SOFTWARE FOR TABULAR DATA PROTECTION

Joe Fred Gonzalez, Jr. and Lawrence H. Cox, National Center for Health Statistics
Joe Fred Gonzalez, Jr., NCHS, 6525 Belcrest Road, Room 915, Hyattsville, MD 20782

KEY WORDS: Statistical disclosure, disclosure limitation

1. Introduction

A major responsibility of the National Center for Health Statistics (NCHS) is the protection of identifiable data collected from survey respondents, persons or establishments. Prior to release of public use files, data that could be used to identify a respondent are perturbed or removed from microdata files. The other mechanism for statistical disclosure is the possible identification of individuals or establishments via tabular data. The National Center for Health Statistics has sponsored the development of disclosure limitation software for two-way tables by OptTek Systems, Inc. This paper will describe features of the software including its different functions: cell suppression, controlled rounding, unbiased controlled rounding, and controlled rounding subject to subtotal constraints.

2. Software Functions

As mentioned in the introduction, this paper will describe features of the software, including four different functions for tabular disclosure limitation. The four sections that follow will describe each of the functions in greater detail. Figure 1 provides an image of the data protection utility screen, and Table 1 contains an image of an array with 20 rows and 10 columns whose cell entries were randomly generated by the software. The four disclosure limitation functions were individually applied to this original data table. A resulting table will be presented for each of the four disclosure limitation function outputs.

2.1 Cell Suppression Function

A multiple-cell suppression technique by Cox [1] is used as the cell suppression function in the Software for Tabular Data Protection (STDP).

Cell suppression hides from publication the values of all cells representing direct disclosure of confidential data on individual respondents (the *disclosure cells*), together with a sufficient number of appropriately selected nondisclosure cells (the *complementary cells*) to ensure that a third party cannot reconstruct or narrowly estimate confidential respondent data by manipulating linear relationships between released and suppressed table values.

The challenge of this cell suppression problem is to select complementary cells that provide sufficient disclosure protection while minimizing the amount of information lost due to suppression. The cell suppression approach used is based on mathematical networks which offer theoretical and practical advantages. A mathematical network is a specialized linear program defined over a mathematical graph.

Table 2 displays the results of applying the *cell suppression* function. The counts in cells (7,3) and (9,7) are the primary suppressions, and the counts in cells (7,2), (7,7), (9,2), (13, 2), and (13, 3) are the complementary suppressions. On a computer screen, the primary suppressions would be highlighted in blue, and the complementary suppressions would be highlighted in red.

2.2 Controlled Rounding Function

The *controlled rounding* function that is used in the STDP is based on the methodology described by Cox and Ernst [2] and by Causey, Cox, and Ernst [3].

The *controlled rounding* function rounds all entries in a one or two-way tabular array A to integer multiples of a positive integer base B subject to the following requirements:

> (1) each entry in A is rounded to an *adjacent integer multiple of B*; that is, an entry a is rounded to either $B[a/B]$ *or* $B([a/B] + 1)$, where [ ] is the greatest integer function, and
>
> (2) the sum of the rounded values for any row (or column) of A equals the rounded value of the corresponding row (or column) total entry.

Requirements (1) and (2) are referred to as controlled rounding of an array A.

Additionally, optimal controlled roundings were achieved by presenting this problem as a capacitated transportation problem whose objective function is minimized with respect to the l_p norm, $1 \# p < 4$, where the objective function is the p^{th} root of the sum of the p^{th} powers of the absolute values of the differences between rounded and unrounded entries of A. That is, the objective function to minimize with respect to l_p norm is:

$$l_p[R(A),\ A] \ ' \ (\sum_{i=1}^{m}\sum_{j=1}^{n} *R(a_{ij}) \& a_{ij}^{*p})^{1/p}.$$

2.2.1 Test Results for Controlled Rounding Function

- Testing was done on a Pentium 4 processor with 261,200 KB of Ram
- Total time to solve the problem is dependent on: number of cells in table; number of rows; and number of columns.
- A table with 50 rows and 50 columns was rounded in less than a minute.
- A table with 100 rows and 100 columns was rounded in 24 minutes.
- A table with 1000 rows and 5 columns was rounded in 1 hour and 40 minutes.

Table 3 displays the results of applying the *controlled rounding* function (to base 5 with power = p =2) .

2.3 Unbiased Controlled Rounding Function

The *unbiased controlled rounding* function that is used in the STDP is based on the methodology described by Cox [4].

First, we assume that we have a two-way table A that is additive, that is, entries sum along rows and columns to all corresponding totals entries.

The objective is to construct a second additive table R(A) whose internal and totals entries, denoted by R(a), are integer multiples of B that are adjacent to the corresponding entries of A, that is, R(a) = B[a/B] or B([a/B]+1), where [a/B] denotes the integer part of a/B. Therefore, the conditions for unbiased controlled rounding are that every entry a of A satisfies the following:

1. R(a) = B[a/B] or B([a/B]+1)
2. R(a) is additive
3. *R(a) - a * < B
4. E(R(a)) = a.

2.3.1 Test Results for Unbiased Controlled Rounding

- A table with 50 rows and 50 columns was rounded in a second.
- A table with 100 rows and 100 columns was rounded in 4 seconds.
- A table with 400 rows and 25 columns was rounded in 5 seconds.
- A table with 2000 rows and 25 columns was rounded in 5 minutes and 45 seconds.

Table 4 displays the results of applying the *unbiased controlled rounding* function (to base 5 with power =p =2).

2.4 Controlled Rounding Subject to Subtotal Constraints

The *controlled rounding subject to subtotal constraints* function that is used in the STDP is based on the methodology described by Cox and George [5]. The methodology used in this function is similar to that used for *controlled rounding* as discussed earlier. Recall that controlled rounding for a two-way table was presented as a capacitated transportation problem. This function extends that methodology to tables with subtotals along one, but not both, dimensions.

Table 5 displays the results of applying the *controlled rounding subject to subtotal constraints* (to base 5 with power = p = 2) function.

3. Future Research and Development

As mentioned earlier, the software developed for this project is a tool which features some of the different mathematical functions for protecting potential disclosure cell values in two-way tables. The ultimate goal of this project is to develop production level software that can be embedded into NCHS data analysis activities, for example, the NCHS Research Data Center (RDC).

References

1. Cox, L.H. (1995). Network models for complementary cell suppression. *Journal of the American Statistical Association* **90**, 1453-1462.

2. Cox, L.H. and L.R. Ernst (1982). Controlled rounding. *INFOR* **20**, 423-432.

3. Causey, B.D., L.H. Cox, and L.R. Ernst (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical*

Association **80**, 903-909.

4. Cox, L.H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association* **82**, 520-524.

5. Cox, L.H. and J. A. George (1989). Controlled rounding for tables with subtotals. *Annals of Operations Research* **20**, 141-157.

Figure 1. CDC Data Protection Utility Screen

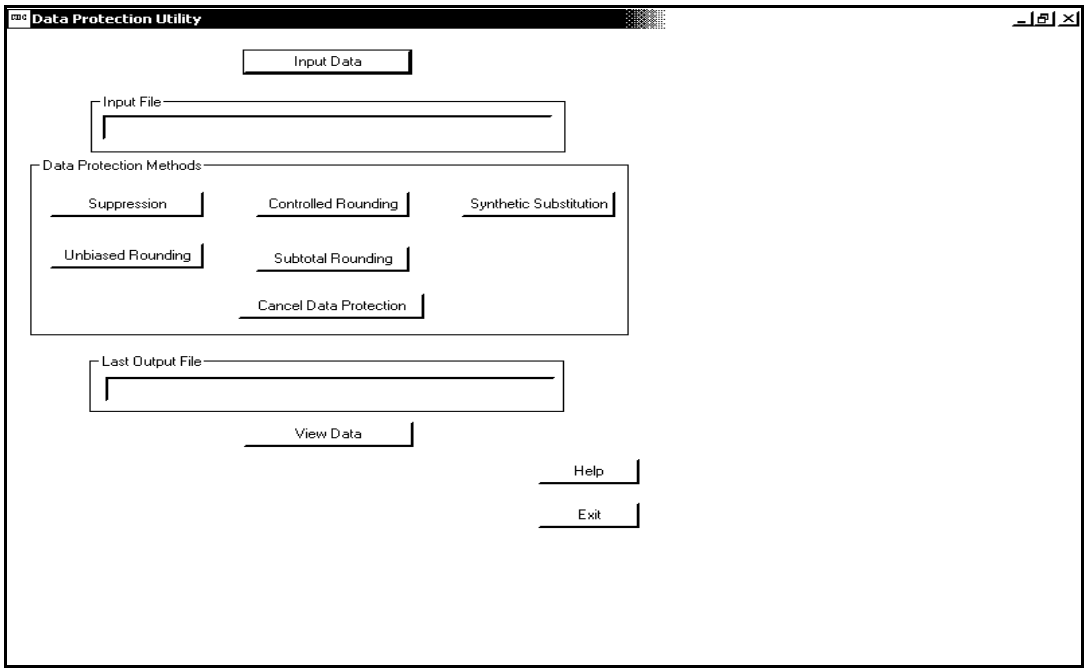
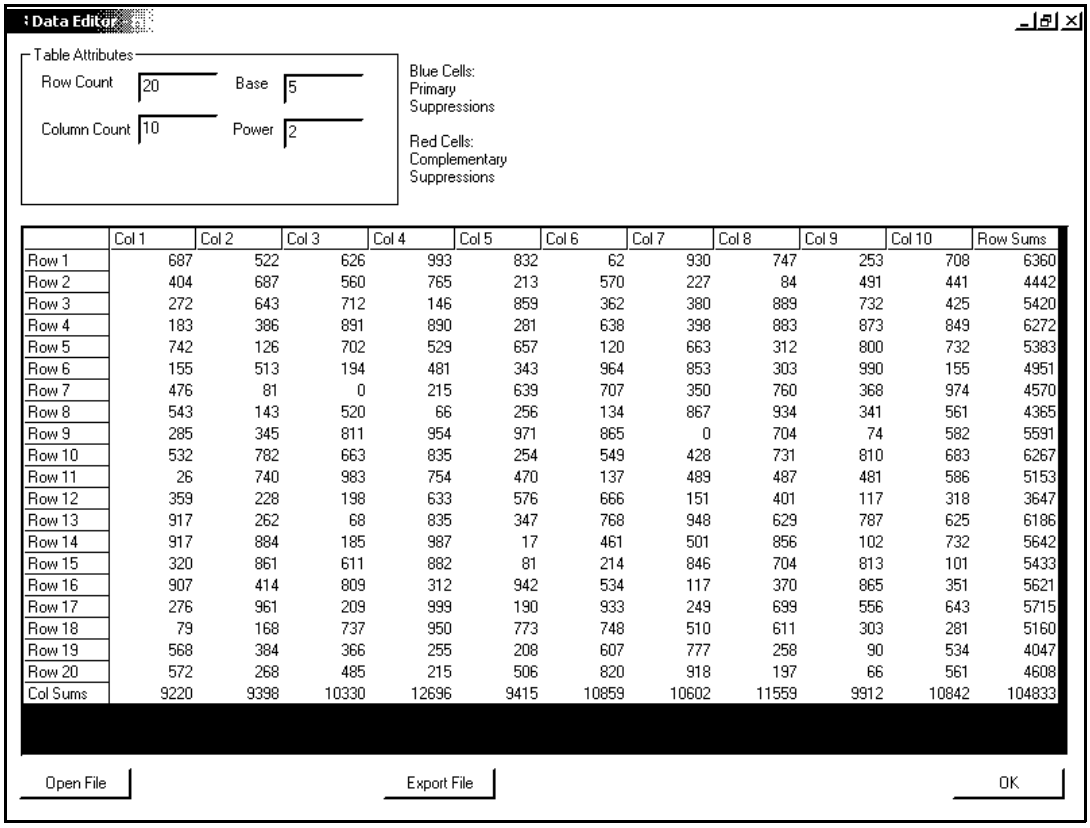
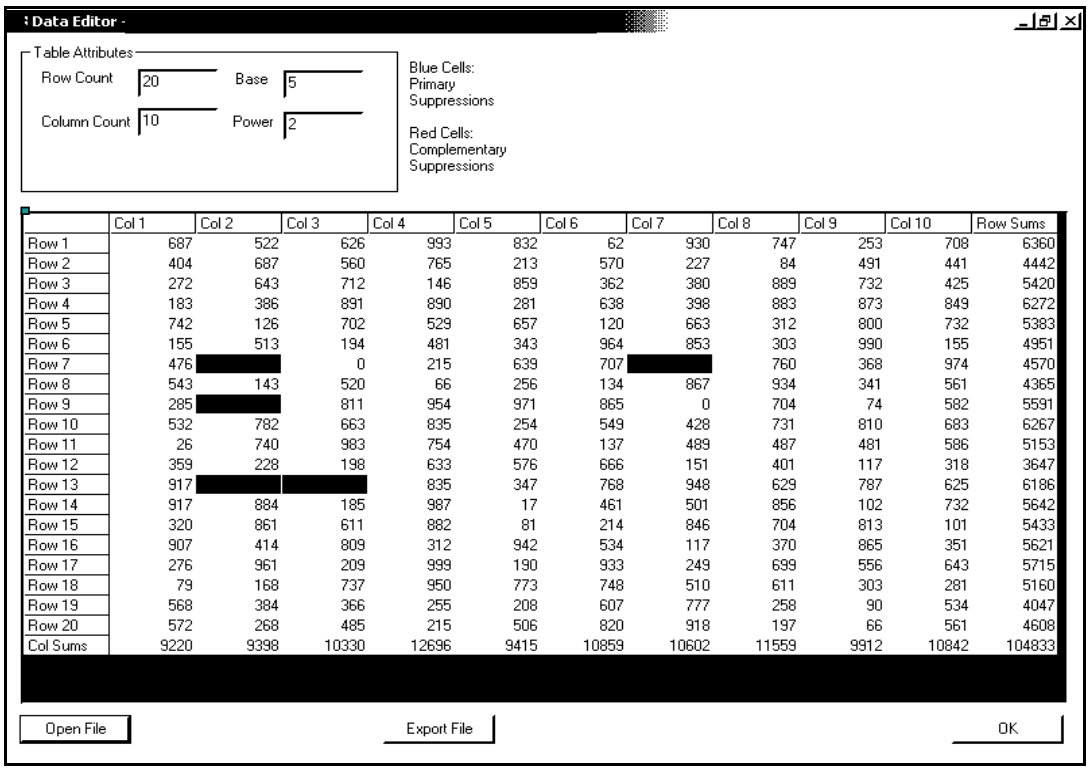


Table 1. Original 20 rows by 10 columns.



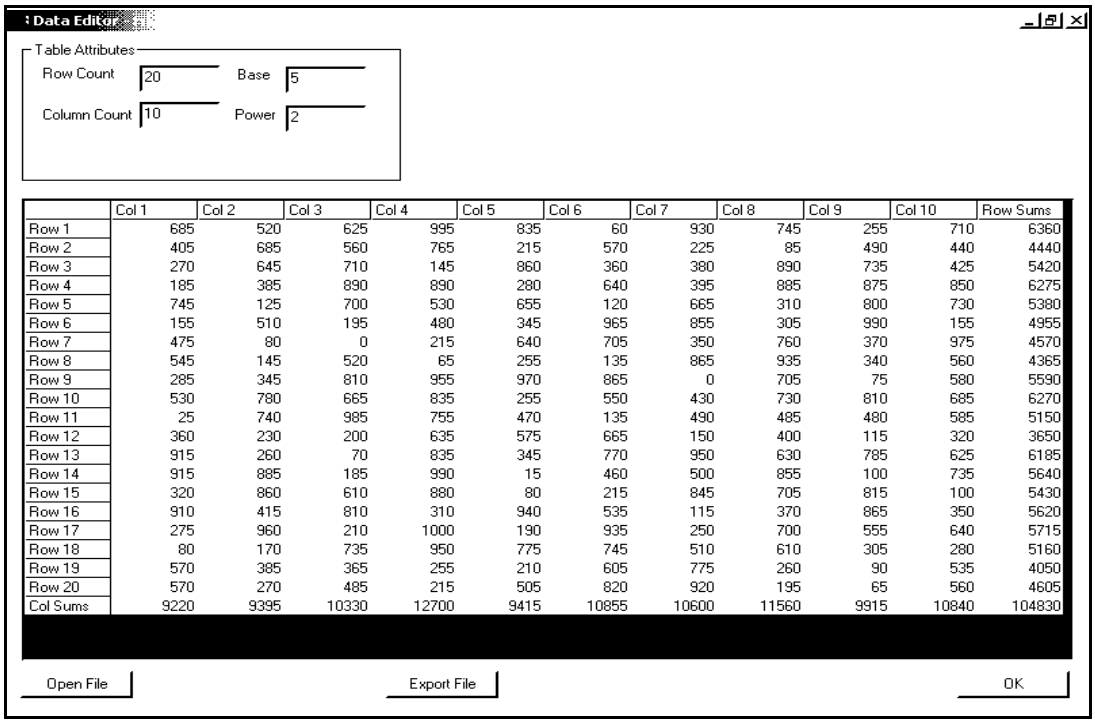
	Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9	Col 10	Row Sums
Row 1	687	522	626	993	832	62	930	747	253	708	6360
Row 2	404	687	560	765	213	570	227	84	491	441	4442
Row 3	272	643	712	146	859	362	380	889	732	425	5420
Row 4	183	386	891	890	281	638	398	883	873	849	6272
Row 5	742	126	702	529	657	120	663	312	800	732	5383
Row 6	155	513	194	481	343	964	853	303	990	155	4951
Row 7	476	81	0	215	639	707	350	760	368	974	4570
Row 8	543	143	520	66	256	134	867	934	341	561	4365
Row 9	285	345	811	954	971	865	0	704	74	582	5591
Row 10	532	782	663	835	254	549	428	731	810	683	6267
Row 11	26	740	983	754	470	137	489	487	481	586	5153
Row 12	359	228	198	633	576	666	151	401	117	318	3647
Row 13	917	262	68	835	347	768	948	629	787	625	6186
Row 14	917	884	185	987	17	461	501	856	102	732	5642
Row 15	320	861	611	882	81	214	846	704	813	101	5433
Row 16	907	414	809	312	942	534	117	370	865	351	5621
Row 17	276	961	209	999	190	933	249	699	556	643	5715
Row 18	79	168	737	950	773	748	510	611	303	281	5160
Row 19	568	384	366	255	208	607	777	258	90	534	4047
Row 20	572	268	485	215	506	820	918	197	66	561	4608
Col Sums	9220	9398	10330	12696	9415	10859	10602	11559	9912	10842	104833

Table 2. Results of Cell Suppression.



	Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9	Col 10	Row Sums
Row 1	687	522	626	993	832	62	930	747	253	708	6360
Row 2	404	687	560	765	213	570	227	84	491	441	4442
Row 3	272	643	712	146	859	362	380	889	732	425	5420
Row 4	183	386	891	890	281	638	398	883	873	849	6272
Row 5	742	126	702	529	657	120	663	312	800	732	5383
Row 6	155	513	194	481	343	964	853	303	990	155	4951
Row 7	476		0	215	639	707		760	368	974	4570
Row 8	543	143	520	66	256	134	867	934	341	561	4365
Row 9	285		811	954	971	865	0	704	74	582	5591
Row 10	532	782	663	835	254	549	428	731	810	683	6267
Row 11	26	740	983	754	470	137	489	487	481	586	5153
Row 12	359	228	198	633	576	666	151	401	117	318	3647
Row 13	917			835	347	768	948	629	787	625	6186
Row 14	917	884	185	987	17	461	501	856	102	732	5642
Row 15	320	861	611	882	81	214	846	704	813	101	5433
Row 16	907	414	809	312	942	534	117	370	865	351	5621
Row 17	276	961	209	999	190	933	249	699	556	643	5715
Row 18	79	168	737	950	773	748	510	611	303	281	5160
Row 19	568	384	366	255	208	607	777	258	90	534	4047
Row 20	572	268	485	215	506	820	918	197	66	561	4608
Col Sums	9220	9398	10330	12696	9415	10859	10602	11559	9912	10842	104833

Table 3. Results of Controlled Rounding.



	Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9	Col 10	Row Sums
Row 1	685	520	625	995	835	60	930	745	255	710	6360
Row 2	405	685	560	765	215	570	225	85	490	440	4440
Row 3	270	645	710	145	860	360	380	890	735	425	5420
Row 4	185	385	890	890	280	640	395	885	875	850	6275
Row 5	745	125	700	530	655	120	665	310	800	730	5380
Row 6	155	510	195	480	345	965	855	305	990	155	4955
Row 7	475	80	0	215	640	705	350	760	370	975	4570
Row 8	545	145	520	65	255	135	865	935	340	560	4365
Row 9	285	345	810	955	970	865	0	705	75	580	5590
Row 10	530	780	665	835	255	550	430	730	810	685	6270
Row 11	25	740	985	755	470	135	490	485	480	585	5150
Row 12	360	230	200	635	575	665	150	400	115	320	3650
Row 13	915	260	70	835	345	770	950	630	785	625	6185
Row 14	915	885	185	990	15	460	500	855	100	735	5640
Row 15	320	860	610	880	80	215	845	705	815	100	5430
Row 16	910	415	810	310	940	535	115	370	865	350	5620
Row 17	275	960	210	1000	190	935	250	700	555	640	5715
Row 18	80	170	735	950	775	745	510	610	305	280	5160
Row 19	570	385	365	255	210	605	775	260	90	535	4050
Row 20	570	270	485	215	505	820	920	195	65	560	4605
Col Sums	9220	9395	10330	12700	9415	10855	10600	11560	9915	10840	104830

Table 4. Results of Unbiased Controlled Rounding

.

Data Editor

Table Attributes

Row Count 20 Base 5

Column Count 10 Power 2

	Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9	Col 10	Row Sums
Row 1	685	525	625	995	830	60	930	750	255	705	6360
Row 2	400	685	560	765	215	570	230	80	490	445	4440
Row 3	270	645	710	145	860	365	380	890	730	425	5420
Row 4	180	390	895	890	280	635	395	885	875	850	6275
Row 5	745	125	700	525	660	120	665	315	800	730	5385
Row 6	155	515	190	480	340	965	855	305	990	155	4950
Row 7	475	80	0	215	640	710	350	760	365	975	4570
Row 8	545	140	520	65	255	135	870	935	340	560	4365
Row 9	285	345	810	955	970	865	0	705	75	585	5595
Row 10	530	785	665	835	255	545	430	730	810	680	6265
Row 11	25	740	985	755	470	140	485	485	485	585	5155
Row 12	360	225	195	635	580	665	150	400	115	320	3645
Row 13	920	260	65	835	350	770	945	630	785	625	6185
Row 14	915	885	185	990	15	460	505	855	100	730	5640
Row 15	320	865	610	880	80	215	845	705	815	100	5435
Row 16	910	410	810	310	945	535	115	370	865	350	5620
Row 17	280	960	210	1000	190	930	250	695	555	645	5715
Row 18	80	165	740	950	770	750	510	610	305	280	5160
Row 19	570	385	370	255	205	605	775	260	90	535	4050
Row 20	570	270	485	215	505	820	915	195	65	565	4605
Col Sums	9220	9400	10330	12695	9415	10860	10600	11560	9910	10845	104835

Open File Export File OK

Table 5. Results of Controlled Rounding Subject to Subtotal Constraints.

Data Editor -

Table Attributes

Row Count 20 Base 5

Column Count 10 Power 2

	Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9	Col 10	Row Sums
Row 1	687	522	626	993	832	62	930	747	253	708	6360
Row 2	404	687	560	765	213	570	227	84	491	441	4442
Row 3	272	643	712	146	859	362	380	889	732	425	5420
Row 4	183	386	891	890	281	638	398	883	873	849	6272
Row 5	742	126	702	529	657	120	663	312	800	732	5383
Row 6	155	513	194	481	343	964	853	303	990	155	4951
Row 7	476	85	0	215	635	707	350	760	368	974	4570
Row 8	543	145	520	65	255	134	867	934	341	561	4365
Row 9	285	345	811	954	971	865	0	704	74	582	5591
Row 10	532	782	663	835	254	549	428	731	810	683	6267
Row 11	26	740	983	754	470	137	489	487	481	586	5153
Row 12	359	228	198	633	576	666	151	401	117	318	3647
Row 13	917	262	68	835	347	768	948	629	787	625	6186
Row 14	917	884	185	987	17	461	501	856	102	732	5642
Row 15	320	861	611	882	81	214	846	704	813	101	5433
Row 16	907	414	809	312	942	534	117	370	865	351	5621
Row 17	276	961	209	999	190	933	249	699	556	643	5715
Row 18	79	168	737	950	773	748	510	611	303	281	5160
Row 19	568	384	366	255	208	607	777	258	90	534	4047
Row 20	572	268	485	215	506	820	918	197	66	561	4608
Col Sums	9220	9404	10330	12695	9410	10859	10602	11559	9912	10842	104833

Open File Export File OK