

Relative Risk and Attributable Risk Analysis

Thomas Kincaid

October 23, 2015

Contents

1	Introduction	1
2	Preliminaries	2
3	Load the survey design and analytical variables data set	2
4	Relative risk analysis	5
5	Attributable risk analysis	9

1 Introduction

This document presents relative risk and attributable risk analysis of a GRTS survey design. The resource employed in the analysis is lakes in the 48 contiguous United States. Data was obtained from the National Lakes Survey (NLA) that was conducted in 2007 by the U.S. Environmental Protection Agency ([2009](#)). Relative risk measures the strength of association between stressor and response variables that can be classified as either "good" (i.e., reference condition) or "poor" (i.e., different from reference condition). Attributable risk measures the percent reduction in the extent of poor condition of a response variable that presumably would result from eliminating a stressor variable. Discussion regarding relative risk in the context of aquatic resource surveys is provided in Van Sickle et. al. ([2006](#)) and Van Sickle and Paulsen ([2008](#)).

2 Preliminaries

The initial step is to use the library function to load the spsurvey package. After the package is loaded, a message is printed to the R console indicating that the spsurvey package was loaded successfully.

Load the spsurvey package

```
> # Load the spsurvey package
> library(spsurvey)
>
```

Version 3.1 of the spsurvey package was loaded successfully.

3 Load the survey design and analytical variables data set

The original NLA data file contains more than 1,000 records. To produce a more manageable number of records, lakes located in the western U.S. were retained in the data that will be analyzed, which produced a data set containing 236 records.

The next step is to load the data set, which includes both survey design variables and analytical variables. The data function is used to load the data set and assign it to a data frame named NLA_2007. The nrow function is used to determine the number of rows in the NLA_2007 data frame, and the resulting value is assigned to an object named nr. Finally, the initial six lines and the final six lines in the NLA_2007 data frame are printed using the head and tail functions, respectively.

Load the survey design and analytical variables data set

```
> # Load the data set and determine the number of rows in the data frame
> data(NLA_2007)
> nr <- nrow(NLA_2007)
>
```

Display the initial six lines in the data file.

```
> # Display the initial six lines in the data file
> head(NLA_2007)
```

	siteID	xcoord	ycoord	wgt	Lake_Origin	Chla	OE5	PTL
1	NLA06608-0001	-1327628.1	3012181	7.594532	Natural	0.240	0.504031	6

```

2 NLA06608-0004 -1084415.8 1668316 9.171940 Man-Made 4.600 1.032252 18
3 NLA06608-0005 -1497348.8 2475338 15.027385 Natural 1.205 0.988630 4
4 NLA06608-0015 -1044530.9 1166122 6.920957 Man-Made 20.000 0.918628 109
5 NLA06608-0033 -1901234.0 2956669 32.549373 Natural 8.920 0.673385 67
6 NLA06608-0042 -874392.3 2436245 9.832508 Man-Made 2.208 0.860663 15
  NTL Turbidity Chla_cond OE5_cond PTL_cond NTL_cond Turbidity_cond
1 151 0.474 Good Poor Good Good Good
2 344 3.810 Poor Good Good Good Good
3 85 0.475 Good Good Good Good Good
4 470 32.700 Good Good Good Good Poor
5 835 12.200 Poor Good Poor Poor Poor
6 213 0.791 Good Good Good Good Good

```

>

Display the final six lines in the data file.

```

> # Display the final six lines in the data file
> tail(NLA_2007)

```

```

      siteID  xcoord  ycoord      wgt Lake_Origin  Chla      OE5 PTL
231 NLA06608-3121 -1599693 2614663 4.035550 Man-Made 14.640 0.709114 46
232 NLA06608-3153 -1970907 3130822 7.938297 Natural 1.499 0.737076 1
233 NLA06608-3157 -1581199 2449359 4.035550 Man-Made 2.208 0.922396 8
234 NLA06608-3265 -1595910 2964913 21.498248 Natural 1.768 0.648352 7
235 NLA06608-3313 -1294482 2232798 3.399432 Man-Made 7.728 0.592139 41
236 NLA06608-3329 -1543474 2998349 3.664951 Natural 3.704 0.991219 10
  NTL Turbidity Chla_cond OE5_cond PTL_cond NTL_cond Turbidity_cond
231 455 5.720 Poor Good Poor Poor Poor
232 116 0.420 Good Good Good Good Good
233 70 1.790 Good Good Good Good Good
234 338 0.561 Good Good Good Good Good
235 316 5.670 Good Poor Good Good Good
236 374 1.050 Poor Good Good Good Good

```

>

The location of lakes that were sampled in the western United States is displayed in Figure 1. The sample sites are displayed using a unique color for the two values of lake origin (natural and manmade).

Plot of NLA Sample Sites Color-Coded by Lake Origin

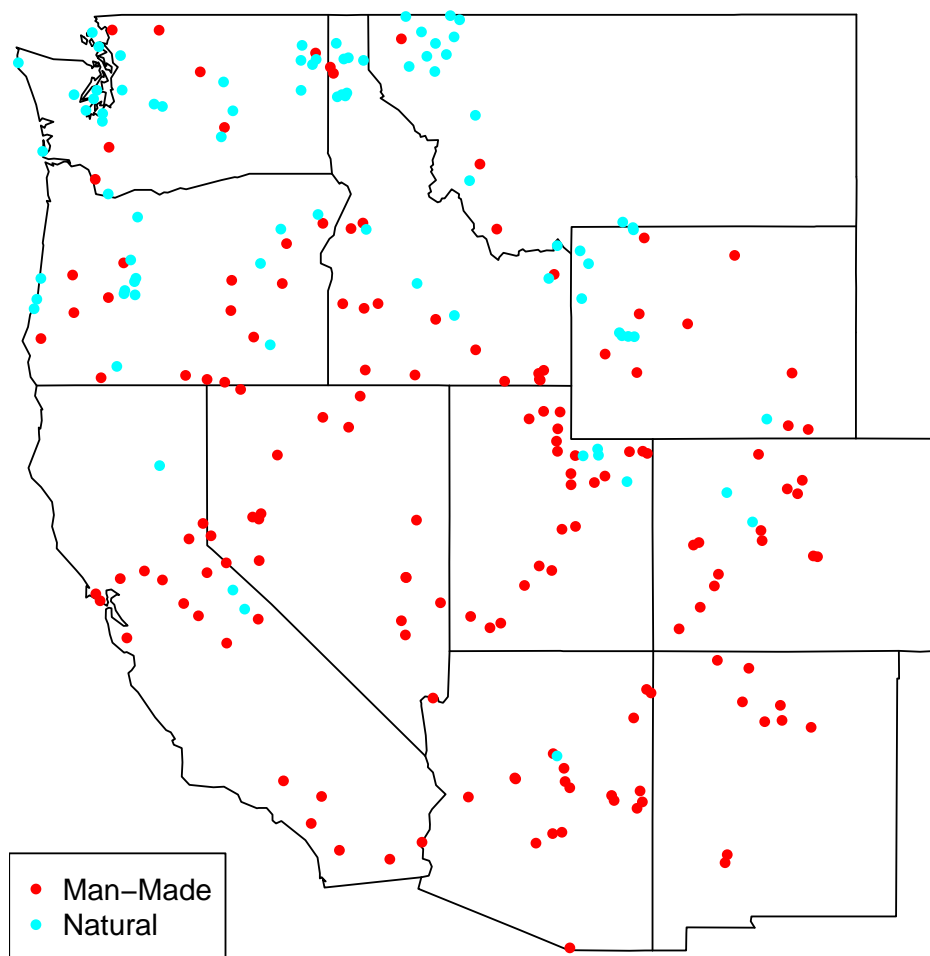


Figure 1: Location of lakes that were sampled in the western United States by the U.S. Environmental Protection Agency during the National Lakes Assessment (NLA) conducted in 2007.

4 Relative risk analysis

Relative risk analysis will be investigated by examining two response variables and three stressor variables. The response variables are chlorophyll-a concentration for each sample site, which is a measure of trophic condition, and an index of macroinvertebrate taxa loss that is based on modeling the ratio of observed and expected loss. The stressor variables are total nitrogen concentration, total phosphorus concentration, and turbidity for each site.

The `relrisk.analysis` function will be used to calculate relative risk estimates. Four data frames constitute the primary input to the `relrisk.analysis` function. The first column (variable) in the four data frames provides the unique identifier (site ID) for each sample site and is used to connect records among the data frames. The `siteID` variable in the `NLA_2007` data frame is assigned to the `siteID` variable in the data frames. The four data frames that will be created are named as follows: `sites`, `subpop`, `design`, and `data.risk`. The `sites` data frame identifies sites to use in the analysis and contains two variables: (1) `siteID` - site ID values and (2) `Use` - a logical vector indicating which sites to use in the analysis. Since we want to include all sampled sites, the `rep` (repeat) function is used to assign the value `TRUE` to each element of the `Use` variable. Recall that `nr` is an object containing the number of rows in the `NLA_2007` data frame. The `subpop` data frame defines populations and, optionally, subpopulations for which estimates are desired. Unlike the `sites` and `design` data frames, the `subpop` data frame can contain an arbitrary number of columns. The first variable in the `subpop` data frame identifies site ID values and each subsequent variable identifies a type of population, where the variable name is used to identify type. A type variable identifies each site with a character value. If the number of unique values for a type variable is greater than one, then the set of values represent subpopulations of that type. When a type variable consists of a single unique value, then the type does not contain subpopulations. For this analysis, the `subpop` data frame contains three variables: (1) `siteID` - site ID values, (2) `Western_US` - which will be used to calculate estimates for all of the sample sites combined, and (3) `Lake_Origin` - which will be used to calculate estimates for each of the two classes of lake origin (natural and manmade). The `rep` function is used to assign values to the `Western_US` variable, and the `Lake_Origin` variable in the `NLA_2007` data frame is assigned to the `Lake_Origin` variable in the `subpop` data frame. The `design` data frame consists of survey design variables. For the analysis under consideration, the `design` data frame contains the following variables: (1) `siteID` - site ID values; (2) `wgt` - final, adjusted, survey design weights; (3) `xcoord` - x-coordinates for location; and (4) `ycoord` - y-coordinates for location. The `wgt`, `xcoord`, and `ycoord` variables in the `design` data frame are assigned values using variables with the same names in the `NLA_2007` data frame. Like the `subpop` data frame, the `data.risk` data frame can contain an arbitrary number of columns. The first variable in the `data.risk` data frame identifies site ID values and each subsequent variable identifies a response or stressor variable. For this analysis, the response variables are `Chlorophylla` and `MacroInvert_OE`, and the stressor variables are `Total_Nitrogen`, `Total_Phosphorus`, and `Turbidity`, which are assigned, respectively, variables `Chla_cond`, `OE5_cond`, `PTL_cond`, `NTL_cond`, and `Turbidity_cond` in the `NLA_2007` data frame.

Create the `sites` data frame.

```
> sites <- data.frame(siteID=NLA_2007$siteID,
+                      Use=rep(TRUE, nr))
```

Create the subpop data frame.

```
> subpop <- data.frame(siteID=NLA_2007$siteID,
+                      Western_US=rep("Western_US", nr),
+                      Lake_Origin=NLA_2007$Lake_Origin)
```

Create the design data frame.

```
> design <- data.frame(siteID=NLA_2007$siteID,
+                      wgt=NLA_2007$wgt,
+                      xcoord=NLA_2007$xcoord,
+                      ycoord=NLA_2007$ycoord)
```

Create the data.risk data frame.

```
> data.risk <- data.frame(siteID=NLA_2007$siteID,
+                         Chlorophyll_a=NLA_2007$Chla_cond,
+                         MacroInvert_OE=NLA_2007$OE5_cond,
+                         Total_Nitrogen=NLA_2007$NTL_cond,
+                         Total_Phosphorus=NLA_2007$PTL_cond,
+                         Turbidity=NLA_2007$Turbidity_cond)
```

Names of the response and stressor variables for which relative risk estimates are desired must be specified. The values "Chlorophyll_a" and "MacroInvert_OE" are assigned to resp_vars. The values "Total_Nitrogen", "Total_Phosphorus", and "Turbidity" are assigned to stress_vars.

Assign response and stressor variable names.

```
> resp_vars <- c("Chlorophyll_a", "MacroInvert_OE")
+ stress_vars <- c("Total_Nitrogen", "Total_Phosphorus", "Turbidity")
```

The relrisk.analysis function is used to calculate relative risk estimates. In the call to function relrisk.analysis, arguments response.var and stressor.var provide the names of columns in the data.risk data frame that contain response variables and stressor variables, respectively. The rep function is used to repeat each of the response variable names in resp_vars once for each of the stressor variable names in stress_vars and the result is assigned to the response.var argument. Similarly, the rep function is used to repeat the set of stressor variable names in stress_vars once for each of the values in resp_var and the result is assigned to the stressor.var argument. The result is that relrisk.analysis will calculate a relative risk estimate for each combination of response and stressor variables.

Calculate relative risk estimates.

```
> RelRisk_Estimates <- relrisk.analysis(sites, subpop, design, data.risk,
+   response.var= rep(resp_vars, each=length(stress_vars)),
+   stressor.var=rep(stress_vars, length(resp_vars)))
```

The relative risk estimates are displayed using the print function. The object produced by relrisk.analysis is a data frame containing twentyone columns. The first five columns identify the population (Type), subpopulation (Subpopulation), response variable (Response), stressor variable (Stressor), and number of response variable (or stressor variable) values used to calculate the relative risk estimate (NResp). The next six columns provide results for the relative risk estimate: the estimate (Estimate), numerator of the estimate (Estimate.num), denominator of the estimate (Estimate.denom), logarithm of the standard error of the estimate (StdError.log), lower confidence bound (LCB95Pct), and upper confidence bound (UCB95Pct). Argument conf for relrisk.analysis allows control of the confidence bound level. The default value for conf is 95, hence the column names for confidence bounds contain the value 95. Supplying a different value to the conf argument will be reflected in the confidence bound names. Confidence bounds are obtained using the logarithm of standard error and the Normal distribution multiplier corresponding to the confidence level. Results are then backtransformed to the original scale to obtain the confidence bound estimates. The next column in the data frame contains the sum of the survey design weights (WeightTotal). The next four columns provide cell counts for the two-by-two table of response variable categories and stressor variable categories and are named CellCounts.rc, where r indicates row number in the table and c indicates column number. Rows contain the response variable categories and column contain the stressor variable categories. By default, number 1 is the "Poor" category, and number 2 is the "Good" category. The final four columns in the data frame contain the cell proportion estimates for the two-by-two table, where columns are named CellProportions.rc using the same convention as the cell count columns. Note that the cell proportion estimates are weighted estimates obtained using the survey design weights.

```
> # Print the relative risk estimates
> print(RelRisk_Estimates)
```

	Type	Subpopulation	Response	Stressor	NResp	Estimate
1	Western_US	Western_US	Chlorophyll_a	Total_Nitrogen	236	2.7274819
2	Lake_Origin	Man-Made	Chlorophyll_a	Total_Nitrogen	152	2.4104712
3	Lake_Origin	Natural	Chlorophyll_a	Total_Nitrogen	84	1.8529034
4	Western_US	Western_US	Chlorophyll_a	Total_Phosphorus	236	2.2817129
5	Lake_Origin	Man-Made	Chlorophyll_a	Total_Phosphorus	152	2.4506118
6	Lake_Origin	Natural	Chlorophyll_a	Total_Phosphorus	84	1.4260122
7	Western_US	Western_US	Chlorophyll_a	Turbidity	236	1.8703252
8	Lake_Origin	Man-Made	Chlorophyll_a	Turbidity	152	1.0599859
9	Lake_Origin	Natural	Chlorophyll_a	Turbidity	84	4.7199940
10	Western_US	Western_US	MacroInvert_OE	Total_Nitrogen	234	2.7177105
11	Lake_Origin	Man-Made	MacroInvert_OE	Total_Nitrogen	151	1.4635342
12	Lake_Origin	Natural	MacroInvert_OE	Total_Nitrogen	83	3.0081542

13	Western_US	Western_US	MacroInvert_OE	Total_Phosphorus	234	1.5114042
14	Lake_Origin	Man-Made	MacroInvert_OE	Total_Phosphorus	151	0.8614559
15	Lake_Origin	Natural	MacroInvert_OE	Total_Phosphorus	83	3.8828510
16	Western_US	Western_US	MacroInvert_OE	Turbidity	234	5.1694904
17	Lake_Origin	Man-Made	MacroInvert_OE	Turbidity	151	2.9228876
18	Lake_Origin	Natural	MacroInvert_OE	Turbidity	83	4.6573413
	Estimate.num	Estimate.denom	StdError.log	LCB95Pct	UCB95Pct	WeightTotal
1	0.5869872	0.21521213	0.4309409	1.1720452	6.347160	4890.777
2	0.7072461	0.29340576	0.2913936	1.3616556	4.267137	2049.445
3	0.3371698	0.18196839	0.7970782	0.3884889	8.837450	2841.333
4	0.5381828	0.23586790	0.4265909	0.9888859	5.264727	4890.777
5	0.7786874	0.31775225	0.2763906	1.4256417	4.212488	2049.445
6	0.2688624	0.18854142	0.7579254	0.3228315	6.298984	2841.333
7	0.5582292	0.29846639	0.3918041	0.8677865	4.031079	4890.777
8	0.5278154	0.49794564	0.3971241	0.4867069	2.308515	2049.445
9	0.9277937	0.19656671	0.4505363	1.9518403	11.414020	2841.333
10	0.3015402	0.11095377	0.4505836	1.1237397	6.572652	4882.983
11	0.3753646	0.25647817	0.4703811	0.5821217	3.679527	2045.360
12	0.1476541	0.04908460	0.7608096	0.6771702	13.362950	2837.623
13	0.2225565	0.14725151	0.5029961	0.5639357	4.050715	4882.983
14	0.2902637	0.33694546	0.6017883	0.2648439	2.802052	2045.360
15	0.1467371	0.03779108	0.7215134	0.9440553	15.969968	2837.623
16	0.5656631	0.10942338	0.3707848	2.4993963	10.692034	4882.983
17	0.5866627	0.20071339	0.4241173	1.2729250	6.711528	2045.360
18	0.2924337	0.06278985	0.8528069	0.8754424	24.776992	2837.623
	CellCounts.11	CellCounts.12	CellCounts.21	CellCounts.22	CellProportions.11	
1	41	33	22	140	0.188105511	
2	29	20	17	86	0.365103884	
3	12	13	5	54	0.060437097	
4	39	35	8	154	0.175320038	
5	26	23	5	98	0.319782522	
6	13	12	3	56	0.071119667	
7	22	52	16	146	0.077115692	
8	19	30	15	88	0.160770986	
9	3	22	1	58	0.016775371	
10	28	26	33	147	0.096304303	
11	23	17	22	89	0.193412993	
12	5	9	11	58	0.026308304	
13	17	37	30	150	0.072616404	
14	11	29	20	91	0.119440241	
15	6	8	10	59	0.038865755	
16	20	34	16	164	0.077364465	
17	19	21	14	97	0.177880988	
18	1	13	2	67	0.004912098	
	CellProportions.12	CellProportions.21	CellProportions.22			

1	0.14624540	0.132353799	0.5332953
2	0.14193999	0.151129239	0.3418269
3	0.14935088	0.118811130	0.6714009
4	0.15903087	0.150442929	0.5152062
5	0.18726135	0.090886158	0.4020700
6	0.13866831	0.193401056	0.5968110
7	0.25723522	0.061027730	0.6046214
8	0.34627289	0.143826029	0.3491301
9	0.19301261	0.001305557	0.7889065
10	0.07551795	0.223070334	0.6051074
11	0.12432343	0.321854037	0.3604095
12	0.04033894	0.151866985	0.7814858
13	0.09920585	0.253666542	0.5745112
14	0.19829618	0.292048553	0.3902150
15	0.02778149	0.226000766	0.7073520
16	0.09445779	0.059403274	0.7687745
17	0.13985543	0.125327293	0.5569363
18	0.06173515	0.011885203	0.9214676

>

The `write.csv` function is used to store the relative risk estimates as a comma-separated value (csv) file. Files in csv format can be read by programs such as Microsoft Excel.

```
> write.csv(RelRisk_Estimates, file="RelRisk_Estimates.csv")
```

5 Attributable risk analysis

The `attrisk.analysis` function will be used to calculate attributable risk estimates. The four data frames used to calculate relative risk estimates can be used for attributable risk estimation. Arguments for the `attrisk.analysis` function are identical to arguments for the `relrisk.analysis` function

Calculate attributable risk estimates.

```
> AttRisk_Estimates <- attrisk.analysis(sites, subpop, design, data.risk,
+   response.var= rep(resp_vars, each=length(stress_vars)),
+   stressor.var=rep(stress_vars, length(resp_vars)))
```

The attributable risk estimates are displayed using the `print` function. The object produced by `attrisk.analysis` is a data frame containing nineteen columns. The data data frame is identical to the one produced by the `relrisk.analysis` function except that it doesn't include the columns named `Estimate.num` and `Estimate.denom`. Since attributable risk is not calculated using a ratio estimator, values for numerator and denominator estimates are not relevant.

```
> # Print the attributable risk estimates
> print(AttRisk_Estimates)
```

	Type	Subpopulation	Response	Stressor	NResp	Estimate
1	Western_US	Western_US	Chlorophyll_a	Total_Nitrogen	236	0.35632857
2	Lake_Origin	Man-Made	Chlorophyll_a	Total_Nitrogen	152	0.42134049
3	Lake_Origin	Natural	Chlorophyll_a	Total_Nitrogen	84	0.13260811
4	Western_US	Western_US	Chlorophyll_a	Total_Phosphorus	236	0.29454985
5	Lake_Origin	Man-Made	Chlorophyll_a	Total_Phosphorus	152	0.37332396
6	Lake_Origin	Natural	Chlorophyll_a	Total_Phosphorus	84	0.10127632
7	Western_US	Western_US	Chlorophyll_a	Turbidity	236	0.10732593
8	Lake_Origin	Man-Made	Chlorophyll_a	Turbidity	152	0.01794367
9	Lake_Origin	Natural	Chlorophyll_a	Turbidity	84	0.06302202
10	Western_US	Western_US	MacroInvert_OE	Total_Nitrogen	234	0.35425260
11	Lake_Origin	Man-Made	MacroInvert_OE	Total_Nitrogen	151	0.19279580
12	Lake_Origin	Natural	MacroInvert_OE	Total_Nitrogen	83	0.26351638
13	Western_US	Western_US	MacroInvert_OE	Total_Phosphorus	234	0.14300096
14	Lake_Origin	Man-Made	MacroInvert_OE	Total_Phosphorus	151	-0.06045589
15	Lake_Origin	Natural	MacroInvert_OE	Total_Phosphorus	83	0.43296859
16	Western_US	Western_US	MacroInvert_OE	Turbidity	234	0.36315945
17	Lake_Origin	Man-Made	MacroInvert_OE	Turbidity	151	0.36830221
18	Lake_Origin	Natural	MacroInvert_OE	Turbidity	83	0.05787783
	StdError.log	LCB95Pct	UCB95Pct	WeightTotal	CellCounts.11	CellCounts.12
1	0.25520134	-0.061431594	0.6096659	4890.777	41	33
2	0.23787429	0.077636614	0.6369686	2049.445	29	20
3	0.20270096	-0.290490261	0.4169900	2841.333	12	13
4	0.23113933	-0.109717615	0.5515437	4890.777	39	35
5	0.20990967	0.054375912	0.5846945	2049.445	26	23
6	0.24370434	-0.448997696	0.4425773	2841.333	13	12
7	0.09551704	-0.076458205	0.2597325	4890.777	22	52
8	0.12710745	-0.259883685	0.2345050	2049.445	19	30
9	0.05046779	-0.034397925	0.1512669	2841.333	3	22
10	0.22734361	-0.008272434	0.5864315	4882.983	28	26
11	0.27828602	-0.392710659	0.5321508	2045.360	23	17
12	0.22902173	-0.153736794	0.5298684	2837.623	5	9
13	0.20135161	-0.271660373	0.4224501	4882.983	17	37
14	0.23045317	-0.665922146	0.3249584	2045.360	11	29
15	0.34363056	-0.112003067	0.7108599	2837.623	6	8
16	0.17937230	0.094866404	0.5519270	4882.983	20	34
17	0.23493409	-0.001118809	0.6014039	2045.360	19	21
18	0.06577750	-0.071758993	0.1718342	2837.623	1	13
	CellCounts.21	CellCounts.22	CellProportions.11	CellProportions.12		
1	22	140	0.188105511	0.14624540		
2	17	86	0.365103884	0.14193999		
3	5	54	0.060437097	0.14935088		

4	8	154	0.175320038	0.15903087
5	5	98	0.319782522	0.18726135
6	3	56	0.071119667	0.13866831
7	16	146	0.077115692	0.25723522
8	15	88	0.160770986	0.34627289
9	1	58	0.016775371	0.19301261
10	33	147	0.096304303	0.07551795
11	22	89	0.193412993	0.12432343
12	11	58	0.026308304	0.04033894
13	30	150	0.072616404	0.09920585
14	20	91	0.119440241	0.19829618
15	10	59	0.038865755	0.02778149
16	16	164	0.077364465	0.09445779
17	14	97	0.177880988	0.13985543
18	2	67	0.004912098	0.06173515

CellProportions.21 CellProportions.22

1	0.132353799	0.5332953
2	0.151129239	0.3418269
3	0.118811130	0.6714009
4	0.150442929	0.5152062
5	0.090886158	0.4020700
6	0.193401056	0.5968110
7	0.061027730	0.6046214
8	0.143826029	0.3491301
9	0.001305557	0.7889065
10	0.223070334	0.6051074
11	0.321854037	0.3604095
12	0.151866985	0.7814858
13	0.253666542	0.5745112
14	0.292048553	0.3902150
15	0.226000766	0.7073520
16	0.059403274	0.7687745
17	0.125327293	0.5569363
18	0.011885203	0.9214676

>

The write.csv function is used to store the attributable risk estimates as a csv file.

```
> write.csv(AttRisk_Estimates, file="AttRisk_Estimates.csv")
```

References

- U.S. Environmental Protection Agency (2009). National Lakes Assessment: A collaborative survey of the nations lakes. Technical report, U.S. Environmental Protection Agency, Office of Water and Office of Research and Development. EPA 841-R-09-001.
- Van Sickle, J. and S. G. Paulsen (2008). Assessing the attributable risks, relative risks, and regional extents of aquatic stressors. *Journal of the North American Benthological Society* 27, 920–931.
- Van Sickle, J., J. L. Stoddard, S. G. Paulsen, and A. R. Olsen (2006). Using relative risk to compare the effects of aquatic stressors at a regional scale. *Environmental Management* 38, 1020–1030.