

## **Kappa test and measure of agreement among three assessments of avian collision risk classifications of wind turbines in Altamont**

Analysis by Julie Yee  
November 29, 2006

Objective: Test and measure of agreement among three assessments of risk classifications, because there had been questions raised about possible biases in the assessment causing turbines to be classified in opposite risk tiers depending on whether turbine size was accounted for.

Methods: Using the risk classifications of each turbine assessed in the January, March, and June versions of the Smallwood and Siegel reports (2005), I calculated Cohen's Kappa, a measure of agreement between any two assessments (for example, between January and March, January and June, and March and June) (Agresti, 2002). The Kappa statistic requires risk ratings to range over the same scale. Since the risk scales differed among assessments (January and March involved a scale from 1 to 5, and June a scale from 1 to 6), then I used grouping to reduce the number of classifications so that all assessments had similar scales.

I evaluated the frequency tables for the risk classifications of the three assessments and noted shifts in the weight of distribution (p. 1, SAS output). In particular, turbines rated using the January assessment tended to have higher rating values (lower risk tier) compared to March (January mean = 4.30; March mean = 4.08). June had the highest mean (5.25), which is an artifact of the larger scale. Assessments having different distributions can lead to apparent differences in the ratings even though all three assessments might classify the top 15% of riskiest turbines similarly. To reduce any disagreements that might be artifacts resulting from distributional differences, I reclassified the ratings to a scale of 4 using these following rules.

January assessment: lump tiers 1 and 2 into top tier  
March assessment: lump tiers 3 and 4 into third tier  
June assessment: lump tiers 1-3 into top tier

Under the 4-tier classification, the new top tier contains 5-8% of turbines, the new top two tiers cumulatively contain 13-16%, and the new top three tiers 43-50%. The new mean tier ratings for January, March, and June are 3.34, 3.31, and 3.30 respectively.

Considering the small proportion of turbines in the top classification, I repeated the analysis for a 3-tier classification, derived by lumping the first two tiers above into one top tier. In other words,

January assessment: lump tiers 1-3 into top tier  
March assessment: lump tiers 1-2 into top tier and 3-4 into next tier  
June assessment: lump tiers 1-4 into top tier

The mean tier ratings using the 3-tier classification are 2.41, 2.36, and 2.38 for January, March, and June respectively.

I conducted 6 sets of Kappa calculations, one for each pair of months and each new tier classification. The Kappa statistic ranges from a negative minimum value (actual minimum depends on analysis) to a maximum of +1. Negative values indicate disagreement (opposite ratings), positive values indicate agreement with a +1 indicating perfect agreement, and zero indicating independence (as if turbines had been randomly scrambled between assessments).<sup>1</sup>

Results: All Kappa calculations are reported on pages 5-7 and 10-12 of SAS software output. All Kappa statistics were significantly greater than 0, indicating a tendency for all assessments to agree. January and March were the two assessments that agreed the least (95% CI for Kappa: 29-34% for 4-tier classification, 31-36% for 3-tier classification). January and June were similar in their levels of agreement (95% CI for Kappa: 34-39% for 4-tier classification, 36-41% for 3-tier classification). March and June had rather high levels of agreement (76-79% for 4-tier classification, 82-84% for 3-tier classification).

Discussion: I haven't found much guidance in the literature on interpreting Kappa values. Agresti (2002) refers to Kappa between 40-60% as "moderately strong agreement." I would loosely expand to that interpretation by describing levels of 70-80% as being strong agreement, and 30-40% as being mild to moderate agreement.

I interpret levels of 30-40% as indicating substantial differences between assessments, but generally more agreement than disagreement. A caveat of this analysis is that Kappa was not calculated separately for the different turbine size classes. While unconfirmed, it's possible that the assessments agreed better for turbines of certain size classes (i.e., mid-size) than turbines of other size classes (i.e. extreme sizes). I don't have turbine size data to correspond with risk tier data, so I didn't explore this possibility.

#### References:

- Agresti, A. 2002. Categorical Data Analysis. John Wiley & Sons Inc., Hoboken, NJ.
- Smallwood, S. and L. Spiegel. January 2005. Assessment to Support an Adaptive Management Plan for the APWRA. California Energy Commission Report.
- Smallwood, S. and L. Spiegel. June 2005. Combining Biology-Based and Policy-Based Tiers of Priority for Determining Wind Turbine Relocation/Shutdown to reduce bird fatalities in the APWRA. California Energy Commission Report.
- Smallwood, S. and L. Spiegel. March 2005. Partial Re-assessment of an Adaptive Management Plan for the APWRA: Accounting for Turbine Size. California Energy Commission Report.

---

<sup>1</sup> There are two types of Kappa statistics. The Simple Kappa treats all disagreements similarly (i.e., a turbine assessed "1" in January and "2" in March is an equivalently-sized disagreement to having it assessed "4" in March). The Weighted Kappa weights "1" and "4" to be a bigger disagreement than "1" and "2". By default, SAS produces both types of Kappa, but for discussion I refer just to Weighted Kappa.

The FREQ Procedure

January

January	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	130	3.20	130	3.20
2	164	4.04	294	7.24
3	359	8.84	653	16.09
4	1097	27.03	1750	43.11
5	2309	56.89	4059	100.00

March

March	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	205	5.05	205	5.05
2	337	8.30	542	13.35
3	411	10.13	953	23.48
4	1098	27.05	2051	50.53
5	2008	49.47	4059	100.00

July

July	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	54	1.33	54	1.33
2	101	2.49	155	3.82
3	152	3.74	307	7.56
4	296	7.29	603	14.86
5	1323	32.59	1926	47.45
6	2133	52.55	4059	100.00

## The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
January	January	150	4.3035230	5.2584605	1.0000000	5.0000000
March	March	150	4.0758808	6.1243428	1.0000000	5.0000000
July	July	150	5.2498152	5.5300809	1.0000000	6.0000000

Tiers scaled to 4 levels

3

January assessment: lump tiers 1 and 2 into top tier

March assessment: lump tiers 3 and 4 into third tier

June assessment: lump tiers 1-3 into top tier

12:59 Thursday, November 30, 2006

The FREQ Procedure

jan	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	294	7.24	294	7.24
2	359	8.84	653	16.09
3	1097	27.03	1750	43.11
4	2309	56.89	4059	100.00

mar	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	205	5.05	205	5.05
2	337	8.30	542	13.35
3	1509	37.18	2051	50.53
4	2008	49.47	4059	100.00

jun	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	307	7.56	307	7.56
2	296	7.29	603	14.86
3	1323	32.59	1926	47.45
4	2133	52.55	4059	100.00

Tiers scaled to 4 levels  
Mean tier value, by assessment

4

12:59 Thursday, November 30, 2006

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
jan	150	3.3355506	4.7677468	1.0000000	4.0000000
mar	150	3.3106677	4.3142095	1.0000000	4.0000000
jun	150	3.3013057	4.6979241	1.0000000	4.0000000

The FREQ Procedure

Table of jan by mar

		jan				mar				
Frequency	Percent					Total				
		1	2	3	4					
Row Pct	Col Pct									
1		57	60	71	106	294				
	1.40	1.48	1.75	2.61	7.24					
	19.39	20.41	24.15	36.05						
	27.80	17.80	4.71	5.28						
2		61	86	153	59	359				
	1.50	2.12	3.77	1.45	8.84					
	16.99	23.96	42.62	16.43						
	29.76	25.52	10.14	2.94						
3		48	89	598	362	1097				
	1.18	2.19	14.73	8.92	27.03					
	4.38	8.11	54.51	33.00						
	23.41	26.41	39.63	18.03						
4		39	102	687	1481	2309				
	0.96	2.51	16.93	36.49	56.89					
	1.69	4.42	29.75	64.14						
	19.02	30.27	45.53	73.75						
Total		205	337	1509	2008	4059				
	5.05	8.30	37.18	49.47	100.00					

Statistics for Table of jan by mar

Test of Symmetry

Statistic (S)	164.5135
DF	6
Pr > S	<.0001

Kappa Statistics

Statistic	Value	ASE	95% Confidence Limits	
Simple Kappa	0.2545	0.0117	0.2316	0.2774
Weighted Kappa	0.3167	0.0124	0.2923	0.3410

Sample Size = 4059

The FREQ Procedure

Table of jan by jun

jan		jun				
Frequency	Percent	1	2	3	4	Total
Row Pct	Col Pct					
1	95	31	62	106	294	
	2.34	0.76	1.53	2.61	7.24	
	32.31	10.54	21.09	36.05		
	30.94	10.47	4.69	4.97		
2	116	71	113	59	359	
	2.86	1.75	2.78	1.45	8.84	
	32.31	19.78	31.48	16.43		
	37.79	23.99	8.54	2.77		
3	69	128	519	381	1097	
	1.70	3.15	12.79	9.39	27.03	
	6.29	11.67	47.31	34.73		
	22.48	43.24	39.23	17.86		
4	27	66	629	1587	2309	
	0.67	1.63	15.50	39.10	56.89	
	1.17	2.86	27.24	68.73		
	8.79	22.30	47.54	74.40		
Total	307	296	1323	2133	4059	
	7.56	7.29	32.59	52.55	100.00	

Statistics for Table of jan by jun

Test of Symmetry

Statistic (S)	158.6692
DF	6
Pr > S	<.0001

Kappa Statistics

Statistic	Value	ASE	95% Confidence Limits	
Simple Kappa	0.2675	0.0117	0.2445	0.2905
Weighted Kappa	0.3646	0.0123	0.3404	0.3887

Sample Size = 4059

The FREQ Procedure

Table of mar by jun

mar		jun				
Frequency	Percent	1	2	3	4	Total
Row Pct	Col Pct					
1	118	43	42	2	205	
	2.91	1.06	1.03	0.05	5.05	
	57.56	20.98	20.49	0.98		
	38.44	14.53	3.17	0.09		
2	189	31	116	1	337	
	4.66	0.76	2.86	0.02	8.30	
	56.08	9.20	34.42	0.30		
	61.56	10.47	8.77	0.05		
3	0	222	1165	122	1509	
	0.00	5.47	28.70	3.01	37.18	
	0.00	14.71	77.20	8.08		
	0.00	75.00	88.06	5.72		
4	0	0	0	2008	2008	
	0.00	0.00	0.00	49.47	49.47	
	0.00	0.00	0.00	100.00		
	0.00	0.00	0.00	94.14		
Total	307	296	1323	2133	4059	
	7.56	7.29	32.59	52.55	100.00	

Statistics for Table of mar by jun

Test of Symmetry

Statistic (S)	292.1219
DF	6
Pr > S	<.0001

Kappa Statistics

Statistic	Value	ASE	95% Confidence Limits	
Simple Kappa	0.7018	0.0087	0.6848	0.7189
Weighted Kappa	0.7757	0.0067	0.7627	0.7888

Sample Size = 4059

Tiers scaled to 3 levels

8

January assessment: lump tiers 1-3 into top tier  
March assessment: lump tiers 1-2 into top tier and tiers 3-4 into second tier  
June assessment: lump tiers 1-4 into top tier

12:59 Thursday, November 30, 2006

The FREQ Procedure

jan	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	653	16.09	653	16.09
2	1097	27.03	1750	43.11
3	2309	56.89	4059	100.00

mar	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	542	13.35	542	13.35
2	1509	37.18	2051	50.53
3	2008	49.47	4059	100.00

jun	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	603	14.86	603	14.86
2	1323	32.59	1926	47.45
3	2133	52.55	4059	100.00

Tiers scaled to 3 levels  
Mean tier value, by assessment

9

12:59 Thursday, November 30, 2006

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
jan	150	2.4079823	3.9172492	1.0000000	3.0000000
mar	150	2.3611727	3.6824642	1.0000000	3.0000000
jun	150	2.3769401	3.8068127	1.0000000	3.0000000

The FREQ Procedure

Table of jan by mar

jan		mar			
Frequency	Percent	Row Pct	Col Pct		
		1	2	3	Total
1		264	224	165	653
		6.50	5.52	4.07	16.09
		40.43	34.30	25.27	
		48.71	14.84	8.22	
2		137	598	362	1097
		3.38	14.73	8.92	27.03
		12.49	54.51	33.00	
		25.28	39.63	18.03	
3		141	687	1481	2309
		3.47	16.93	36.49	56.89
		6.11	29.75	64.14	
		26.01	45.53	73.75	
Total		542	1509	2008	4059
		13.35	37.18	49.47	100.00

Statistics for Table of jan by mar

Test of Symmetry

Statistic (S)	123.5402
DF	3
Pr > S	<.0001

Kappa Statistics

Statistic	Value	ASE	95% Confidence Limits	
Simple Kappa	0.2914	0.0124	0.2671	0.3158
Weighted Kappa	0.3377	0.0129	0.3124	0.3631

Sample Size = 4059

The FREQ Procedure

Table of jan by jun

jan		jun			
Frequency	Percent	Row Pct	Col Pct		
		1	2	3	Total
1		313	175	165	653
		7.71	4.31	4.07	16.09
		47.93	26.80	25.27	
		51.91	13.23	7.74	
2		197	519	381	1097
		4.85	12.79	9.39	27.03
		17.96	47.31	34.73	
		32.67	39.23	17.86	
3		93	629	1587	2309
		2.29	15.50	39.10	56.89
		4.03	27.24	68.73	
		15.42	47.54	74.40	
Total		603	1323	2133	4059
		14.86	32.59	52.55	100.00

Statistics for Table of jan by jun

Test of Symmetry

Statistic (S)	82.2891
DF	3
Pr > S	<.0001

Kappa Statistics

Statistic	Value	ASE	95% Confidence Limits	
Simple Kappa	0.3141	0.0125	0.2895	0.3387
Weighted Kappa	0.3832	0.0127	0.3582	0.4082

Sample Size = 4059

The FREQ Procedure

Table of mar by jun

mar		jun			
Frequency	Percent	Row Pct	Col Pct		
		1	2	3	Total
1		381	158	3	542
		9.39	3.89	0.07	13.35
		70.30	29.15	0.55	
		63.18	11.94	0.14	
2		222	1165	122	1509
		5.47	28.70	3.01	37.18
		14.71	77.20	8.08	
		36.82	88.06	5.72	
3		0	0	2008	2008
		0.00	0.00	49.47	49.47
		0.00	0.00	100.00	
		0.00	0.00	94.14	
Total		603	1323	2133	4059
		14.86	32.59	52.55	100.00

Statistics for Table of mar by jun

Test of Symmetry

Statistic (S)	135.7789
DF	3
Pr > S	<.0001

Kappa Statistics

Statistic	Value	ASE	95% Confidence Limits	
Simple Kappa	0.7923	0.0083	0.7761	0.8085
Weighted Kappa	0.8315	0.0067	0.8183	0.8447

Sample Size = 4059