We're Just Ordinary People: A Look at Kim Davis and Media Bias

Kevin Durham, Cory Nichols, Adam Soto, Kareem Williams
SMU Data Science MSDS 6110 - Immersion

1. INTRODUCTION

Marriage equality has been a hot-topic in the United States for decades. Over the summer of 2015, the Supreme Court ruled in favor of same-sex marriage, once and for all ending the debate about the legitimacy of heterosexual and homosexual marriage rights. Unfortunately, all of the United States did not take this news kindly and many refused to accept the justice's ruling.

One such dissenter was Kim Davis, a Rowan county clerk in Kentucky. Kim was in charge of issuing and overseeing marriage licenses for all of Rowan County. She refused to allow her office to issue marriage licenses to same-sex couples and was soon after sued by four gay couples. Kim was then ordered by the District Court of Eastern Kentucky to issue licenses as directed by the law. She then continued to refuse and was consequently jailed for contempt.

Kim Davis' story was met with an avalanche of support and comparable outrage. The media was similarly divided and the coverage of the Kim Davis marriage equality situation proved out the split.

Given such a contentious subject, it is anecdotally surmised that there will be obvious coverage bias from news outlets around the United States. In California, one would likely expect coverage of the Kim Davis situation to be more biased from a news outlet in Georgia. A Georgian would likely expect biased coverage from a news outlet in San Francisco.

To determine if there is indeed geographic bias, 40 randomly selected articles were analyzed prior to Kim Davis' incarceration on September 3rd, 2015 using logistic and multiple linear regression across different sampling and scoring methodologies. We are primarily focused on the former, logistic regression using binary responses for article bias. Multiple linear regression findings are presented as a supplement to the primary analysis. Further, we are interested in covariates such as outlet reach, gender of the article author and the word count of the article. These covariates are not critical and will not guide our analysis, rather, we wish to control for their effects.

2. OBSERVATIONAL STUDY DESIGN

A total of 60 online articles with Kim Davis as the subject were randomly selected from across United States media outlets using Google Filtered Search. To truly test whether a news outlet was biased, only opinion pieces from non-Associated Press writers were selected before Kim Davis' incarceration on September 3rd, 2015, and no earlier than July 1st, 2015. 60 random numbers between 1 and 200 were used to identify the articles in the search result to be included in the initial sample level.

Of the 60 articles, 34 articles were then identified from southern news outlets. 26 non-southern news articles representing the "other" level of the regional factor were made up of California, Washington, Washington DC, Nevada and Oregon.

Finally, a total of 40 articles were selected to represent the final sample for analysis. To control for selection bias, article sampling was assigned to four different individuals researchers. Each individual selected 10 articles (5 for each region) across Southern and Non-Southern regions. Figure 1 shows a visual breakdown of the sampling plan:

Figure 1

Kim Davis Media Bias Sampling Plan

1) Select 60 Articles (Outlet Sponsored, Non-AP Opinion Pieces) Via Search and Random Number Assignment

N = 60

2) Categorize Into South and Other

Total South: 26	Total Other: 34
States: AL, KY, TX, LA	States: NV, CA, WA, OR

2) Randomly Select: n = 20 from each Regional Strata

South		Other	
Regional	4	Regional	9
Local	16	Local	11
Total	**20**	**Total**	**20**

Article selection proved to be difficult, as many news outlets, especially local outlets, do not maintain article presence online after a month. Therefore, the 60 articles initially chosen proved to be the majority of what could be expected to be found via web search for opinion based, non-associated press (outlet owned) articles on September 20th 2015. Further, even though Google Search has a search filter, selection of the final sample required culling a few articles that were not deemed opinion based by the selector. This subjectivity does lead to a less powerful inference in the study, however, association and odds can still be investigated and inferred with caution.

Additional covariates were not balanced in this study, as they are not our primary focus. Further exposure to variables and categorizations is found in section three.

3. EXPLORATORY DATA ANALYSIS UNDER THE LOGISTIC MODEL

Logistic regression using binary indicators is the primary means of data analysis for this study. Bias is represented as a binary (0,1) value with four explanatory variables: word count, location, reach and author gender. Full exposition to response and explanatory variables follows below.

3.1 THE RESPONSE VARIABLE

Variable 1	Levels	Description
BIAS	Binary (2)	Article Bias

Article bias (1) was decided based on a five point rubric:

1. **Does the headline predisposition the reader?**
2. **Is the article unbalanced from an evidence standpoint?**
3. **Does the article directly slander or compliment the subject or important interests associated with subject matter? (e.g. one example was LGBT activists were called 'militants')**
4. **Are the visuals overly positive or negative?**
5. **Overuse of buzzwords and categorizations? E.g. LGBT, Christian Agenda, Homosexual Agenda.**

Should three out of five points of the rubric be answered with a "yes", the article was considered biased and recorded as an event (1). Each article received a grade from four researchers to control for bias and rubric interpretation. Group scores were collated, with majority rules deciding on the result. There were no ties.

3.2 EXPLANATORY VARIABLES

Variable 1	Levels	Description
LOCATION	2	Region – South or "Other"

LOCATION has two levels, and is represented by "South" and "Other." South is made up of Texas, Louisiana, Kentucky and Alabama. Other is comprised of Washington, Washington DC, Oregon, California and Nevada. Because LOCATION is the explanatory variable we're primarily interested in, the levels are balanced at 20 observations each.

Level Labels:
1: South **2:** Other

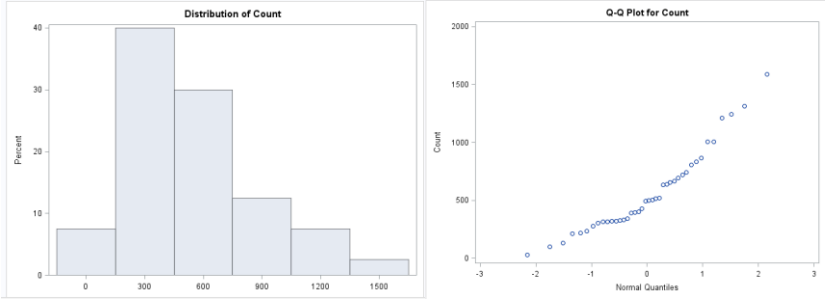
Variable 2	Levels	Description
count	Quantitative	Article Word Count

Word counts were captured as a covariate for each article. The word count analysis did not include the title of the author or extraneous text from ads or copyright information. Based on data distributions, there is a right (positive skew) for word count; however, it is not severe enough to justify applying a transformation to the variable given our observation count of 40. Further, we are using logistic regression, which is more robust to departures from normality than linear regression.

Figure 2 Means Table for Word Count

Analysis Variable : Count				
N	Mean	Std Dev	Minimum	Maximum
40	563.2750000	355.1435959	30.0000000	1591.00

Figure 3: Distribution Analysis for Word Count



Variable 3	Levels	Description
GENDER	3	Male, Female, Group Author(s)

The GENDER variable represents the sex of the author of the article. Gender was obtained via names and biographies of the writer of the article. In some cases (8), the news "staff" (a group) wrote the article, and no author gender could be determined. Females wrote 13 of the articles while males wrote 19 articles.

Level Labels:

1: Female **2:** Female **3:** Staff

Variable 4	Levels	Description
REACH	2	The population reach of the news outlet

To obtain a reach factor, each article was further sub-categorized based on local or regional reach, indicating the audience size for the outlet. County-level news articles, for instance, were considered local articles. The San Francisco Chronicle, on the other hand, was considered a regional news outlet. In order to confirm categorization of articles to a

reach level, www.stateofthemedia.org was used to identify outlet reach across states in the US.

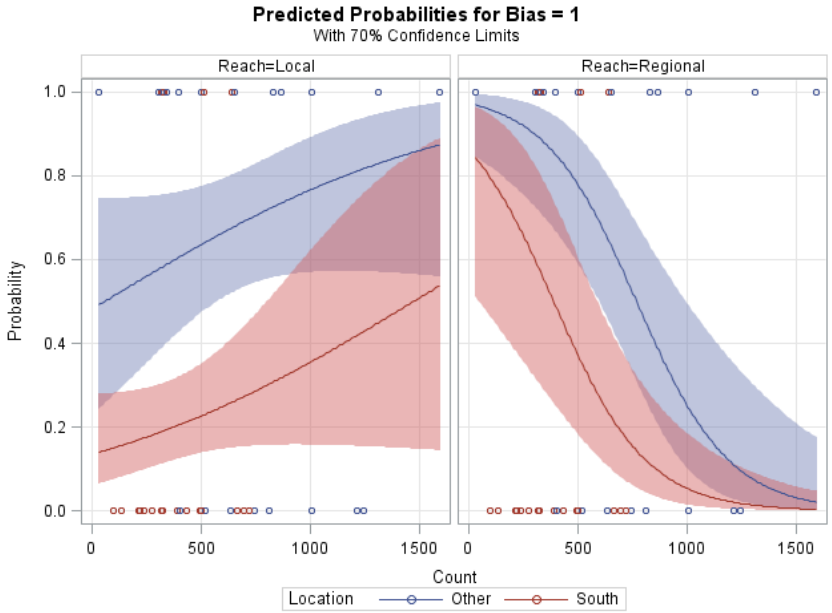
3.3 LOGISTIC REGRESSION ASSUMPTIONS VALIDATION

Utilizing binary logistic regression assumes the following:

1. Response is binary
2. Independent error terms
3. Linear relationship between **log odds or logistic function** and quantitative independent variables
4. Multivariate normality (very loose requirement)
5. Relatively large sample size (guidelines of N = p*10)

A sample size of 40 with four predictor variables, randomly selected news articles and a binary response variable help to ensure the logistic model is appropriate for analyzing media bias in the Kim Davis situation. Further, multivariate normality is somewhat relaxed in the case of logistic regression versus a typical linear regression. As noted in section 3.2, word count is not non-normal enough to cause issues.

Figure 4: Effect Plots For Word Count By Location Sliced by Levels of Reach



Effect plots indicate a mostly linear relationship between the logistic function results and word count for both local and regional news articles. Therefore it is safe to move forward with logistic regression using word count in the model.

Other relationships are also readily apparent from the effects plots; however, these relationships will be analyzed and explained thoroughly in section 4.

4. LOGISTIC REGRESSION

A binary response allows for the investigation of bias for each level of LOCATION while controlling for other covariates. We wish to investigate if there is a significant LOCATION effect in the logistic regression model for bias in news articles written about the Kim Davis gay rights event. Covariates reach, word count and author gender are included as part of our primary analysis. Formally, our null and alternative hypotheses can be read as follows:

$$H_0: LOCATION = 0 \; in \; presence \; of \; all \; covariates$$
$$H_a: LOCATION \neq 0 \; in \; presence \; of \; all \; covariates$$

Less formally, we are interested in the slope of location being statistically significant in the logistic regression model while also considering other covariates. In this case, the model with location is actually a parallel lines model in its simplest form with no interactions or quadratic terms considered. However, a rich model will be considered initially to account for interactions between independent variables region, reach, count and gender.

Let variables in the initial logistic regression model **Lc** equal LOCATION*count, **Rc** equal REACH*count and **Gc** equal GENDER*count. Thus, the initial logistic regression model is:

$$LOGIT(BIAS) = B_0 + LOCATION + REACH + GENDER + count + Lc + Rc + Gc$$

The Akaike's Information Criterion and Beta = 0 test were used in SAS to model the initial and subsequent logistic regression models in order to find a desired fit for modeling bias. A table summary is found in figure 5:

Figure 5: Logistic Regression Model Summary

MODEL PASS 1-3	AIC INTERCEPT ONLY	AIC FULL MODEL	Model Beta = 0 Test Significant
1 {LOGIT(BIAS) | LOC REACH GENDER count Lc Rc Gc }	56.5	59.4	No
2 {LOGIT(BIAS) | LOC REACH count Lc Rc}	56.5	53.8	No
3 {LOGIT(BIAS) | LOC REACH count Rc }	56.5	52.5	Yes

Gender was highly insignificant (p=0.87) in the initial model, thus gender and Gc (interaction of gender and count) was dropped from the analysis. Lc, or the interaction of location and count was also insignificant in the second model (p=0.4); however, AIC indicated that the independent variables in the model produced a better fit to the data than the intercept by itself.

Finally, modeling bias a function of location, reach, word count and the interaction of word count and reach proved to be significant based on the overall model beta = 0 test,

indicating that one or more of the independent variables are statistically significant. Thus, the third model was chosen as the final fit to the media bias data set.

Indeed, maximum likelihood estimates indicate all variables except word count are significant in model three. However, there is a significant interaction present between word count and the reach of the news outlet, indicating longer articles from local news outlets have greater odds of being biased. Therefore, word count is kept in the model in order to appropriately represent its presence in the significant interaction of word count and reach. The maximum likelihood estimates are given below in figure 6:

Figure 6

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.8765	0.9607	0.8324	0.3616
Location	Other	1	0.8933	0.4269	4.3794	0.0364
Reach	Local	1	-1.8426	0.9202	4.0091	0.0453
Count		1	-0.00172	0.00145	1.4011	0.2365
Count*Reach	Local	1	0.00298	0.00134	4.9827	0.0256

In order to interpret outputs in the form of odds, we must exponentiate two times the maximum likelihood (MLE) coefficients in order to obtain the normal odds ratios.

Based on the MLE coefficients from the third and final model, location shows that bias in non-south news outlets are 5.97 times the odds of south news outlets covering the Kim Davis story. Outlets with smaller reach (local outlets) are 0.025 times as likely to be biased as regional news outlets, indicating regional and national news outlets have greater (almost 40 times) odds of bias than do local news outlets in the presence of covariates location, article word count and the interaction of word count and reach. The interaction effect is significant, however slight, in the presence of covariates previously mentioned. Indicating longer local articles have greater odds to be biased.

We can easily determine the probability of article bias using the previously mentioned logistic regression model. Two examples are given below

Figure 7

Probability Model Examples:

Under the proposed logistic regression model:

{LOGIT(BIAS) | LOC REACH count Rc }

An article about Kim Davis before September 3rd 2015, written in a non-southern state by a local newspaper with 250 words would have a:

$$e\text{^}(2 * 0.8765 + 2 * 0.8933 + (2 * -1.8426) + (2 * -0.00172) * 250 + (2 * 0.0029 * 250)) \Big/ (1 + e\text{^}(2 * 0.8765 + 2 * 0.8933 + (2 * -1.8426) + (2 * -0.00172) * 250 + (2 * 0.00298) * 250)))$$

=

62% probability of being biased

If the article has regional reach and is written in a non-southern state there is a:

$$e\text{^}((2 * 0.8765 + 2 * 0.8933)))/(1 + e\text{^}((2 * 0.8765 + 2 * 0.8933)))$$

=

If the selected model is used to predict bias in an article, a probability threshold of 36% would be optimal to maximize correct classification. This data is based on the training data set itself and is used to classify the training set directly. Direct reference to the classification table can be found in appendix 2, figure 1. Therefore, appropriate train and test procedures would need to be employed to ensure model fit for prediction is optimal.

Figure 8

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
10.6213	8	0.2241

Finally, a Hosmer and Lemeshow Goodness of Fit test indicates the maximum likelihood estimate model is appropriate for our analysis (p value of 0.22 on chi-squared distribution) and no quasi-MLE model is necessary.

5. AN ADDITIONAL APPROACH

As part of the data analysis process, feedback was acquired from colleagues in regard to additional analysis methods that could handle answering the hypothesized question about media bias. One such approach proposed utilizing an unbiased, algorithmic approach for determining article sentiment. This approach was based on IBM's Alchemy API, an algorithmic approach with an interface that provides document sentiment scoring of text. In this case, sentiment was considered bias.

Alchemy API was used to score each of the 40 articles previously analyzed with logistic regression analysis. Scores were given on a negative one to positive one scale, with result values scoring out to 5 decimal places. For instance, a viable score could be 0.36423. Because of the large range of continuous values, a multiple linear regression was fit to the data in the hopes of not only obtaining a judgment on article bias, but also the severity of the bias. After analyzing the data, a threshold was set for what would be considered article bias. If an article scored below -0.25 or above 0.25, the article could be considered biased. Specifically, this means that 25% or more of the article could be considered as supportive or detracting from the subject of the article, in this case Kim Davis.

After retrieving the values from Alchemy API and running manual and automated model selection methods, no viable multiple linear regression model was obtained.

After much analysis, reasons for rejection of the MLR approach proved to be:

1) F and p values for the overall model that are high (1.45 and 0.247 respectively) and point to a MLR model that is not significant.
2) Alchemy API does not have the capability of assessing underlying bias and structural bias of articles detectable by humans
3) p-values for the individual variables, Count and Outreach, as well as the Pearson correlation coefficients showed non-significance and no to slight correlation

6. CONCLUSION

In general, binary logistic regression analysis indicates that online articles written by news outlets outside of southern states have greater odds and probability of being biased. Specifically, non-southern news outlets have a six times greater odds to be biased when covering the Kim Davis situation. Further, word count and reach of the news outlet are also significant covariates to consider when determining media bias in the Kim Davis situation, whereas gender proved to be a non-factor. This study is not all

encompassing from a variables perspective and many other variables could be considered in future studies on media bias.

While there is no explicit reason for why this seems to be the case, a larger look at the regional demographics may help explain the result. The south is often regarded as the "Bible Belt" with the inference that religious values are more deeply held by a larger number of people in the region resulting in a "consensus" among like-minded people. By extension, there are fewer people with a contrary position and an even smaller number of these are writers for news outlets and their opinions don't get distributed.

Cities in states outside of the "south" grouping are likely populated with a greater mix of cultures, values and perspectives and it's not unreasonable to consider that perhaps this multiculturalism leads to more bias and opinions about what could be seen as oppressive behavior by Kim Davis. More tolerance and acceptance in these cultural centers incites a more vocal response toward those seen to be less tolerant and in this study, Kim Davis' behavior is viewed as being intolerant of homosexual couples. Supporting this conjecture is San Francisco which is not in the south, is a large city with a diverse population in addition to a large and well known gay community. Bias views against Kim Davis are not surprising and even expected. The populations of New York, Los Angeles and Chicago share many attributes with population of San Francisco.

Due to limited quantities of articles from non-associated press authors, online publications and researcher influence in cleaning the sample for study, cautious inference can be made only to the states considered in the study. Further, inference cannot be made about bias outside of the Kim Davis gay rights situation in particular. This study does not prove that California news outlets are more biased than news outlets in Georgia, for instance.

APPENDIX 1 – SAS CODE

```
data test;
input Location $ Reach $ Bias Count Gender $;
DATALINES;
Other Regional    0     1241 F
Other Regional    0     1211 M
Other Regional    1     500 S
South Local 0     430 F
South Local 0     493 M
Other Local 0     519 F
Other Regional    1     322 S
```

```
Other Local 1     1591 M
Other Regional    0     808 F
Other Regional    1     867 S
Other Local 0     741 F
Other Local 1     831 F
South Local 0     315 M
South Local 0     99 F
South Local 1     315 F
South Local 0     501 S
South Regional    0     133 S
South Local 0     235 F
South Local 0     211 M
South Local 0     388 M
South Local 1     513 M
South Local 0     663 M
South Local 1     639 F
South Local 0     322 M
South Regional    0     717 M
South Local 0     277 S
South Local 1     323 F
South Regional    1     332 S
Other Local 0     402 S
Other Regional    1     655 F
Other Local 1     1007 M
Other Regional    1     304 F
Other Local 1     339 M
Other Local 1     1312 M
South Local 0     218 M
South Local 0     694 M
Other Local 1     396 M
Other Regional    0     1005 M
Other Local 1     30 M
Other Local 0     632 M
;

data logitout;
set test;
logcount = log(count);
sqrtcount = sqrt(count);
IF reach = 'Regional' THEN reaind=1; ELSE reaind=0;
IF location = 'Other' THEN locind = 1; ELSE locind =0;
IF gender = 'M' THEN genind = 1; ELSE genind =0;
react = reaind*count;
locct = locind*count;
genrct = genind*count;
run;

PROC MEANS data= test;
VAR count;
run;
PROC UNIVARIATE data = test;
VAR count;
HISTOGRAM;
QQPLOT;
RUN;

/* data does not veer from normality enough */
```

```
PROC UNIVARIATE data = logitout;
VAR logcount sqrtcount;
HISTOGRAM;
QQPLOT;
RUN;
/* Begin Model Investigations */
PROC LOGISTIC data = logitout DESCENDING;
CLASS LOCATION REACH GENDER;
MODEL bias = LOCATION REACH GENDER count react locct genrct;
RUN;
/* GENDER highly insignificant */
/* LOCATION interaction with count highly insignificant */
/* STEPWISE method for model selection - not optimal because of
experimentwise error rate */
PROC LOGISTIC data = test DESCENDING;
CLASS LOCATION REACH GENDER;
MODEL bias = location reach count gender / SELECTION = stepwise IPLOTS
INFLUENCE CL LACKFIT;
effectplot interaction (x = location sliceby = reach) / at(location = 'South'
'Other');
RUN;

/* THIS IS THE FINAL MODEL */
PROC LOGISTIC data = test DESCENDING;
CLASS LOCATION REACH GENDER;
MODEL bias = location reach | count / IPLOTS INFLUENCE CL LACKFIT CTABLE;
effectplot fit /obs(jitter(y=0.02));
effectplot slicefit (sliceby=location)/at(reach=all) clm alpha = .3;
output out = logits predprobs = I p=probpreb;
RUN;
```

MLR MODEL CODE

* EXCLUDING REDUNDANT CODE TO SAVE PAGES

```
/*Looking without Observation 3 from above*/
data news 3;
inputs Obs  Gender $    Location $  Outlet $    Alchemy     Count;
datalines;
1     F     Other Regional    -0.233345   1241
2     M     Other Regional    -0.0479982  1211
4     F     South Local -0.36221    430
5     M     South Local -0.0776174  493
6     F     Other Regional    -0.361426   519
7     S     Other Regional    -0.135748   322
8     M     Other Local -0.324497   1591
9     F     Other Regional    -0.35414    808
10    S     Other Regional    -0.341999   867
11    F     Other Regional    -0.342009   741
12    F     Other Local -0.0281417  831
13    M     South Local -0.126394   315
14    F     South Local -0.135021   99
15    F     South Local -0.0355124  315
16    S     South Local -0.15003    501
17    S     South Regional    -0.217917   133
18    F     South Local -0.22534    235
```

```
19    M     South Local -0.429925   211
20    M     South Local -0.475668   388
21    M     South Local 0.0767464   513
22    M     South Local -0.332041   663
23    F     South Local -0.199439   639
24    M     South Local -0.276852   322
25    M     South Regional    -0.310649   717
26    S     South Local -0.487447   277
27    F     South Local -0.201707   323
28    S     South Regional    -0.492445   332
29    S     Other Local -0.259158   402
30    F     Other Regional    -0.266939   655
31    M     Other Local -0.204525   1007
32    F     Other Regional    -0.316329   304
33    M     Other Local -0.0890564  339
34    M     Other Local -0.182837   1312
35    M     South Local -0.254262   218
36    M     South Local 0.0613189   694
37    M     Other Local -0.540062   396
38    M     Other Regional    -0.165739   1005
39    M     Other Local 0.147505    632
40    F     Other Local -0.199439   702
;

data news 4;
set news 3;
if Gender='M' then Gen Cat = 0;
      else if Gender='F' then Gen_Cat = 1;
      else Gen Cat = 2;
if Location='South' then Loc_Cat = 0;
      else Loc Cat = 1;
if Outlet='Regional' then Out_Cat = 0;
      else Out_Cat = 1;
;

proc corr data = news 4;
var Alchemy Count Gen_Cat Loc_Cat Out_Cat;
run;

proc reg data = news 4;
model Alchemy = Count Gen_Cat Loc_Cat Out_Cat / selection = stepwise;
run;

data news 5;
set news 2;
log alchemy = log(abs(Alchemy));
log_count = log(Count);
;

proc reg data = news 5;
model log_alchemy = log_count Gen_Cat Loc_Cat Out_Cat / selection = stepwise;
run;

ods graphics on;
proc reg data = news 5;
model log_alchemy = Out_Cat / VIF R;
run;
```

```
quit;
ods graphics off;
```

APPENDIX 2 – FIGURES

Figure 1 – Classification Table For Logistic Regression Model 3

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.120	17	0	23	0	42.5	100.0	0.0	57.5	.
0.140	17	1	22	0	45.0	100.0	4.3	56.4	0.0
0.160	15	2	21	2	42.5	88.2	8.7	58.3	50.0
0.180	15	3	20	2	45.0	88.2	13.0	57.1	40.0
0.200	13	8	15	4	52.5	76.5	34.8	53.6	33.3
0.220	13	11	12	4	60.0	76.5	47.8	48.0	26.7
0.240	13	13	10	4	65.0	76.5	56.5	43.5	23.5
0.260	13	14	9	4	67.5	76.5	60.9	40.9	22.2
0.280	12	14	9	5	65.0	70.6	60.9	42.9	26.3
0.300	11	15	8	6	65.0	64.7	65.2	42.1	28.6
0.320	11	17	6	6	70.0	64.7	73.9	35.3	26.1
0.340	11	17	6	6	70.0	64.7	73.9	35.3	26.1
0.360	11	17	6	6	70.0	64.7	73.9	35.3	26.1
0.380	11	17	6	6	70.0	64.7	73.9	35.3	26.1