8TH ANNUAL
MIT SLOAN
SPORTS ANALYTICS CONFERENCE
FEBRUARY 28 - MARCH 1, 2014 HYNES CONVENTION CENTER PRESENTED BY ESPN

Knowing what we don't know in NCAA Football ratings: Understanding and using structured uncertainty

Daniel Tarlow, Thore Graepel, Tom Minka
Microsoft Research
Cambridge, United Kingdom
Email: {dtarlow,thoreg,minka}@microsoft.com

Abstract

There is a great deal of uncertainty in the skills of teams in NCAA football, which makes ranking teams and choosing postseason matchups difficult. Despite this, standard approaches (e.g., the BCS system) estimate a single ranking of teams and use it to make decisions about postseason matchups. In this work, we argue for embracing uncertainty in rating NCAA football teams. Specifically, we (1) develop a statistical model that infers uncertainty in and correlations between team skills based on game outcomes, (2) make proposals for how to communicate the inferred uncertainty, and (3) show how to make decisions about postseason matchups that are principled in the face of the uncertainty. We apply our method to 14 years of NCAA football data and show that it produces interesting recommendations for postseason matchups, and that there are general lessons to be learned about choosing postseason matchups based on our analysis.

1 Introduction

Ranking NCAA football teams is challenging for a number of reasons. Primarily, there are many teams and few games played. In addition, most games are played within leagues, and cross-league games rarely match up the strongest teams in the respective leagues. Finally, game outcomes are noisy; that is, observing team A winning over team B does not guarantee that A is the stronger team. Yet at the end of each season, we hope that the result of one game (or starting next year, a four-team tournament) will allow us to confidently proclaim a champion. Current approaches to ranking NCAA football make little acknowledgement of uncertainty in team rankings, but there can be significant controversy about the final decisions. In extreme cases (e.g., in 2003-04) a single champion cannot be agreed upon even after all bowl games have been played.

The point of this work is to argue that we should embrace, quantify, and communicate uncertainty in the ranking of NCAA football teams, and that when it comes time to make postseason matchup decisions, the uncertainty should be taken into account. We take a computer-based statistical approach. There are disadvantages to such approaches, including a difficulty in understanding "style points" or qualitative properties of the game (although maybe these are the aspects that most easily deceive humans). However, statistical approaches have many advantages including understanding how skills apply transitively along the graph of games played, and they can potentially be less susceptible to biases. In this work, we emphasize another benefit: the ability to capture sophisticated notions of uncertainty, and to make rational decisions in the face of uncertainty.

We make three specific contributions: (1) a statistical model that infers structured uncertainties in team skills, capable of concluding, for example, that we are quite certain about relative within league strengths but very uncertain about certain cross-league strengths; (2) proposals for how to visualize this uncertainty to more accurately convey what we do and do not know about relative team strengths; and (3) showing how to make principled decisions under this uncertainty, specifically proposing a method for choosing postseason matchups that acknowledges the uncertainty and chooses matchups that are expected to minimize controversy; that is, to make us as confident as possible that the team crowned champion is indeed the best. Experimentally we apply our method to NCAA football seasons from 2000-01 through 2013-14 and consider how to choose championship games and four-team tournaments so as to minimize expected controversy. We synthesize some general lessons that we believe important to consider when choosing postseason matchups.

2 Why we should track rating uncertainty?

A central argument in this work is that we should quantify our uncertainty about the playing strength of teams. Why would this be important? Any statistical estimation of ratings from finite data is subject to uncertainty, typically with

more data leading to less uncertainty. Tracking that uncertainty at the individual team level is very useful: First, ratings can be given with confidence intervals to indicate the remaining uncertainty. Second, additional data can be incorporated easily because the uncertainty of the current estimate influences the degree to which new data should influence the estimate. Third, we can take into account knowledge of uncertainty to arrange additional matches such that they are most informative about questions of interest, such as which is the best team.

In addition, we argue for the importance of *pairwise uncertainty*, or covariance information, which quantifies how knowledge of one team's rating depends on knowledge about another team's rating. Suppose, there are three teams, A and B in league 1, and C in league 2, and that teams A and B play a match which A wins. We can conclude that the rating of A is likely to be greater than that of B. There is no correlation (yet) between any of the cross-league pairs such as A and C or B and C. However, we do know that if the rating of B is likely to be high, then the rating of A is also likely to be high: A positive correlation between the ratings of A and B has been induced. As a consequence, if team B wins over, say, team C (in a cross-league match) we can conclude—in a statistical sense—that team A is also stronger than team C, mediated by its correlation with team B. Hence, keeping track of pairwise correlations allows rating information to flow along the edges of the match graph and helps us assess the expected gain in information for the rating of a third team A from observing a match outcome between two teams B and C.

3 Statistical Ranking Model

Ranking under uncertainty has found great success both inside and outside the sports world. One of the most popular approaches is to assume that there is an unknown real-valued rating (or "skill") variable associated with each player (or team), and the difference between two players's rating determines the probability of a player winning. The origin of this approach lies in the well-known Elo model for chess (Elo, 1978), which relies on point estimates of ratings based on game outcomes. Elo has subsequently been extended to include a Bayesian uncertainty estimate in the Glicko system (Glickman, 1999), and has been generalized to multiple teams in the TrueSkill system, which is used within Microsoft's XBox Live service to assess player skills for leaderboards and matchmaking (Herbrich, 2007). TrueSkill represents its knowledge about team ratings as a probability distribution over team ratings, which incorporates the system's uncertainty.

Our first contribution is to adapt the TrueSkill algorithm to NCAA Football. We modify the model to include an extra rating for home field advantage and decompose each team's rating into the sum of a league rating and an individual team rating. More importantly, we extend the inference so as to represent a full covariance matrix (all pairwise correlations) between team ratings.[1]

We are working within the framework of Bayesian inference to implement our models. The philosophy of Bayesian modelling is as follows. First, create a probabilistic model of how the observed data (game outcomes) has been generated, given the (unknown) latent variables (team ratings). In the inference step, the latent variables are equipped with prior belief distributions and the generative process is inverted: Update the prior beliefs over latent variables in light of observed data, resulting in posterior belief distributions over latent variables. Inference is performed automatically using the Infer.NET system (Minka, 2012), which uses a Bayesian inference algorithm known as Expectation Propagation (EP) (Minka, 2001). It would be possible to use other inference algorithms, e.g., Markov Chain Monte Carlo, but for TrueSkill-like models, EP is fast and accurate.

Pseudocode specifying the full probabilistic model is given in Fig. 1, roughly according to Infer.NET syntax. Hyperparameters of the model were chosen in accordance with (Herbrich, 2007). Given this generative process we observe the values of `team1Won[g]` for all games `g`, and infer a probability distribution over team and league ratings, and home team advantage strength. The posterior belief distribution is a multivariate Gaussian distribution characterized by its mean vector (with as many components as there are teams), and its covariance matrix, which describes the correlation structure of the posterior. See Fig. 2 (a) for an example inferred covariance matrix. There is a row and column for each team, both of which have been sorted according to league structure (the last, large block corresponds to teams not in the upper FBS division). There are strong within league correlations (the on-diagonal blocks), and some correlations across certain leagues, but also many leagues with little correlation (dark off-diagonal blocks) where we are quite uncertain about relative skills.

[1] We also add a constraint that the average skill is equal to the prior mean, to break a symmetry in the model.

```
for team in allTeams:
   rating[team] = GaussianFromMeanAndVariance(25, (25/3)^2)
for league in allLeagues:
   leagueRating[league] = Gaussian(0, (25/3)^2)
homeAdvantage = GaussianFromMeanAndVariance(5, 25)
for g in range(numGames): // loop over games
   game = allGames[g]
   team1Rating = rating[game.team1] + leagueRating[game.team1League]
   team2Rating = rating[game.team2] + leagueRating[game.team2League]
   if game.homeTeam == game.team1:  // team1 was home
      team1HomeAdvantage = homeAdvantage
   else if game.homeTeam == game.team2:  // team2 was home
      team1HomeAdvantage = -homeAdvantage
   else:  // neutral site
      team1HomeAdvantage = 0
   // How well teams played in the given game; whoever performed best won
   team1Performance = team1Rating + team1HomeAdvantage + Gaussian(0, (25/6)^2)
   team2Performance = team2Rating + Gaussian(0, (25/6)^2)
   team1Won[g] = (team1Performance > team2Performance)
```

Figure 1: Probabilistic program pseudocode specifying our model of team skills and game outcomes.

4 Communicating Structured Uncertainty

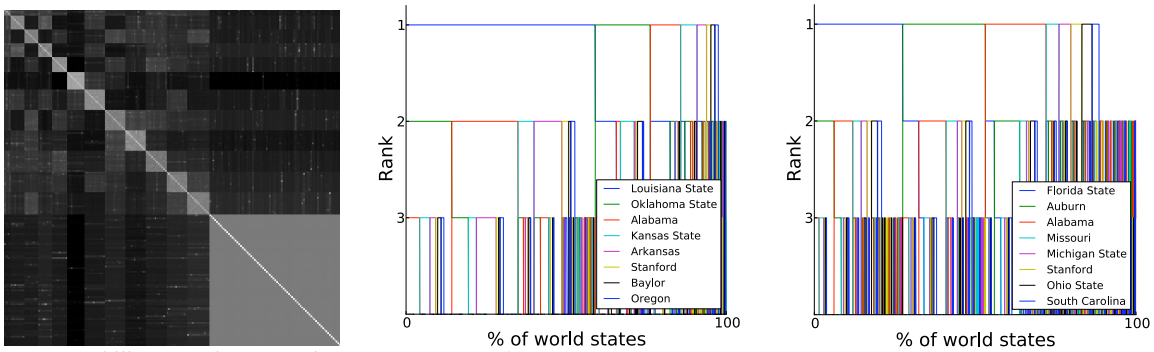
A primary challenge is how to communicate the results of the inference in a manner that is as comprehendible as possible. We explore methods of visualization here, making two suggestions. First is to infer and visualize the covariance matrix over team skills. Our second proposal is to produce a plot that serves two purposes: (1) it shows the probability under the inferred posterior of different teams being the top-ranked team; (2) it allows us to visualize correlation in beliefs over team rankings. The plot is constructed as follows. First, draw a large number (e.g., 10000) of samples from the posterior distribution over team skills. For each sample, sort the skills in descending order to get a sample ranking. Produce a team order by sorting teams by the number of samples in which they are the top-ranked team. Finally, sort the samples under a comparator that places a sample $\boldsymbol{s}$ less than a sample $\boldsymbol{s'}$ if the top-ranked team in $\boldsymbol{s}$ comes earlier in the team order than the top ranked team in $\boldsymbol{s'}$. If they agree on the top ranked team, sort under the same comparator but using the second-ranked team, and recurse until some disagreement is found or all elements have been examined, at which point the samples are equal.

We can then take these ordered samples and plot them. We use the x-axis to range over samples, and the y-axis shows the ordering of teams for a given sample. Intuitively, the x-axis represents the different possible states of the world with more space given to rankings that are more likely under the model. Results appear in Fig. 2 (b) and (c). In Fig. 2 (b) we plot the least controversial year (more on controversy in the next section). In the top left region of the plot, we see that Louisiana State (LSU) is clearly the most likely to be ranked #1; however, there is some probability assigned to world states where some other team has the greatest skill, showing that we are never fully certain about ranking the top teams. The empty space on the far right of the top row (and elsewhere) denotes probability that some team other than the ones that appear in the legend was ranked #1. In Fig. 2 (c), we illustrate how these plots can be used to assess correlations in rankings. Florida State is most likely to be ranked #1, but by looking at the second rank in the world states where either Auburn or Alabama is ranked #1, we see that Florida State is less likely than Alabama or Auburn, respectively, to be ranked #2. This shows that the model believes the skills of Auburn and Alabama are highly correlated. More results appear in the Appendix.

5 Decision Making Under Uncertainty

Here we consider how to use the inferred posterior distributions over skills to decide on postseason matchups. We adopt the approach of Bayesian decision theory (see e.g., Gelman et al. (2003)), which proceeds as follows:

- Enumerate the space of all possible decisions and outcomes. In our case, decisions are postseason matchups $\boldsymbol{m}^p$, and outcomes are the team that is deemed the champion t^* and the skills of all teams $\boldsymbol{s}$.
- Define a loss function $\Delta(t^*, \boldsymbol{s})$ that measures how unhappy we are with a given outcome. The loss function we use is a measure of controversy: if the team crowned champion is not the team with greatest skill, then we suffer a loss of 1; otherwise we suffer no loss.
- Choose the postseason matchups that minimize expected loss conditional upon the matchups.



(a) 2011-12 Skills Posterior Covariance (b) 2011-12 (c) 2013-14

Figure 2: (a) Example learned covariance matrix over team skills. (b,c) Visualization of distribution of rankings for (b) 2011-12, the least controversial year we considered, and (c) 2013-14, a year where correlations in team skills can be seen (between Auburn and Alabama). More results in Appendix. Best viewed in color.

The crux then is to compute expected controversy. To do so, we must sum over the possible game results $\boldsymbol{r}^p$ of the chosen matchups and integrate over the uncertainty in skills:

$$\text{ExpectedControversy}(\boldsymbol{m}^p) = \sum_{\in \text{PossibleResults}(\quad)} \int p(\boldsymbol{r}^p \mid \boldsymbol{m}^p, \boldsymbol{m}^r, \boldsymbol{r}^r) p(\boldsymbol{s} \mid \boldsymbol{m}^p \boldsymbol{r}^p, \boldsymbol{m}^r, \boldsymbol{r}^r) \Delta(t^*(\boldsymbol{r}^p), \boldsymbol{s}) d\boldsymbol{s}$$

where $\boldsymbol{m}^r$ and $\boldsymbol{r}^r$ are regular season matchups and results, $p(\boldsymbol{s} \mid \boldsymbol{m}^r, \boldsymbol{r}^r, \boldsymbol{m}^p, \boldsymbol{r}^p)$ is the updated posterior distribution over skills that comes from observing the results of the postseason matchups and re-running inference, and $t^*(\boldsymbol{r}^p)$ denotes the team that is (deterministically) crowned the champion given the postseason results. Since the expected controversy numbers range from 0 to 1, we report them as a percentage values.

To better understand the decision-making procedure, it is productive to consider an example. Let us take the 2013-14 season, where the model believes that the top two teams according to posterior mean are Auburn and Alabama, while Florida State has third largest posterior mean. If the skills of all three teams were uncorrelated and had the same posterior variance, then the minimum expected controversy matchup would choose the two teams with largest posterior mean—Auburn and Alabama. However, as discussed above, the model believes that the skills of Auburn and Alabama are highly correlated. If we were to match them up in a championship game, we would identify more surely which was stronger, but there would still be significant chance that Florida State was the best team, and thus the matchup would have a high expected controversy score. In contrast, if we match up Auburn against Florida State, then the two possible outcomes are Florida State wins or Auburn wins. If Florida State wins, the model lowers its estimation of the skill of Auburn, and as a consequence of the correlation, it also lowers its estimation of the skill of Alabama. The result is that we believe Florida State is stronger than both teams. If Auburn wins, then the question of whether Florida State is better than Auburn and Alabama is also resolved in the negative. Thus there is less expected controversy, and indeed, we will see in the next section that Auburn versus Florida State is the favored matchup.

The computation of expected loss for a given matchup can be done efficiently using Infer.NET. To choose a matchup, our strategy is to prune the space of all possible matchups and only consider those that might plausibly have minimum expected controversy (we restrict attention to matchups involving teams that are in the top 10 according to mean posterior skill), then to find the minimum by enumerating all plausible matchups.[2] This is tractable for choosing a single championship game, or for choosing matchups for a four-team tournament. To scale to matchups involving more teams, additional approximations would likely be required.

[2] We make one other minor approximation for computational efficiency's sake, which is to change the loss so that we suffer a loss of 1 if the team crowned champion is not the team with greatest skill *amongst the 40 best teams*, as measured by the mean posterior skill.

Table 1: Choosing championship matchups. Expected controversy score is reported in parentheses.

Year	Model Top 2	Minimum Controversy	BCS Choice
2000-01	Oklahoma v Washington (69.6%)	Oklahoma v Florida St. (66.3%)	Oklahoma v Florida St. (66.3%)
2001-02	Miami v Oregon (70.1%)	Miami v Nebraska (69.1%)	Miami v Nebraska (69.1%)
2002-03	Miami v Ohio St. (54.7%)	Miami v Ohio St. (54.7%)	Miami v Ohio St. (54.7%)
2003-04	Oklahoma v LSU (82.4%)	Oklahoma v USC (80.6%)	Oklahoma v LSU (82.4%)
2004-05	Oklahoma v USC (62.1%)	Oklahoma v USC (62.1%)	Oklahoma v USC (62.1%)
2005-06	Texas v USC (47.6%)	Texas v USC (47.6%)	Texas v USC (47.6%)
2006-07	Ohio St. v Florida (67.3%)	Ohio St. v Florida (67.3%)	Ohio St. v Florida (67.3%)
2007-08	Kansas v Virginia Tech (83.0%)	Kansas v Ohio St. (79.1%)	LSU v Ohio St. (88.7%)
2008-09	Oklahoma v Texas (79.8%)	Oklahoma v Utah (73.6%)	Florida v Oklahoma (80.4%)
2009-10	Alabama v Florida (66.1%)	Alabama v Cincinnati (56.6%)	Alabama v Texas (59.1%)
2010-11	Auburn v Oregon (53.0%)	Auburn v Oregon (53.0%)	Auburn v Oregon (53.0%)
2011-12	LSU v Oklahoma St. (47.3%)	LSU v Oklahoma St. (47.3%)	Alabama v LSU (52.7%)
2012-13	Florida v Notre Dame (65.3%)	Florida v Notre Dame (65.3%)	Alabama v Notre Dame (73.6%)
2013-14	Auburn v Alabama (80.8%)	Auburn v Florida St. (67.1%)	Auburn v Florida St. (67.1%)

Table 2: Choosing four-team tournaments. Expected controversy score is reported in parentheses.

Year	Model Top 4	Minimum Controversy
2000-01	Oklahoma v Washington, Florida St. v Miami (58.0%)	Oklahoma vs VA Tech, Washington vs Florida St. (57.6%)
2001-02	Miami v Oregon, Nebraska v Colorado (63.0%)	Miami v Colorado, Oregon v Nebraska (62.1%)
2002-03	Miami v Ohio St., Georgia v USC (51.4%)	Miami v Iowa, Ohio St. v Georgia (50.6%)
2003-04	Oklahoma v LSU, USC v Georgia (76.2%)	Oklahoma v LSU, USC v Miami (OH) (73.3%)
2004-05	Oklahoma v USC, Texas v Auburn (53.4%)	Oklahoma v Utah, USC v Auburn (52.8%)
2005-06	Texas v USC, Penn St. v Ohio St. (47.6%)	Texas v West Virginia, USC v Penn St. (45.3%)
2006-07	Ohio St. v Florida, Michigan v USC (61.6%)	Ohio St. v Boise St., Florida v Michigan (58.6%)
2007-08	Kansas v Virginia Tech, Missouri v LSU (78.5%)	Kansas v Hawaii, Missouri v Ohio St. (71.0%)
2008-09	Oklahoma v Texas, Texas Tech v Utah (66.7%)	Oklahoma v Boise St., Utah v Alabama (64.5%)
2009-10	Alabama v Florida, Cincinnati v Texas (49.2%)	Alabama v Texas, Cincinnati v TCU (47.3%)
2010-11	Auburn v Oregon, Stanford v Arkansas (52.7%)	Auburn v Wisconsin, Oregon v TCU (48.4%)
2011-12	LSU v Oklahoma St., Alabama v Kansas St. (47.3%)	LSU v Stanford, Oklahoma St. v Alabama (44.4%)
2012-13	Florida v Notre Dame, Alabama v Ohio St. (56.5%)	Florida v Ohio St., Notre Dame v Alabama (56.3%)
2013-14	Auburn v Alabama, Florida St. v Missouri (63.4%)	Auburn v Ohio St., Alabama v Florida St. (60.8%)

Choosing Championship Matchups. As a first experiment, we train on regular season games then use the method described in the previous section to find the single championship game that minimizes expected controversy. Each season is treated independently. For comparison, we also consider matching up the two teams with greatest posterior mean skill, and matching up the two teams that were chosen by the BCS system to play for the championship. In Table 1 we report the matchups and expected controversies.

The first thing to note about the results is that there are several seasons (5 of 14) where the three methods agree: our model and the BCS ranking system agree on which are the top two teams, and the matchup that minimizes expected controversy is to match these two teams up. This happens in 2002-03, 2004-05, 2005-06, 2006-07, and 2010-11. In 2000-01, 2001-02, and 2013-14, the model disagrees with the BCS ranking over who the top two teams are, but nevertheless thinks that the matchup chosen by the BCS system is the best pairing of teams. In 2003-04, the model agrees with the BCS system that Oklahoma and LSU were the two strongest teams, but it thinks that pairing Oklahoma versus USC would have led to less controversy. We wonder if this matchup would have avoided the split championship controversy that followed. Finally, there are a number of years where our recommended matchups have a more fundamental disagreement with the BCS. In 2007-08, 2008-09, and 2009-10, the minimum expected controversy criterion suggests that teams with strong records but from smaller conferences or less traditionally strong football powers (Kansas, Utah, and Cincinnati, respectively) be given a shot against the more traditionally strong team. We believe that a legitimate case can be made for these matchups.

Choosing a Four-Team Tournament. We now turn attention to the problem of choosing four teams to participate in a final playoff. We define the expected controversy measure as before, this time computing the

probability of each of the 8 possible tournament outcomes and the controversy measure conditional upon each outcome. Results appear in Table 2.

In these experiments, we see that the tournament matchup with minimum expected controversy rarely involves the four teams that the model thinks are strongest according to the mean skill estimates. Only in 2001-02 and 2012-13 are the set of teams the same. With the additional budget of games, our method usually thinks it is best to include at least one team that we are less certain about, like Utah in 2004-05, Hawaii in 2007-08, and Boise State in 2006-07 and 2008-09.

How much better is a four-team playoff than a single championship game? By comparing the expected controversy numbers in Table 1 versus Table 2, we see that the four-team playoff decreases expected controversy by around 5-10%. For comparison, note the range of expected controversies across the most and least controversial is about 40%.

Which was the most controversial year? By looking at the values of the expected controversy scores, we can ask which seasons were more or less controversial. Our method estimates 2003-04 was the most controversial year, which is reasonable, given that it is the year in which the title was split between LSU and USC. The least controversial year (measured by how controversial we estimate the BCS choice to be) was USC vs. Texas in 2004-05.

6 Conclusions

We have presented principled methods for inferring, visualizing, and making decisions based upon uncertain estimates of NCAA football team ratings. The key aspect of our approach is inferring a posterior distribution representing structured uncertainty about team skills and their correlations. At a high level, we argue that when rating NCAA football teams, it is better to face the uncertainty head-on than to pretend it does not exist. When the time comes for the College Football Playoff committee to make decisions, we hope they will think about the uncertainty, and recognize that simply matching up the four teams that are strongest according to one consensus ranking is likely not the optimal decision procedure.

It would almost certainly be possible to improve the individual components of our methods. For example, there is no notion of time or special treatment of overtime wins and losses in the model. We believe this to explain, for example, why in the four-team experiment in 2007-08, the model is less forgiving of LSU (BCS #1) for its two regular season losses than the BCS system was. The loss function used for our decision procedure is a simple proxy for how controversial a decision is, and more refined notions of controversy, or other criteria for the goals of postseason matchups could be developed. Our claims are not that our specific instantiations are optimal. Rather, we are arguing for quantifying our uncertainty, and making decisions in a way that acknowledges the uncertainty.

Our model is simple and produces good results; when it comes to producing a single most likely rating, it mostly agrees with the computer systems that have been used in the BCS formula. It then has an advantage over alternative methods because it captures pairwise uncertainty in team skills, which we argue is essential when it comes time to make postseason matchup decisions—for example, to decide when teams with strong records from weaker conferences deserve a shot at the championship.

In future work, we would like to explore other applications of the methods developed here. The methodology is particularly applicable when there exist boundaries (whether they be temporal, geographic, or otherwise) that limit the flow of information about relative strengths of teams or players in different regions.

References

[1] Elo, A. E. (1978). *The rating of chess players: Past and present*. Arco Publishing, New York.

[2] Glickman, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics*, 48:377–394.

[3] Herbrich, R., Minka, T., and Graepel, T. (2007). TrueSkill™: A Bayesian skill rating system. In Scholkopf, B., Platt, J. C., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19 (NIPS-06)*, pages 569–576. MIT Press.

[4] Minka, T., Winn, J., Guiver, J., and Knowles, D. (2012). Infer.NET 2.5. Microsoft Research Cambridge. http://research.microsoft.com/infernet.

[5] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis, Second Edition*. Chapman and Hall/CRC.

Appendix

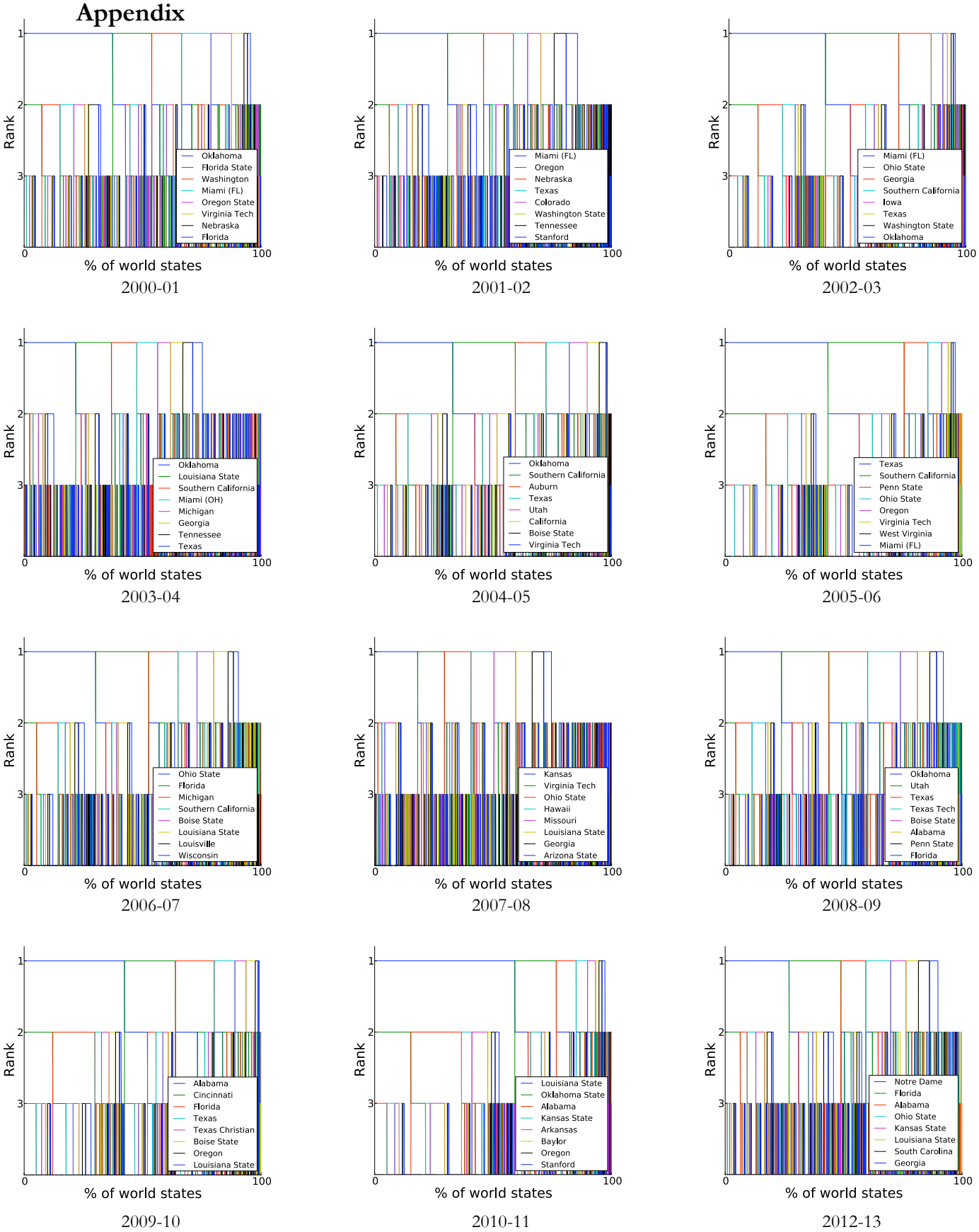


Figure 3: Visualization of rank distributions for years not shown in main paper. Best viewed in color.