce a ranking of n items requires the choice among $n!$ possible rankings, whereas the same information can be elicited with only $\binom{n}{2}$ paired comparisons. This is not trivial since $n! \gg \binom{n}{2}$; for example, with $n = 20$, we have $20! = 2.4 \times 10^{18}$, or 2 quintillion possible rankings, whereas $\binom{20}{2} = 190$ paired comparisons is large but still manageable in a single survey (and may even be reduced; see Mitliagkas, Gopalan, Caramanis, and Vishwanath 2011). For these reasons, Converse and Presser (1986, p.28) comment on a historical example with only 13 items: "Tasks of this scope were soon seen as much too difficult..., and in our own time, rank orders of this size are all but invisible in the literature". Thus, if full ranking is used, the best practice has been "not to use lists longer than three or four items" (Gideon, 2012).

Second, Thurstone's approach only requires simple questions that are easy to understand, concrete, and specific. With it, we ask a respondent which among a pair of legislative districts is more compact, and then repeat this simple question multiple times with different pairs of districts. Then, after eliciting information in this manner, the researchers combine these binary decisions into a ranked scale (using Guttman scaling or a more sophisticated approach accounting for measurement error; e.g., Mitliagkas, Gopalan, Caramanis, and Vishwanath 2011). The method assumes all respondents will use the same unidimensional scale to make their choices for all their paired comparisons (an issue we return to). The supposed advantage of this approach is that respondents are asked only what they know (a paired comparison) and researchers do what they are better at, which is taking on the complicated task of inferring the underlying full ranking from all the elicited information.

To apply this method, we conducted multiple iterated rounds of pre-testing and cognitive debriefing while adjusting question wording and how the districts appeared[1]. But

[1] All districts are visualized at maximally high resolution to ensure that no features such as coastline are

despite dozens of trials over many months, testing numerous variations, and with a wide range of research subjects, online and in person, our inter- and intracoder reliability statistics were rarely much above random chance. To see what we found, consider a simple experiment with 40 respondents (in this case on Amazon's Mechanical Turk), each asked to choose the more compact district from each of twenty pairs, producing a 20-length binary decision vector. This survey enabled us to compare the percent agreement among the 20 decisions for each of $\binom{40}{2} = 780$ pairs of respondents. Figure 1 gives a histogram of these percent agreements (in blue, marked "paired", computed as a density estimate). For comparison, we also generate a placebo test, under the null hypothesis of no agreement, by randomly generating 780 pairs of 20-length vectors and computing from them the percent agreement and plotting its histogram (white with a black outline, marked "Random"). (We discuss the "Ranking" figure in the next section.)

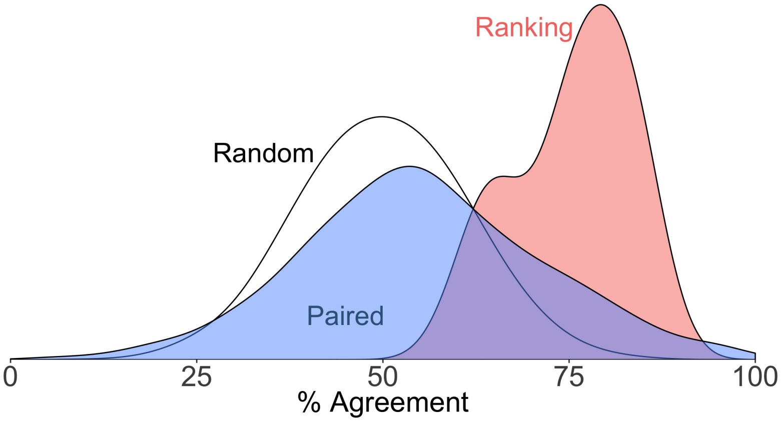

Figure 1: Intercoder Reliability of Thurstone's Paired Comparisons (blue histogram), full ranking (salmon histogram), and a random placebo distribution (white histogram), all using density estimation.

As expected when comparing coin flips, the random placebo percent agreement is centered at 50%. In contrast, the paired comparison percent agreement histogram is shifted farther to the right than the placebo histogram, but the mean only moves to 54%, leaving the two distributions with considerable overlap. Put differently, the best we could do with the method of paired comparisons, even before the step of turning paired decisions into

lost.

rank orders, is results with unacceptably low levels of intercoder reliability.

We now rule out the possibility that these results are due to different people having incompatible notions of compactness by studying intracoder reliability. To do this, we waited two weeks, randomly shuffled the order of the 20 paired comparison questions, and administered the survey to the same people. (Of the 40 people, only one mentioned, on post-survey cognitive debriefing, that "some" of the districts may have been the same as the first week.)

These results appear in Figure 2 (also as a blue histogram marked "Paired") and are more distinct from the random placebo test (in white with a black outline marked "Random") than with intercoder reliability in Figure 1, as would be expected. The mean of the paired comparison histogram is now at 65% agreement, although the overlap with the random distribution is still large. (We discuss the third histogram in the next section.)

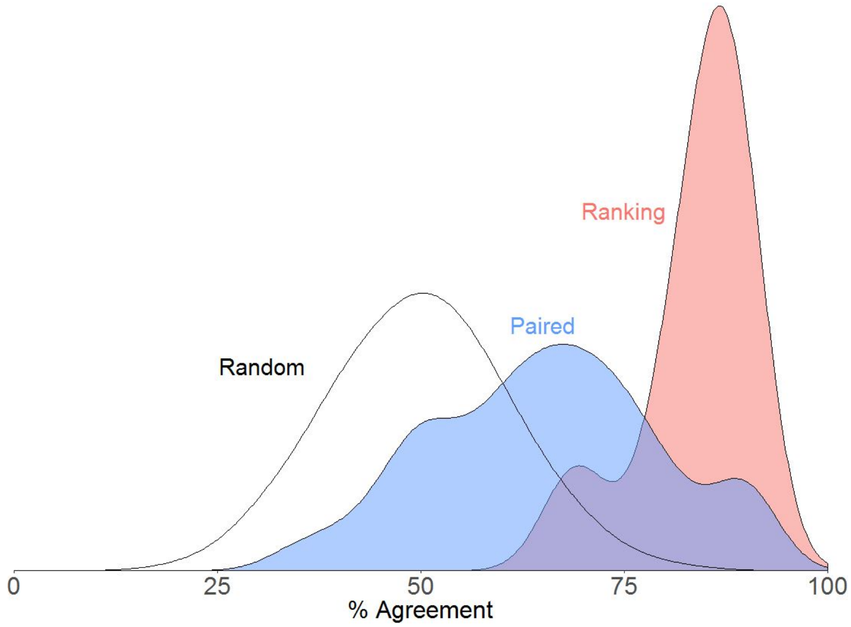

Figure 2: Intracoder Reliability of Thurstone's Paired Comparisons (blue histogram), full ranking (salmon histogram), and a random placebo distribution (white histogram), all using density estimation.

We thus conclude that these standard, best practice approaches are inadequate, at least for our application, and turn to an alternative. See Section 3.

Appendix C Compactness Data and Software

We offer additional details here of how we collected data for our experiments and how other researchers can do the same. We then outline data we make available on the compactness of numerous state legislative and congressional districts.

Data Collection To construct training and test sets for our various experiments, we use a set of 20,160 district shapes, including all congressional districts 1823–2013 and the last two cycles of state legislative districts. We obtained the shape files and other geographic data for congressional districts from Lewis, DeVine, Pitcher, and Martis (2013) and state legislative districts from McMaster, Lindberg, and Van Riper (2003).

To ensure we have variation in districts according to existing measures, we begin with a preliminary compactness ranking by ordering these districts based on an average of each district's Reock, Polsby-Popper, and Convex Hull scores. We create six groups of districts using systematic random sampling — to ensure a spread over the entire range of compactness — using a random start without replacement across groups — to avoid overlap among the groups. For the cross-validation in Section 4.1, we drew 100 districts. For our out-of-sample validations in Section 4.2, we collected 20 districts (to accommodate respondent time constraints).

We tested a variety of different instructions to our respondents. Here is a simple version we used for our online administration for full ranking. [We found the sentences in square brackets below useful for respondents, such as some from Mechanical Turk, who are not as familiar with the concept of compactness or the idea of legislative districts. Experiments we conducted among those familiar indicate that these passages do not affect the resulting rankings.]

> The law requires that legislative districts for the US congress and many state legislatures be "compact". The law does not say exactly what district compactness is, but generally, people think they know it when they see it. [One dictionary definition of compactness is "joined or packed together closely and firmly united; dense; arranged efficiently within a relatively small space." Some characteristics of districts people view as noncompact are wiggles, arms, noncontiguous segments, river-like features, or being much longer

> than wide. Compact districts look more densely packed, like rectangles, circles, or hexagons.]
>
> Here's your task: Below is a group of legislative districts, randomly ordered. Order the districts from most compact (at the top left) to least compact (at the bottom right) according to your own best judgement, by dragging and dropping. [We have many individuals performing this task, and the more your ranks are similar to others', the better you will have done.]

For paired comparisons, we changed the second paragraph to ask respondents to choose the more compact district of the two presented to them.

Our undergraduate respondents ranked 100 districts in a conference room with a long set of connected tables. We printed out pictures of each district, along with an identifying number, on a card measuring 4.25×5.5" (one quarter of a standard 8×11.5" paper). We asked each respondent to order the cards from most to least compact. As described in Section 3.1, we experimented with different sets of instructions, and with respondents working alone and in pairs, but we found no difference in intercoder or intracoder reliability as a result.

We asked the Mechanical Turk workers who ranked 100 districts to print out twenty-five sheets of paper with four districts each, and then to cut each in quarters and to follow the same instructions we gave our undergraduates. We asked for and received cell phone photos from the Turkers at each stage, to help ensure the task was completed as designed.

The undergraduates and Mechanical Turk respondents each took about 45–90 minutes to rank 100 districts. In order to reach a larger number of respondents, and especially to avoid charges of diverting public officials from performing their duties, we conducted our out-of-sample predictions with 20 districts. We chose this number by repeated experimentation with undergraduates, until we were able to get the time necessary to complete the task to under ten minutes. Most took 7–10 minutes.

Data Availability and Future Research For each of 20,160 congressional and state legislative districts, we compute the degree of compactness (Section 3.2) and an uncertainty estimate (see Section E). We make all these data, as well as the ranking data we collected to generate our model, publicly available as a companion to this paper, as well as software to estimate compactness in other districts or geographies. We think further analyses of these data may shed light on many venerable political science questions, such as compactness' relationship with balance between the parties, the existence of partisan gerrymandering, and the extent of racial fairness.

These data suggest many important questions worthy of further analysis. To illustrate, we examine compactness in four states frequently mentioned in the press as examples political gerrymandering. In Maryland's 2016 congressional elections, Republicans received 37% of the state's vote but only one of seven congressional seats. In Pennsylvania, despite winning approximately 46% of the two-party vote share in 2016, Democrats won only 5 of 18 congressional districts. In North Carolina, Democrats won 47% of the vote in 2016, but won only 3 of 13 congressional seats. Similarly, in Ohio, the Democratic vote was 42% while Democrats hold only 3 of 16 seats. A full partisan symmetry analysis would need to be conducted to evaluate whether these results were fair to the political parties (Gelman and King, 1994; King and Browning, 1987), but this prima facie evidence certainly suggests further analysis is worthwhile.

Our model predicts the rank a district would be given by a human coder (given only the shape of the district), with rank 1 being most compact and higher numbers indicating higher levels of noncompactness. We thus compute this *noncompactness* measure, using our methods, for each congressional district in each of these four states, for every new redistricting since 1893. We then average compactness for all districts within each state and, in Figure 3, plot the averages over time.

Interestingly, noncompactness dramatically increases in Ohio and Pennsylvania beginning in the mid-1960s, shortly after *Baker v. Carr* (1962) mandated redistricting to achieve equal district populations. Maryland and North Carolina, in contrast, show no such increase. Is this because these states had high noncompactness levels to begin with? Could noncompactness have been at an effective maximum? Did redistricters from the majority parties in Ohio and Pennsylvania take advantage in ways those in North Carolina and Maryland did not? Did the progress (or overreaching) on behalf of minorities in two of the states take a different path than in the other two? Or might the differences be due to other factors, such as local political subdivisions, communities of interest, or natural fea-

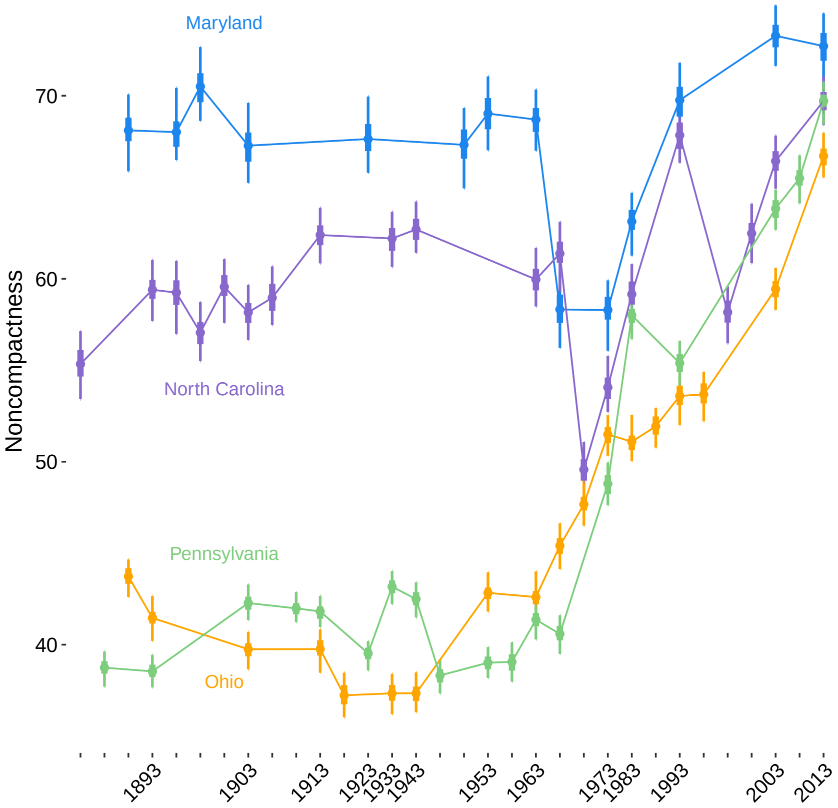

Figure 3: Time series plots of average district compactness (and 95% confidence intervals) in congressional districts for four states often claimed in the media to be political gerrymanders.

tures of the states being taken into account in districting in different ways? We encourage future researchers to delve into these and the numerous other questions these data suggest.

Appendix D Ensemble Modeling

Since we have six distinct training sets, we construct our ensemble using six distinct training sets. We do so in three steps: (1) fit each of four models described below to each of the training sets; (2) calculate each model's predictions for our universe of 20,160 districts; and (3) average each of the 24 predictions elementwise to produce a final ensembled compactness measure. We offer more detailed information about each step in our replication data file and information about each model here:

Linear regression with variable selection We chose covariates via cross-validation, iteratively dropping the worst-performing covariate and observing the increase in cross-validation accuracy. We followed this procedure until the cross-validation accuracy began to decrease.

The selected main variables are: Polsby-Popper, Boyce-Clark, Convex Hull, Significant Corners, Significant Corner coordinate variance ratio, X Symmetry, Y Symmetry, District Area, Point coordinate variance ratio, Variation in Line Segment Length. As well, included are the following interactions: Polsby-Popper * Convex Hull, Polsby-Popper * X Symmetry, Polsby-Popper * Y Symmetry, X Symmetry * Y Symmetry, Polsby-Popper * Significant Corners, Convex Hull * Significant Corners, Polsby-Popper * X Symmetry * Y symmetry, and Corner coordinate variance ratio * point coordinate variance ratio.

Random Forest Random Forests, which consist of bootstrap-aggregated decision trees, are among the most commonly used machine learning models in practice. We train our random forest using 2,000 trees and the default settings in the `randomForest` library (Liaw and Wiener, 2002).

AdaBoosted decision trees ADTs are structurally similar to random forests, but with each tree trained on a version of the data reweighted based on the previous tree's residuals (Kaufman, Kraft, and Sen, 2018). We use 2,000 trees, an interaction depth of 3, and otherwise default settings in the `gbm` library (Ridgeway, 2015).

SVM Support vector machine regression is also widely applicable and requires little tuning. We train using the default settings for the `e1071` library (Meyer, Dimitriadou, Hornik, Weingessel, and Leisch, 2017), which includes the radial kernel.

All together, we produce 24 models and predictions: four methods each for six different and non-overlapping training sets. We therefore have 24 predictions for each of the 20,160 districts in our test set; we average the predictions for each district and produce a final compactness score which we include in our replication file.

Appendix E Uncertainty Estimation

Prior approaches to compactness do not define theoretical quantities of interest separate from their proposed empirical measures. As a result, the statistical properties of these measures have not been defined or evaluated. And without this key distinction, estimates of uncertainty (based on deviations from a quantity of interest) have not been introduced.

Our theoretical quantity of interest is perceived compactness, which we theorize is common across educated people. Like all existing compactness measures, our proposed measure is a deterministic function of only district shape. We treat our measure as a prediction of perceived compactness and evaluate its uncertainty based predictive accuracy. Uncertainty estimates are then a function of (a) measurement error in eliciting views of compactness from any individual, (b) actual variation across individuals in their views, and (c) predictive inaccuracy.

We offer uncertainty measurements for both a single compactness measure and the difference in two compactness measures. For a single measure, we plot all our data used to evaluate out-of-sample our compactness predictions in Figure 4 (left panel), with our predicted compactness horizontally by the absolute deviation from the truth vertically. We then sort these data into 20 bins defined on the horizontal axis. Then we calculate for each bin the quantiles of the absolute deviations from the out-of-sample truth. We record the 20 points that are at the 50% quantile and the 20 at the 95% quantile. Each fairly closely follows a quadratic curve and so we fit a polynomial regression and add these to the graph (black for 50% and red for 95%). The height of the black curve then represents the average amount of uncertainty we should expect and the height of the red curve indicates, for any given prediction, half the width of the 95% predictive interval. The red curve happens to have a relatively simple and easy-to-use form. Let c denote predictive compactness. Then half the 95% confidence interval is simply $c - 2 - 0.01c^2$. So for a highly noncompact district with a score of 90, the 95% interval is ± 7.[2]

Finally, Figure 4 (right panel) gives uncertainty estimates for differences between two

[2] We also perform this procedure treating positive and negative errors separately, producing two separate quadratics rather than one. This less efficient procedure produces similar but less conservative predictive intervals, and so stick to the procedure in the text.

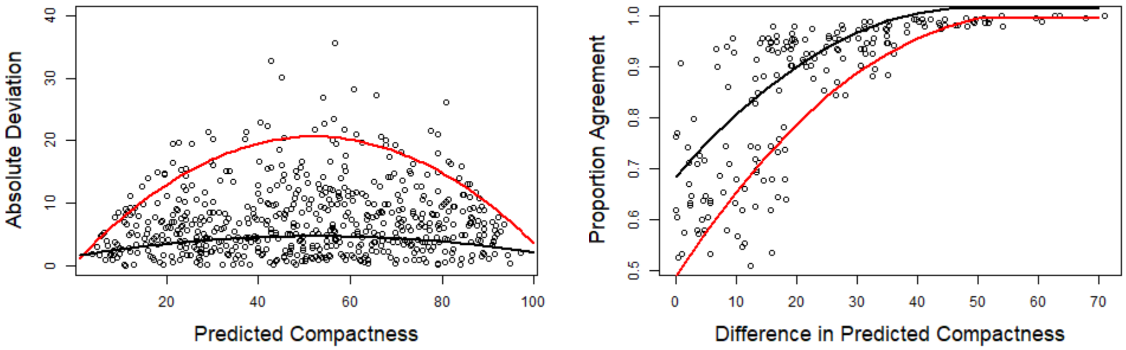

Figure 4: Uncertainty Intervals for a single compactness score (left panel) and for the difference in two compactness scores (right panel). Each graph plots the uncertainty on average (black line) and that which bounds 95% of likely outcomes (red line).

predicted compactness values. We do this by computing the percent agreement on those two districts (vertically) by absolute differences in predicted compactness for two districts (horizontally). We then again create 20 bins on the horizontal axis and compute the 50% and 95% quantiles, and fit smoothed lines (which are also quadratics, except to the top right with few data points). For example, the red line indicates, for a difference of 10 in predicted compactness between districts i and j, on average 80 evaluators out of 100 will agree that district i is more compact than district j, and only at a difference of 5 out of 100 will fewer than 75 judges out of 100 agree.

Appendix F Feature Correlations

This table gives correlations between our measure and related and component measures. We do that here the wrong way, by stacking up all data from all districts in our data set, and allowing bias due to causal heterogeneity. We do it the right way, one legislature at a time, in the next section of this appendix.

We code all measures so that higher levels indicate noncompactness. The Pearson column gives the raw correlations between our measure and others. The Spearman column is the Spearman correlation between our measure, ranked, and the other measures, also ranked. The Pearson_avg column is the average Pearson correlation across each state-chamber-year. We also examine our ensemble as a weighted average of its component measures. We produce three unweighted averages of various compactness measures and calculate their correlation to our ensemble measure. The bottom three rows all indicate these naive averages of various sets of measures. The OldFeatures measure is the naive average of all previously existing compactness measures, after they have been direction-corrected. AllFeatures includes all measures we use in our ensemble; SmartFeatures uses measures we consider to be good predictors of compactness. The SmartFeatures measures consist of X-Symmetry, Y-symmetry, Reock, Polsby-Popper, Convex Hull, and the Bounding Box measure. The OldFeatures measures are the Varcord Ratio, Boyce-Clark, Length/Width, Jaggesness, Convex Hull, Bounding Box, Reock, Polsby-Popper, and Schwartzberg.

feature	Pearson	Spearman	Pearson_avg
Convex Hull	0.89	0.89	0.89
Polsby-Popper	0.92	0.92	0.92
Reock	0.62	0.62	0.62
Bounding Box	0.85	0.85	0.85
Y Symmetry	0.60	0.60	0.60
X Symmetry	0.58	0.58	0.58
Corners	0.31	0.31	0.31
Boyce-Clark	0.34	0.34	0.34
Length/Width	0.04	0.04	0.04
Jaggedness	0.08	0.08	0.08
Parts	0.09	0.09	0.09
Bounding Circle Area	0.01	0.01	0.01
Cornervar Ratio	0.03	0.03	0.03
OldFeatures	0.92	0.92	0.92
AllFeatures	0.85	0.85	0.85
SmartFeatures	0.94	0.94	0.94

Why Not to Use Single Measures or Unweighted Averages

Given the results in the previous table, it may be tempting to use a simpler measure than our ensemble, perhaps Polsby-Popper, since it correlates so highly. This would be inappropriate, since the overall correlation masks important heterogeneity, which indicates that the correlations in the table are biased.

For example, it would be easy to design a map such that any two measures contradict each other in rankings, or another map where they agree. As such, we cannot be confident that for any given map, that any measure evaluated in any way based on prior data will correlate highly with our ensemble. Yet, we know from the paper that our measure — and no other existing measure — will reflect the degree of compactness that human beings know when they see. To illustrate this point, we perform a lengthy analysis in which we systematically evaluate 778 legislatures, every unique state-level legislative map (one state's chamber in one session, including upper chamber, lower chamber, and Congressional maps) for which we have geospatial data.

To take one example, Polsby-Popper correlates with our ensemble at 0.99 in Indiana, but -0.7 in Maryland. Without also calculating our ensemble measure, districters would draw the wrong conclusions about compactness using only Polsby-Popper.

In the tables that follow, we then systematically explore this relationship and perform two analyses: (1) for any given chamber, we record which measure correlates most highly, and (2) for any given measure, we record the percent of the cases that each measure correlate negatively with our measure.

The measure that correlates highest with our measure in the largest number of legislatures is Convex Hull, but this occurs in only 49.6% of the data sets — followed by the Polsby-Popper in 33.3%, Grofman in 7.1%, X-axis Symmetry in 2.1%, Reock and Y-axis Symmetry at 1.7% each, and Boyce-Clark at 1.4%; even measures such as the area of the minimum bounding circle and the number of discontiguous polygons correlate most highly in some situations. In other words, any existing measure can come out on top in approximating our measure depending on the particular features of the set of district shapes that make up the legislature, and so none of these measures alone can be used as a simpler replacement with our measure of what people know when they see, without checking the relationship first. This means that, using any measure but ours, will result in data sets where a human being looking at the districts will draw one conclusion and the measure will suggest the opposite: that is, the human observer will conclude that the measure is wrong. As we show in the paper, this is highly unlikely to occur with our new measure.

Indeed, every measure correlates negatively a significant portion of the time with our measure. Convex Hull and Grofman both do so in around 1% of the cases; Polsby-Popper, around 1.5%. Reock does so in 4% of the cases. X symmetry and Y symmetry correlate negatively in 6.5% and 4% of the cases, respectively. Significant Corners and Boyce Clark correlate negatively in 12% and 13% of the cases, respectively; Length/Width, a stunning 40%.

The conclusion here reinforces that in the paper: If one wishes to measure the type of compactness that all types of people seem to perceive — from undergraduates and Mechanical Turk workers to justices, legislators, and other public officials — then the measure we develop in our paper (coded from these perceptions) is the right choice. No other existing measure is trying to estimate that quantity of interest, nor do any of these measures happen to pick up the same information.

Feature	Proportion of Negative Correlations
Convex Hul	0.009
Polsby-Popper	0.014
Reock	0.039
Grofman	0.012
X-axis Symmetry	0.067
Y-axis Symmetry	0.044
Significant Corners	0.122
Boyce-Clark	0.131
Length/Width	0.404
Jaggedness	0.446
Parts	0.170
Circle Area	0.546
Cornervar Ratio	0.456
OldFeatures	0.014
AllFeatures	0.017
SmartFeatures	0.005

Feature	Highest Correlation	Lowest Correlation	Smallest Absolute Correlation
Convex Hull	0.996	0.222	0.222
Polsby-Popper	0.996	-0.770	0.001
Reock	0.989	-0.408	0.001
X-axis Symmetry	0.981	-0.635	0.001
Y-axis Symmetry	0.974	-0.833	0.014
Significant Corners	0.947	-0.719	0.009
Boyce-Clark	0.926	-0.795	0.000
AllFeatures	0.991	-0.446	0.033
SmartFeatures	0.992	0.037	0.037
OldFeatures	0.997	-0.453	0.069

References

Angel, Shlomo and Jason Parent (2011): “Non-compactness as voter exchange: Towards a constitutional cure for gerrymandering”. In: *Northwestern Interdisciplinary Law Review*, vol. 4, p. 89.

Converse, Jean M. and Stanley Presser (1986): *Survey Questions: Handcrafting the Standardized Questionnaire*. Thousand Oaks, CA: Sage Publications.

Gelman, Andrew and Gary King (May 1994): “A Unified Method of Evaluating Electoral Systems and Redistricting Plans”. In: *American Journal of Political Science*, no. 2, vol. 38, pp. 514–554. URL: j.mp/unifiedEc.

Gideon, Lior (2012): “The art of question phrasing”. In: *Handbook of survey methodology for the social sciences*. Springer, pp. 91–107.

Harris, Chris and Mike Stephens (1988): “A combined corner and edge detector.” In: *Alvey vision conference*. Vol. 15. 50. Citeseer, pp. 10–5244.

Kaufman, Aaron, Peter Kraft, and Maya Sen (2018): “Improving Supreme Court Forecasting Using Boosted Decision Trees”. In: URL: j.mp/sctfore.

King, Gary and Robert X Browning (Dec. 1987): “Democratic Representation and Partisan Bias in Congressional Elections”. In: *American Political Science Review*, no. 4, vol. 81, pp. 1252–1273. URL: j.mp/parSym.

King, Gary and Langche Zeng (2006): “The Dangers of Extreme Counterfactuals”. In: *Political Analysis*, no. 2, vol. 14, pp. 131–159. URL: j.mp/dangerEC.

Kong, Xianshu, Hazel Everett, and Godfried Toussaint (1990): “The Graham scan triangulates simple polygons”. In: *Pattern Recognition Letters*, no. 11, vol. 11, pp. 713–716.

Lewis, Jeffrey B, Brandon DeVine, Lincoln Pitcher, and Kenneth C Martis (2013): “Digital boundary definitions of united states congressional districts, 1789–2012”. In: *Data file and code book*. URL: j.mp/jblkmaps.

Liaw, Andy and Matthew Wiener (2002): “Classification and Regression by randomForest”. In: *R News*, no. 3, vol. 2, pp. 18–22.

MacEachren, Alan M (1985): “Compactness of geographic shape: Comparison and evaluation of measures”. In: *Geografiska Annaler. Series B. Human Geography*, pp. 53–67.

McMaster, Robert B, Mark Lindberg, and David Van Riper (2003): “The national historical geographic information system (NHGIS), Version 11.0”. In: *Proceedings 21st International Cartographic Conference*, pp. 821–828.

Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch (2017): *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.6-8. URL: https://CRAN.R-project.org/package=e1071.

Mitliagkas, Ioannis, Aditya Gopalan, Constantine Caramanis, and Sriram Vishwanath (2011): “User rankings from comparisons: Learning permutations in high dimensions”. In: *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*. IEEE, pp. 1143–1150.

Nielsen, Frank and Richard Nock (2008): “On the smallest enclosing information disk”. In: *Information Processing Letters*, no. 3, vol. 105, pp. 93–97.

Ridgeway, Greg (2015): "gbm: Generalized Boosted Regression Models". In: R package version 2.1.1. URL: https://CRAN.R-project.org/package=gbm.