Seven Degrees of Separation:
Assessing Regional Miscalibration of Elo Ratings in Chess

Michael Bodek

Quantitative Social Sciences Capstone Project

March 12th, 2019

Abstract

The Elo rating system is a rank system that uses pairwise results to measure the relative ability of thousands of competitors. Though originally designed for chess, it is now used in settings as diverse as sports, education, and tweet ranking. The premise of the Elo rating system is that it is self-correcting in that points flow from weaker competitors to stronger competitors after each comparison. However, I hypothesize that the Elo system cannot accurately measure relative ability in a sparse network where there are many degrees of separation between the players. I test this theory using data from national youth chess championships, where competitors have Elo ratings derived primarily from play in their home region and are playing opponents from different parts of the country for one of the first times. A regression testing whether result equals Elo win probability reveals that this is not the case for five of nine region dummies. The analysis is repeated for two separate age groups. Across the two age groups, all nine region dummies have the same sign, and there is heavy overlap among the significant regions. Pairwise comparison of regions confirms the previous results showing that the level of miscalibration is greater than what would occur purely due to noise. These results demonstrate a miscalibration of local Elo rating pools. Holding ability constant, a player's Elo rating would converge to a different value depending on the local cluster in which he competes.

Introduction

The Elo rating system is a rating system currently used in a wide variety of settings including chess, soccer, and scrabble. In recent years, it has been applied to areas as diverse as providing individualized education (Pelanek, 2016), ranking posts in online forums (Sarma et al., 2010), and even rating fabric quality (Tsang et al., 2016). The premise of the Elo rating system is that through examining pairwise results, it is possible to estimate the relative ability of thousands of competitors.

In this project, I am analyzing the accuracy of Elo ratings among youth chess players. More specifically, I am looking into whether Elo ratings can become miscalibrated between different regions when there is minimal interregional play. Answering this question yields information of both practical and academic value. From the practical side, the United States Chess Federation uses Elo ratings as the sole selection criterion for several awards and invitational events. A finding that players from some states or regions are underrated relative to their peers with the same ability would cause a rethinking of this selection process. From the

academic side, finding a miscalibration of Elo ratings between regions would show that Elo ratings cease to be an effective predictor of performance when there are local clusters and little intergroup play. This could inspire future studies on fragmented Elo networks.

Through my analysis, I empirically demonstrate the miscalibration of Elo ratings between regions using data from national youth chess championships. The participants in these tournaments each have an established Elo rating that is primarily derived from play in their home region. At national championships, they are all playing players outside of their local cluster for either the first, or one of the first times. Aggregating all games between players from two different regions, it is possible to measure both actual ability from tournament performance, and predicted ability from Elo ratings. I find a mismatch between the two which lends support to the idea that Elo ratings do not calibrate well between regions.

Literature Review

The Elo rating system was developed by physicist Arpad Elo (Elo, 1978) to evaluate the relative strength of chess players, and it has since been applied to a wide range of settings including soccer, education (Pelanek, 2016), and even cybersecurity (Pieters et al., 2012). Elo ratings use a player's wins and losses to infer innate ability. Through analysis of pairwise results it is possible to differentiate thousands of competitors in a large system as follows:

Given player A with rating R_a, and player B with rating R_b, player A's expected result in a head-to-head matchup is given by the logistic curve $1 / (1 + 10 ^ ((R_b – R_a) / 400))$. The result of a game provides new information as to a competitor's true ability, so ratings update either up or down after each game according to the formula $R_{new} = R_{old} + K * (\text{Result} - \text{Result}_{expected})$, where K is an arbitrary weight. This updating procedure brings ratings closer to true ability; a player playing better than expected will have her rating rise and vice-versa (Glickman 1995).

For example, assume player A has a rating of 1900 and player B has a rating of 1700. By the first formula above, player A has a win probability of .76 and player B a win probability of .24. Should player B win a game against player A, his rating will increase to 1700 + .76 * K, while player A's rating will decrease to 1900 - .76 * K. Player B's victory indicates that he was underrated relative to player A, so the system corrects itself. After many games, the Elo rating system can approximate ability quite accurately.

In fact, a theoretical examination mathematically proved that in a dense interconnected network, Elo ratings converge to true ability fairly quickly (Jabin and Junca, 2015). Even working under the assumption that a player's skill increases with each game, this relationship still holds (During et al., 2018). In other words, the Elo rating system can accurately measure time-invariant playing ability. However, Jabin and Junca note that the convergence between ratings and ability could break down when competitors are only able to compete against a local subset of the population (page 421).

One of the early studies on the Elo rating system suggests that ratings are more accurate when there are fewer degrees of separation linking all the competitors. In this situation ratings can update more quickly throughout the network. On the contrary, if there are several isolated groups of players with minimal inter-group play, ratings might not converge in the same way in each local cluster. That is to say, players of the same ability might end up with different ratings depending on with whom they primarily compete. In practice, such a fragmented Elo network can occur if players are limited to competing in a given geographic area (Glickman and Jones 1999).

An empirical analysis of the network of tournament chess players revealed that this network is in fact fragmented, and it has a statistically significant clustering coefficient

(Almeira et al., 2017). That means the level of clustering in the network is much higher than what would occur purely due to random chance. The ideal behavior underlying Jabin and Junca's proof does not line up with the actual network structure of a real-world Elo system. This opens the door to the possibility of Elo rating mismatches between local clusters of chess players.

Several authors have moved beyond Elo ratings and attempted to evaluate chess players based on move quality. They went through each move of top-level games and compared the selected player's moves to moves suggested by a supercomputer, in order to create match rates. The match rates were compared to Elo ratings, and they revealed among other things that Elo ratings do not inflate over time, (Dangauthier et al., 2007; Ferreira, 2012; Regan and Haworth, 2011) and interestingly, that professional players from Canada play better than their international Elo ratings predict (Regan et al., 2012).

Regan et al. suggest that top Canadian players are underrated due to fewer opportunities to play in tournaments. However, this could also be a result of the difficulty Elo rating systems have in assigning initial ratings to players who have yet to play a tournament game. New players entering the rating pool in a given cluster can have a distortionary effect on the ratings of their opponents. In practice, this has pulled ratings down (Fenner et al., 2012). However, regardless as to the cause, the underrating of Canadian players lends support to the idea that ratings can become miscalibrated.

In this paper, I build upon the idea of rating miscalibraion between local clusters and focus on players of a wider skillset. Using data from US national youth chess championships, I show that the United States Chess Federation Elo rating system is miscalibrated between regions.

This contributes to the literature through more generally demonstrating rating imbalances in a clustered Elo rating system.

Data and Methods

Tournament data from national youth chess championships between 2004 and 2018 were collected from the United States Chess Federation results page. The results of each game along with player information from the National Elementary School Chess Championship were compiled into one dataset, and the National Junior High Chess Championship results were compiled into a separate dataset. Thus, there are two parallel datasets for the two age groups.

In addition to result, for each game in the datasets there is information on both players' ratings, grade, state, and school. Results were coded as 1 for a win, 0.5 for a draw, and 0 for a loss. The win probability for each player-game dyad was calculated according to the Elo rating formula $1 / (1 + 10 \text{ ^ } ((\text{Rating}_b - \text{Rating}_a) / 400))$. Each of the 50 states was mapped to one of nine non-overlapping regions. A detailed description of the mapping from state to region can be found in Appendix A. States with more than 5 competitors playing in the tournament-year were coded as being "populous" chess playing states.

For both age groups, beginner players defined as having an Elo rating below 800 were removed from the dataset as beginners have a higher variance to their ratings (the average Elo is 1400). Then expert players defined as having an Elo rating above 2000 were removed from the dataset, as they are more interconnected in the Elo network. Since highly-ranked players are likely to have more experience playing outside of their home states, the expert players were removed to greater highlight the effect of isolated rating pools. Finally, elementary school-aged players in the Junior High Championship were removed as they are unobservably different from

the other competitors in the tournament. After the removals, the Elementary School dataset contains 12766 rows which correspond to 6383 unique games, and the Junior High dataset contains 7048 rows which correspond to 3524 unique games.

I am interested in studying whether chess players from some regions are systemically overrated or underrated relative to their peers in other parts of the country with the same ability. In other words, have Elo ratings in chess become miscalibrated between regions. To model this, region dummies are the main independent variables, and factors such as grade and income represent controls. The dependent variable is by how much a player overperforms or underperforms his or her Elo rating, calculated as (score - win probability). This quantity has a mean of zero and it ranges from -1 to +1. The further the measure of Elo outperformance is from zero, the more actual result differs from expected result. The null hypothesis is that the region dummies have no effect on Elo outperformance, while the alternative is that they have some effect.

The model is a linear regression with (score - win probability) as the dependent variable, and region dummies and controls the independent variables. Two identical models are constructed, one for the Elementary School dataset, and one for the Junior High dataset. These models can test whether being from a certain region is associated with a difference in performance relative to Elo rating.

Separately, a second model tests direct miscalibration between two regions. For all of the 36 (more specifically 9 * 8 / 2) possible combinations of two regions, if there are more than 50 games between the regions, a t-test is performed to see whether it is possible to reject the null that actual score equals expected Elo win probability. A rejection of the null would indicate that

players from one region outperform expectations when playing players from another. Again, this procedure is applied twice, once for each age group.

Results

The main dependent variable (score - win probability), has a mean of zero. In other words, on average score and win probability are the same, as is expected for the Elo rating system. In order to show miscalibration between regional rating pools, it is enough to show that accounting for the effects of one region moves this quantity away from zero, or that score is not the same as win probability. The Elo rating system is designed to be a catch-all statistical method of ranking players. Thus, any demographic factors causing players to systematically over or underperform their Elo ratings would reveal an imperfection in the system.

To test whether rating pools can become miscalibrated between regions I first performed a short regression of Elo overperformance on region. Results from the regression can be found in Table 1 in Appendix B. From the regression, six of the nine region dummies are significant at a .05 level in the Elementary School dataset, as is the same in the Junior High dataset. Five of the region dummies are significant in both datasets and every single dummy variable has the same sign in both the Elementary School and the Junior High data. The similarity in the results for both of the age groups suggests that the significance represents miscalibrated regional rating pools as opposed to chance correlation. That the same regional disparity exists both for elementary school and junior high players indicates that the Elo rating system is not in sync between different regions.

For example, the coefficient on the West South Central dummy variable is -.109 in the Elementary School data and -.105 in the Junior High data. This means that a player from that

region on average has an actual win probability about ten percentage points lower than what Elo ratings predict. Looking across all the region coefficients from the two datasets, they are all within two percentage points of each other except for Pacific, where the Elementary School coefficient is much greater.

However, there is a simple explanation for the large difference between the two Pacific coefficients. As mentioned in the New York Times (McClain, 2010) many tournaments in Washington state are not rated using the national Elo system. Thus, it makes sense that younger players from Washington state have inaccurate ratings. In fact, removing Washington from the Elementary School dataset brings its Pacific coefficient to zero and does not significantly impact the other coefficients. Table 2 in Appendix B compares the results of the regression with and without observations from Washington. As in the previous results, these coefficients are similar to those in the Junior High dataset.

The similarity of the coefficients between the two datasets indicates with a high degree of confidence that players from a significant region are underrated or overrated relative to players from another region. In other words, if a player moved from one region to another, their ratings would converge to a new value without any changes in ability.

After testing this hypothesis in the short regression, I added controls to further establish the validity of the results. Each of the controls were first regressed on their own against Elo outperformance. Whether a player comes from a populous chess playing state was the only control that was related to the outcome. All the other variables such as income, grade, and rural ended up just being noise.

Additionally, for large rating differences, the win probability asymptotically approaches either 1 or 0. Thus, it becomes harder to interpret these probabilities at the extremes and they are thought to be less accurate. To check this, I regressed Elo overperformance against $(Rating_a - Rating_b)$ and found significance. I also regressed $(Rating_a - Rating_b)$ on region and found significance again. Since this quantity is related to both the independent and dependent variables it must be included to avoid omitted variable bias.

The final model has Elo overperformance as the dependent variable, and region as the main input variable. Whether a player comes from a state with many chess players and rating difference between the two players are controls. Results from the model can be found in the coefficient plots in Figures 1 and 2 as well Table 3 in Appendix B. Note the similarities between the two coefficient plots.

Fig 1. Region dummy coefficient estimates and 95% confidence intervals for Elementary School

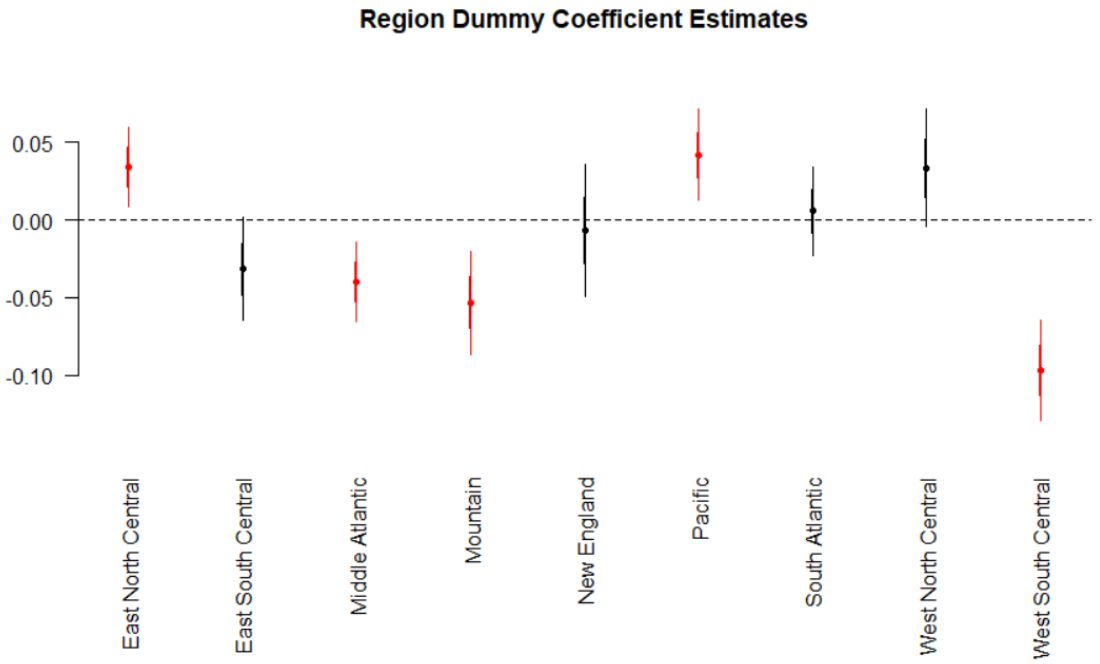
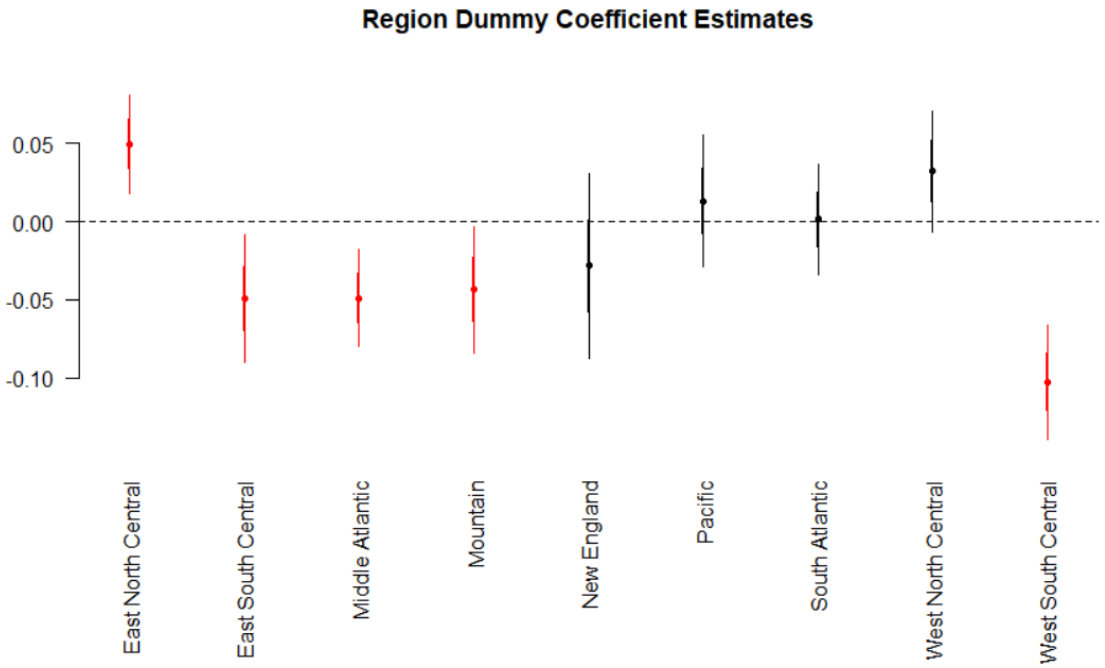


Fig 2. Region dummy coefficient estimates and 95% confidence intervals for Junior High



As in the short regression, after adding the controls all nine region dummies have the same sign in both datasets. This verifies that the coefficients represent a true relationship and are not simply noise. The probability of all nine signs being identical due to chance is 1 / (2^9).

In the Elementary School dataset, five of the region dummies are significant, and five are as well in the Junior High dataset. Four of the five significant dummy variables are the same in both datasets. The Pacific dummy is significant in the Elementary School data but not the Junior High data for reasons mentioned above. The East South Central dummy is significant in the Junior High data but not the Elementary School data, however in the latter it has a p-value of .056. Furthermore, all the variables have relatively similar coefficients. This indicates a disparity across regional rating pools. The coefficients reveal that players from some regions are systematically underrated or overrated relative to players from another.

Ceteris paribus players with the same ability from two different regions would have two different Elo ratings. For most regions the difference between expected and actual performance is on the order of magnitude of 2-3 percentage points, however, for the West South Central region, the difference is as large as 10 percentage points. For a visual description of the miscalibration see Figures 3 and 4 for the significant regions in the Elementary School and Junior High data, and 5 and 6 for all the regions. Again, note the similarities between the Elementary School and Junior High figures.

Fig. 3. Elo disparity for significant regions from the Elementary School dataset

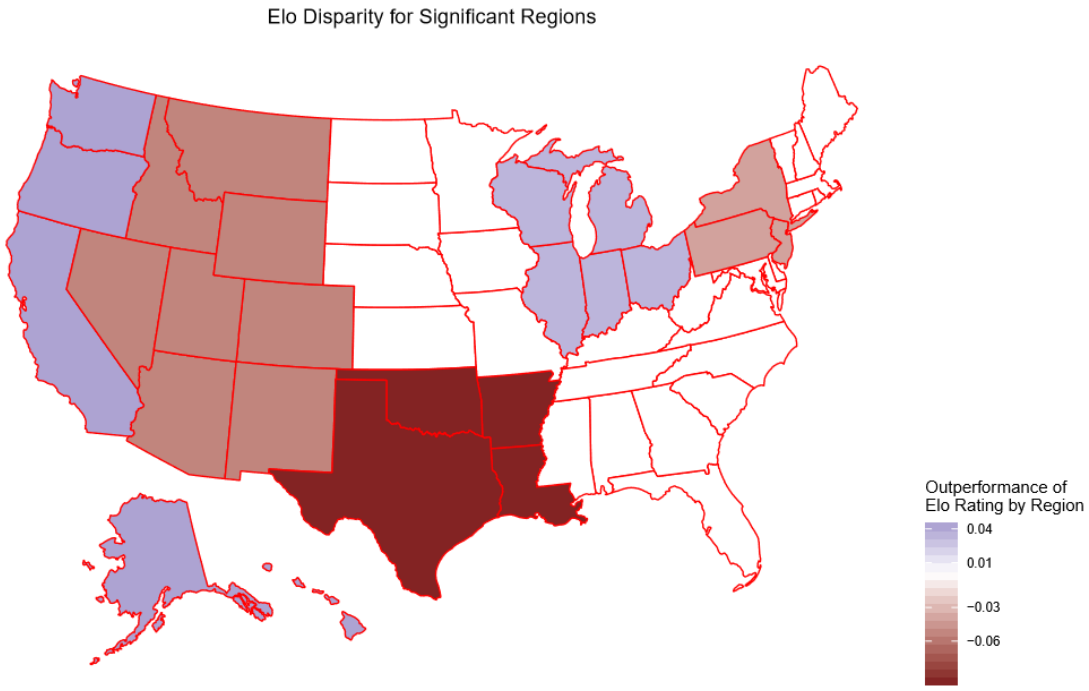


Fig. 4 Elo disparity for significant regions from the Junior High dataset

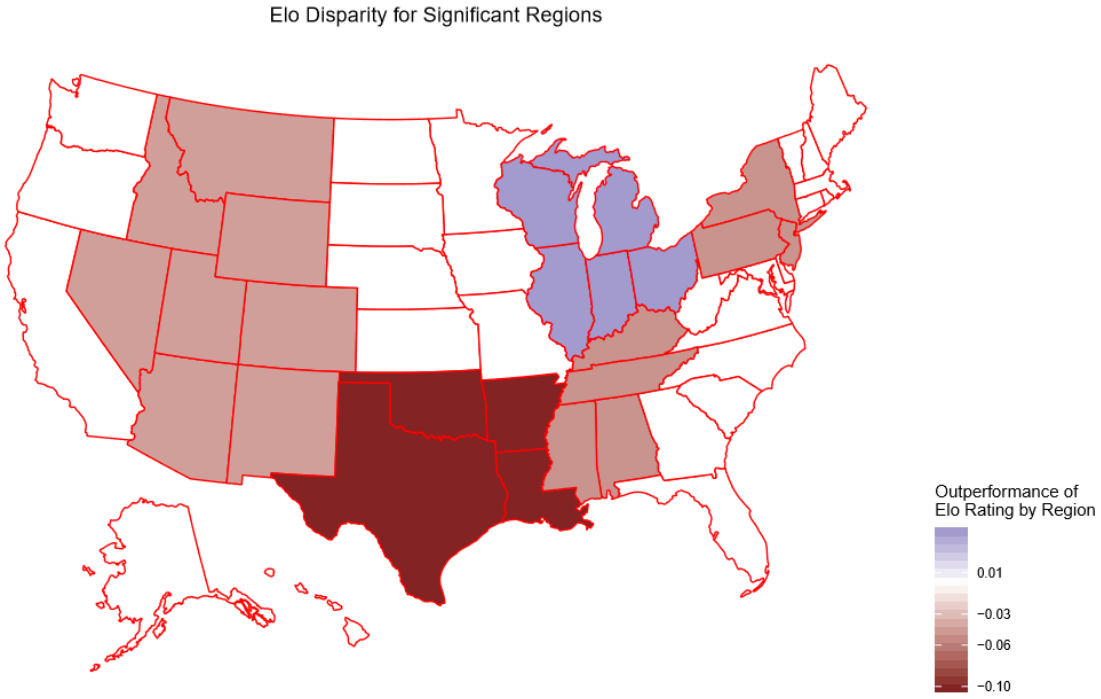


Fig. 5. Elo disparity for all regions from the Elementary School dataset

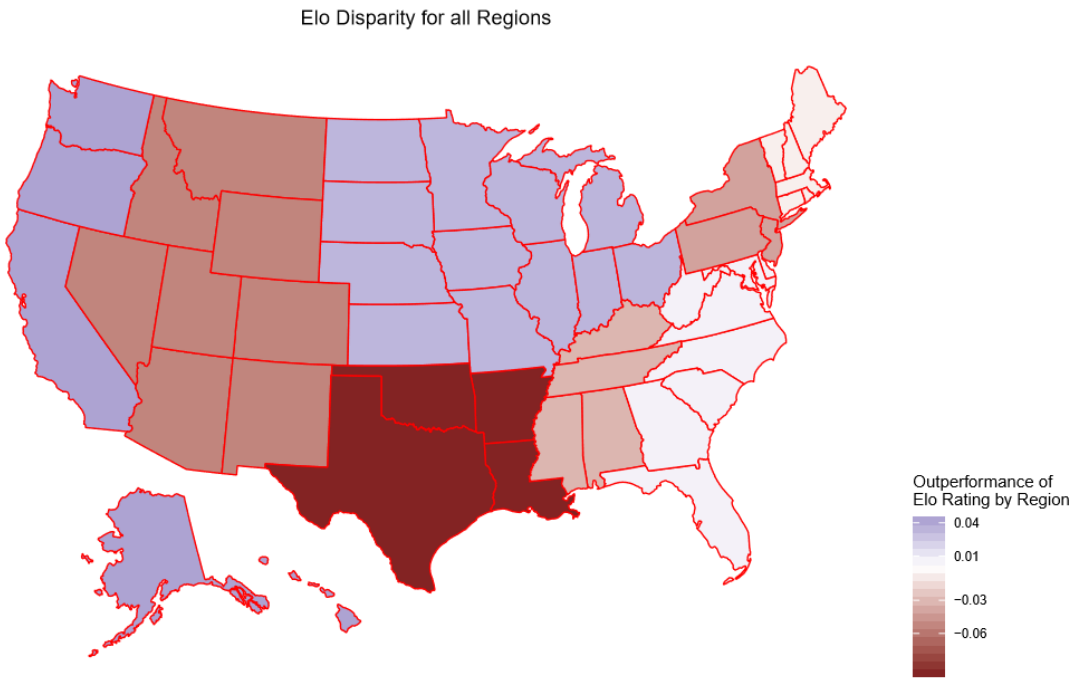
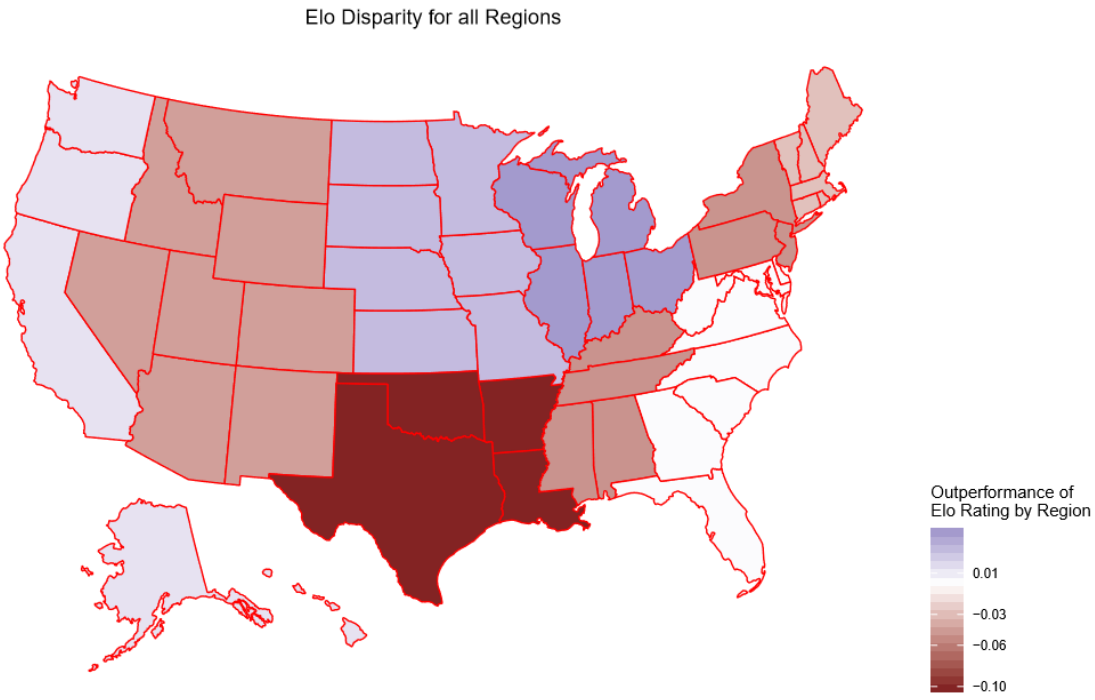


Fig. 6. Elo disparity for all regions from the Junior High dataset



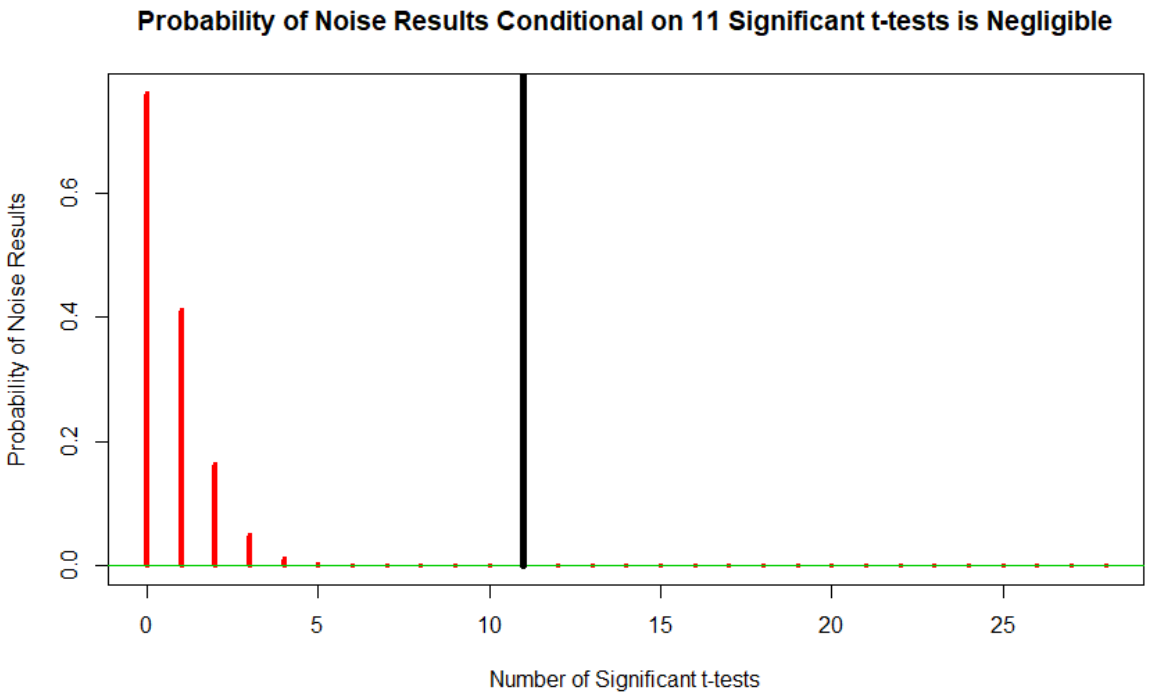
The regressions above show miscalibration between regional rating pools and a national average. They produce results indicating that players from some regions outperform their Elo ratings and that others underperform them. Most importantly, they demonstrate similar effects across both of the age groups, indicating that the results are not spurious correlation.

The regression model is now supplemented by a series of pairwise analyses that show whether regions are miscalibrated directly with each other, as opposed to simply a national average. For each of the 36 possible combinations of two regions, the data is subset to only include games between players from the two selected regions. For all the subsets with more than 50 games, a t-test is performed to test the null that score equals win probability.

In the Elementary School dataset, there are 28 region pairs with more than 50 games between them. Thus 28 t-tests are performed. The null hypothesis is rejected at the .05 level 11 out of 28 times. Details of each of the 28 t-tests can be found in Table 4 in Appendix C.

The frequent rejection of the null indicates with a high degree of confidence that there is miscalibration in rating pools between regions. Looking at the binomial distribution function, the probability of all false positives given 28 trials, 11 or more success, and p = .05 is as small as 3.5 * 10^-9 (Figure 7). Thus rejecting 11 out of 28 null hypotheses shows that Elo ratings are not always an accurate measure of comparing players across disparate geographic areas. This lends support to the broad idea that Elo ratings will not perform as accurately when there are many degrees of separation between the players.

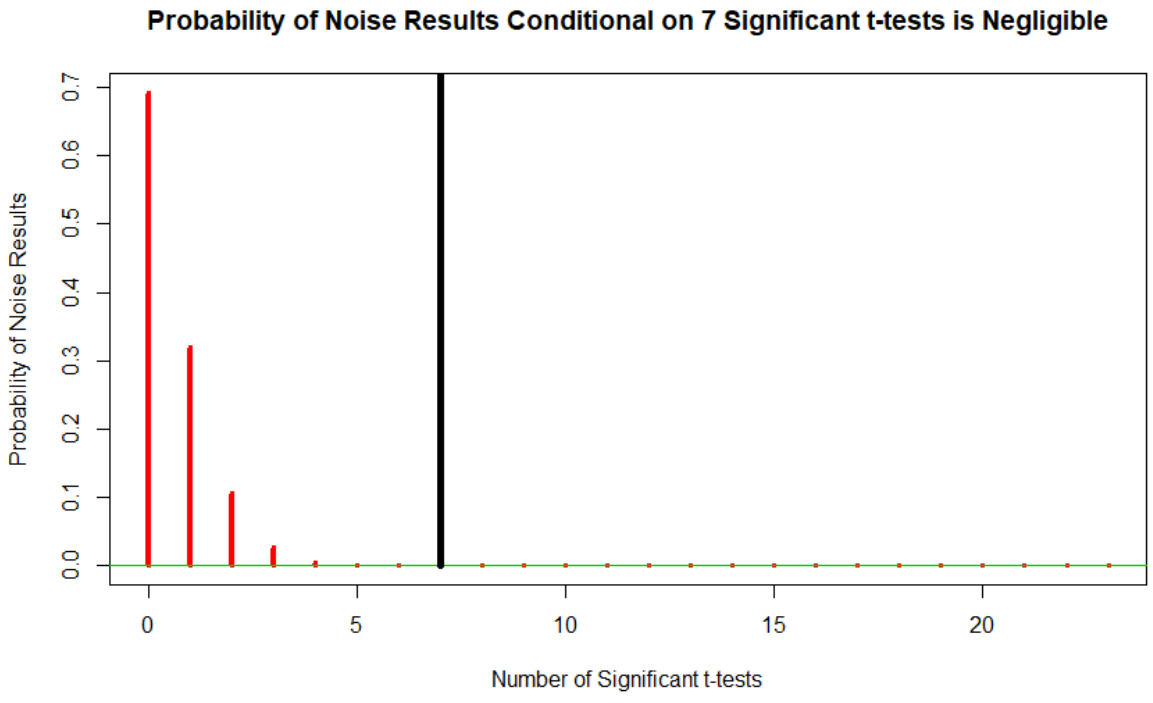
Fig. 7. Pairwise t-test results from Elementary School data shown against a binomial distribution



To color the discussion with an example, there are 113 games between players from the West South Central region and the South Atlantic region. Players from the West South Central have a mean score of .39, but their Elo ratings gave them a prior win probability of .53. The large difference between the two indicates that the South Atlantic rating pool is underrated relative to the West South Central rating pool. This result makes intuitive sense: in the regressions the coefficient on West South Central was -0.1, indicating that players from that region scored on average ten percentage points below expectations.

A similar analysis was conducted for the Junior High dataset. There are 23 region pairs with more than 50 games between them, and 7 of the 23 t-tests are significant. Details of each individual t-test can be found in Table 5 in Appendix C. Again, using the binomial distribution, the probability of seeing this many significant results due to noise is 9.7 * 10^-6 (Figure 8). Once more, the t-tests show with a high degree of certainly that the regions are miscalibrated.

Fig. 8. Pairwise t-test results from Junior High data shown against a binomial distribution



Overall the regressions and the t-tests provide strong statistical evidence for the miscalibration of regional rating pools. Five of nine region dummy variables are statistically significant in each of the datasets, four of the five are significant in both, and all nine region dummies have the same sign. The similarity of the two sets of results indicates that performance does not line up with Elo predicted ability. The consistent underrating and overrating of regional rating pools means that a player who moves across the country would have his or her rating change without an underlying change in true ability.

Furthermore, the t-tests show that many pairs of regions are miscalibrated, when in a perfectly functioning Elo rating system only approximately 1/20 of the tests would be significant due to random noise. In a network where players are not very interconnected, the accuracy of an Elo rating system begins to decrease. The analyses above empirically show that Elo ratings earned primarily from local competition are not comparable at the national level.

Discussion and Conclusion

This project examines whether an Elo rating system can become miscalibrarted when there are many local groups and minimal intergroup play. Data is used from national youth chess championships whose participants all have Elo ratings primarily derived from play in their home region. At national championships they are playing players from different parts of the country for one of the first times.

Analyzing the data, players from five of the nine regions significantly overperform or underperform their Elo ratings. Pairwise comparisons of the games between players from two regions reveal that actual score differs from expected score by an amount much greater than what would be expected due to random noise. Furthermore, similar results are found both for the

elementary school age group and the junior high age group. This lends strong support to the hypothesis that regional rating pools are miscalibrated. The near identical results between the two datasets are unlikely to occur purely due to chance.

These results indicate that Elo ratings do not perform as well when there are local clusters and little play between clusters. Thus, they are not as accurate in a sparse network. For Elo ratings to accurately update, all the players need to be connected to the rest of the network with few degrees of separation.

This project expands the literature about how network structure affects Elo rating systems, however it has its limitations. It assumes that youth chess players from different parts of the country are far apart in the network, though this might not be the case. Within a region there can be various degrees of connectivity which the analysis from this project fails to discern. It would be interesting for a future study to map the network of chess players and examine how Elo rating accuracy is a function of connectivity, rather than using regions as a proxy. In that study, local rating pools can be constructed through a network as opposed to a geographic analysis.

Furthermore, it also could be interesting to see how Elo rating outperformance is a function of prior games played. This project eliminates beginner and expert level players from the data and assumes that all the intermediate players played a similar number of games. However, future projects might be able to include the number of prior games for each player, and closely examine its relationship with Elo rating overperformance.

Nonetheless, the results produced in this paper are of immense value. They show that players with the same ability would have different Elo ratings depending on in which region they live. This means that Elo ratings are not a stand-alone measure of which players are best. More

broadly it demonstrates that a single Elo rating system should not be used to compare different leagues or groups that do not often compete. I would be interested in seeing how this result compares to an analysis of other use cases of the Elo rating system. FIFA for example compares many soccer teams using a single Elo rating system, and understanding the accuracy of teams' Elo ratings could have broad implications on everything from seeding teams to betting markets.

References

Almeira, Nahuel, Ana L. Schaigorodsky, Juan I. Perotti, and Orlando V. Billoni. "Structure Constrained by Metadata in Networks of Chess Players." Scientific Reports7, no. 1 (2017).

Dangauthier, Pierre, Ralf Herbrich, Tom Minka, and Thore Graepel. "TrueSkill Through Time: Revisiting the History of Chess." Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007.

Düring, Bertram, Marco Torregrossa, and Marie-Therese Wolfram. "Boltzmann and Fokker–Planck Equations Modelling the Elo Rating System with Learning Effects." Journal of Nonlinear Science, 2018.

Elo, Arpad E. The Rating of Chessplayers, past and Present. London: Batsford, 1978.

Fenner, T., M. Levene, and G. Loizou. "A Discrete Evolutionary Model for Chess Players Ratings." IEEE Transactions on Computational Intelligence and AI in Games4, no. 2 (2012): 84-93.

Ferreira, Diogo R. "Determining the Strength of Chess Players Based on Actual Play." ICGA Journal35, no. 1 (2012): 3-19.

Glickman, Mark E. "A Comprehensive Guide to Chess Ratings"." American Chess Journal, 1995. http://www.glicko.net/research/crs.pdf.

Glickman, Mark E., and Jones, Albyn C., "Rating the chess rating system", Chance12, no. 2, (1999): 21-28.

Jabin, Pierre-Emmanuel, and Stéphane Junca. "A Continuous Model For Ratings." SIAM Journal on Applied Mathematics75, no. 2 (2015): 420-42.

McClain, Dylan L. “In Children’s Chess, Debate on Ratings.” New York Times, 15 June 2010, www.nytimes.com/2010/06/16/us/16chess.html.

Pelánek, Radek. "Applications of the Elo Rating System in Adaptive Educational Systems." Computers & Education98 (2016): 169-79.

Pieters, Wolter, Sanne H.g. Van Der Ven, and Christian W. Probst. "A Move in the Security Measurement Stalemate." Proceedings of the 2012 Workshop on New Security Paradigms - NSPW 12, 2012.

Regan, Kenneth W., and Guy McC. Haworth. "Intrinsic Chess Ratings." Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, San Francisco, 2011.

Regan, Kenneth W., Bartlomiej Macieja, and Guy McC. Haworth. "Understanding Distributions of Chess Performances." Lecture Notes in Computer Science Advances in Computer Games, 2012, 230-43.

Sarma, Anish Das, Atish Das Sarma, Sreenivas Gollapudi, and Rina Panigrahy. "Ranking Mechanisms in Twitter-like Forums." Proceedings of the Third ACM International Conference on Web Search and Data Mining - WSDM 10, 2010.

Tsang, Colin S.c., Henry Y.t. Ngan, and Grantham K.h. Pang. "Fabric Inspection Based on the Elo Rating Method." Pattern Recognition51 (2016): 378-94.

Appendix A: States in Each Region

Region	States
East North Central	IL IN MI OH WI
East South Central	AL KY MS TN
Middle Atlantic	NJ NY PA
Mountain	AZ CO ID MT NM NV UT WY
New England	CT MA ME NH RI VT
Pacific	AK CA HI OR WA
South Atlantic	DE FL GA MD NC SC VA WV
West North Central	IA KS MN MO ND NE SD
West South Central	AR LA OK TX

Appendix B: Regression Tables

Table 1: Short Regression

	Elo Outperformance:	
	Elementary School	Junior High
	(1)	(2)
East North Central	0.024**	0.030**
	(0.011)	(0.013)
East South Central	−0.028*	−0.042**
	(0.017)	(0.021)
Middle Atlantic	−0.050***	−0.064***
	(0.013)	(0.016)
Mountain	−0.061***	−0.042**
	(0.017)	(0.020)
New England	−0.016	−0.007
	(0.021)	(0.030)
Pacific	0.037**	0.002
	(0.015)	(0.021)
South Atlantic	0.005	0.007
	(0.014)	(0.018)
West North Central	0.042**	0.048**
	(0.019)	(0.020)
West South Central	−0.109***	−0.105***
	(0.016)	(0.019)
Observations	12,766	7,048
R^2	0.012	0.013
Adjusted R^2	0.011	0.012
Residual Std. Error	0.388 (df = 12757)	0.383 (df = 7039)
F Statistic	18.929*** (df = 8; 12757)	11.737*** (df = 8; 7039)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 2: The Effect of Removing Washington (Elementary School)

	Elo Ourperformance:	
	All States	Washington Excluded
	(1)	(2)
East North Central	0.024**	0.029**
	(0.011)	(0.011)
East South Central	−0.028*	−0.023
	(0.017)	(0.017)
Middle Atlantic	−0.050***	−0.047***
	(0.013)	(0.013)
Mountain	−0.061***	−0.063***
	(0.017)	(0.017)
New England	−0.016	−0.018
	(0.021)	(0.022)
Pacific	0.037**	−0.004
	(0.015)	(0.017)
South Atlantic	0.005	0.004
	(0.014)	(0.014)
West North Central	0.042**	0.043**
	(0.019)	(0.019)
West South Central	−0.109***	−0.097***
	(0.016)	(0.017)
Observations	12,766	11,366
R^2	0.012	0.008
Adjusted R^2	0.011	0.007
Residual Std. Error	0.388 (df = 12757)	0.386 (df = 11357)
F Statistic	18.929*** (df = 8; 12757)	11.346*** (df = 8; 11357)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 3: Full Regression Model

	Elo Outperformace:	
	Elementary School	Junior High
	(1)	(2)
East North Central	0.034***	0.049***
	(0.013)	(0.016)
East South Central	−0.032*	−0.049**
	(0.016)	(0.020)
Middle Atlantic	−0.040***	−0.049***
	(0.013)	(0.016)
Mountain	−0.053***	−0.043**
	(0.017)	(0.020)
New England	−0.007	−0.028
	(0.021)	(0.029)
Pacific	0.042***	0.013
	(0.015)	(0.021)
South Atlantic	0.006	0.001
	(0.014)	(0.018)
West North Central	0.033*	0.032*
	(0.019)	(0.019)
West South Central	−0.096***	−0.102***
	(0.016)	(0.018)
Rating Difference	−0.0002***	−0.0002***
	(0.00001)	(0.00001)
Populous	−0.018*	−0.026**
	(0.010)	(0.012)
Observations	12,766	7,048
R^2	0.044	0.047
Adjusted R^2	0.043	0.046
Residual Std. Error	0.382 (df = 12755)	0.377 (df = 7037)
F Statistic	58.609*** (df = 10; 12755)	34.818*** (df = 10; 7037)
Note:		*p<0.1; **p<0.05; ***p<0.01

Appendix C: Results from Pairwise t-tests

Table 4. 11 out of 28 pairwise t-tests significant from Elementary School data

Region 1	Region 2	N	Avg Score	Avg Win Probability	p−Value	Significant?
Pacific	New England	52	0.49	0.48	0.8319	0
Pacific	East North Central	142	0.47	0.45	0.5292	0
Pacific	Middle Atlantic	548	0.57	0.48	0	1
Pacific	Mountain	137	0.56	0.48	0.0228	1
Pacific	East South Central	136	0.63	0.58	0.1332	0
West South Central	Pacific	138	0.38	0.57	0	1
West South Central	East North Central	88	0.52	0.6	0.0289	1
West South Central	South Atlantic	113	0.39	0.53	0.0013	1
West South Central	Middle Atlantic	374	0.47	0.53	0.0011	1
West South Central	Mountain	73	0.51	0.54	0.449	0
West South Central	East South Central	72	0.45	0.57	0.0175	1
New England	Middle Atlantic	142	0.57	0.58	0.9474	0
South Atlantic	Pacific	234	0.47	0.5	0.3309	0
South Atlantic	New England	60	0.51	0.49	0.6688	0
South Atlantic	East North Central	186	0.51	0.51	0.9374	0
South Atlantic	Middle Atlantic	645	0.51	0.45	0.0001	1
South Atlantic	Mountain	139	0.48	0.45	0.272	0
South Atlantic	East South Central	150	0.53	0.49	0.168	0
Middle Atlantic	East North Central	443	0.46	0.53	0.0004	1
Middle Atlantic	East South Central	336	0.57	0.58	0.4952	0
West North Central	Pacific	80	0.47	0.43	0.4319	0
West North Central	East North Central	65	0.49	0.4	0.0954	0
West North Central	South Atlantic	96	0.56	0.52	0.3275	0
West North Central	Middle Atlantic	181	0.54	0.42	0	1
West North Central	East South Central	58	0.56	0.52	0.4036	0
Mountain	East North Central	100	0.52	0.54	0.6992	0
Mountain	Middle Atlantic	296	0.49	0.54	0.0105	1
East South Central	East North Central	81	0.44	0.47	0.5284	0

Table 5. 7 out of 23 pairwise t-tests significant from Junior High data

Region 1	Region 2	N	Avg Score	Avg Win Probability	p−Value	Significant?
Middle Atlantic	East North Central	274	0.45	0.53	0.0003	1
Middle Atlantic	East South Central	200	0.58	0.62	0.1555	0
South Atlantic	East North Central	104	0.51	0.47	0.301	0
South Atlantic	Middle Atlantic	272	0.54	0.46	0.0022	1
South Atlantic	Mountain	71	0.51	0.51	0.9228	0
South Atlantic	Pacific	62	0.48	0.46	0.6961	0
South Atlantic	East South Central	78	0.53	0.48	0.2055	0
Mountain	East North Central	56	0.48	0.54	0.3096	0
Mountain	Middle Atlantic	163	0.44	0.46	0.3532	0
West South Central	East North Central	67	0.44	0.5	0.1284	0
West South Central	Middle Atlantic	279	0.4	0.43	0.1373	0
West South Central	South Atlantic	115	0.42	0.53	0.0024	1
West South Central	East South Central	65	0.38	0.41	0.4039	0
West South Central	West North Central	88	0.44	0.67	0	1
Pacific	East North Central	53	0.57	0.52	0.372	0
Pacific	Middle Atlantic	118	0.59	0.52	0.0302	1
Pacific	Mountain	67	0.57	0.55	0.7856	0
East South Central	East North Central	67	0.46	0.53	0.0803	0
West North Central	East North Central	84	0.46	0.43	0.455	0
West North Central	Middle Atlantic	161	0.45	0.37	0.0117	1
West North Central	South Atlantic	87	0.55	0.47	0.023	1
West North Central	Mountain	74	0.43	0.41	0.4942	0
West North Central	Pacific	72	0.41	0.35	0.1667	0