

# Applicability of Latin Hypercube Sampling to Create Multi Variate Synthetic Micro Data

Ramesh A. DANDEKAR<sup>1</sup>, Michael COHEN<sup>2</sup> and Nancy KIRKENDALL<sup>1</sup>  
*1Energy Information Administration, U.S. Department of Energy, Washington DC*  
*2National Academy of Science, Washington DC*  
*(Ramesh.Dandekar@EIA.DOE.GOV)*

**Abstract:** We use Latin Hypercube Sampling to produce a synthetic data set that reproduces many of the essential features of an original data set while providing disclosure protection. We demonstrate that this procedure is feasible for large data sets. We believe this technique should be included in the current collection of approaches to the problem of disclosure avoidance for large data sets, which currently includes top coding, smearing, and swapping. The use of Latin Hypercube Sampling, along with the goal of reproducing the rank correlation structure instead of the Pearson correlation structure, is relatively unique and given its properties is expected to offer advantages for certain applications and data sets. We have made the first steps in understanding the relative advantages and disadvantages of this approach to disclosure protection, but more work remains.

Keywords: Statistical Disclosure Limitation, Confidentiality, Latin Hypercube Sampling, Rank Correlation

## 1. Introduction

The utility of public-use microdata, generated either synthetically or through use of other techniques, is assessed by how well inferences based on this modified data set mimic those derived through analysis of the real data. This assessment naturally includes univariate as well as multivariate analyses. We use Latin Hypercube Sampling of McKay et al. (1979) along with the rank correlation refinement of Iman and Conover (1982) to reproduce both the univariate and multivariate structure (in the sense of rank correlation) of the original data set. In doing this, much of the univariate and multivariate structure of the real data is retained. At the same time, the data is sufficiently altered to provide considerable protection from disclosure.

In the following, we first describe the basic procedure. We then discuss some refinements including smoothing of the empirical cumulative distribution function and application of

the technique to subpopulations. We also present a real data application that helps to suggest some general advantages of the approach.

## 2. LHS and the Restricted Pairing Algorithm

Univariate statistics of interest are generally simple functions of the first few moments such as the mean, standard deviation, and coefficient of skewness. While interest usually resides in these specific aspects of a distribution, a comprehensive picture of the univariate properties of any variable can be represented by the cumulative distribution function (cdf). Latin Hypercube Sampling (LHS), developed by McKay, et al.(1979), can be used to generate a synthetic data set for a group of uncorrelated variables in which the univariate characteristics of the original data are reproduced almost exactly. In the case where the variables are not uncorrelated, the restricted pairing algorithm of Iman and Conover (1982), which makes use of LHS, has the goal of producing a synthetic data set that reproduces the rank correlation structure of the real data, which has advantages and disadvantages in comparison with procedures that attempt to reproduce the Pearson correlation structure. One benefit of rank correlation is that for nonnormal data, the rank correlation is a more useful summary of the relatedness of two variables that are monotonically but not linearly related. This approach would also have an advantage for heavily skewed distributions, for which the Pearson correlation can be dominated by a small percentage of the data. Finally, while the full procedure that we are advancing is slightly oriented to treatment of continuous variables, the treatment of discrete or categorical data is straightforward and is discussed below

The basic LHS procedure, in combination with the restricted pairing algorithm, proceeds as follows. Let  $\mathbf{V}$  be an  $n$  by  $m$  data matrix containing  $n$  samples of  $\{i/(n+1)\}$ ,  $i = 1, \dots, n$ , for  $m$  different variables,  $j = 1, \dots, m$ . Thus, initially  $\mathbf{V}$  consists of  $m$  identical columns. To induce a desired rank correlation matrix for the variables in  $\mathbf{V}$ , proceed as follows:

1. Let  $\mathbf{T}$  be the  $m$  by  $m$  target rank correlation matrix we wish to impose on data matrix  $\mathbf{V}$ .
2. Shuffle the columns of matrix  $\mathbf{V}$  to remove perfect correlation to get  $\mathbf{V}^*$ .<sup>1</sup>
3. Let  $\mathbf{C}$  be the  $m$  by  $m$  correlation matrix for data matrix  $\mathbf{V}^*$  imposed by the shuffling.
4. Compute the  $m$  by  $m$  lower triangular matrix  $\mathbf{Q}$ , such that  $\mathbf{Q}\mathbf{Q}' = \mathbf{C}$ , where ' denotes transpose.
5. Compute the  $m$  by  $m$  matrix  $\mathbf{P}$ , such that  $\mathbf{P}\mathbf{P}' = \mathbf{T}$ .

---

<sup>1</sup>The algorithm yields slightly different rank correlation matrices depending on the starting values of the shuffled matrix  $\mathbf{V}$ .

6. Compute the  $m$  by  $m$  matrix  $S$ , such that  $S = PQ_{inv}$ , where  $Q_{inv}$  is the inverse of matrix  $Q$ . (This creates  $S$  which operates by “removing” the correlation structure of  $C$  and instituting the correlation structure of  $T$ .)
7. Compute the  $n$  by  $m$  matrix  $R$ , such that  $R = (V^*)S'$ . (Note that  $R$  has a correlation structure that approximates  $T$ .)
8. By columns, replace the values of  $R$  with their corresponding ranks ranging from  $1$  to  $n$ .
9. By columns, arrange the values of matrix  $V^*$  in the same rank sequence as the corresponding columns of  $R$ , to arrive at  $V^{**}$ .
10. Finally, for each column, replace each rank by the inverse cdf for that variable evaluated at  $i/(n+1)$  to arrive at the modified data matrix.

The resultant matrix has a rank correlation structure that resembles  $T$ .

### **3. Iterative Refinement**

For a population with a large number of observations and/or a large number of variables, it is essential to mimic the multivariate covariance structure to the extent possible. With a large number of observations, correlations are known to great accuracy, and if the procedure introduces a comparable amount of error, the value of the synthetic data is reduced in comparison to use of the original data set. With a large number of variables, the procedure as outlined above tends to get further away from the target correlation matrix for individual correlations. Since microdata users could potentially be interested in any subset of the variables, reproducing each of the rank correlations gains importance. Iterative refinement of the rank correlation matrix, developed by Dandekar (1993), achieves that objective by reducing the gap between the rank correlations of the actual and the synthetic data.

Iterative refinement of the rank correlation matrix is based on the fact that the original algorithm yields slightly different results depending on the starting values of the shuffled matrix  $V^*$ . Iterative refinement is implemented as follows:

- Perform all the steps of the restricted pairing algorithm as described above.
- Let  $C^*$  be the revised  $m$  by  $m$  correlation matrix of data matrix  $V^{**}$  in the procedure above.
- Repeat steps 4 through 9 of the restricted pairing algorithm described above until the Mean Absolute Deviation (**MAD**) between the off-diagonal elements of  $T$  and updates of  $C^*$  cannot be reduced further within an acceptable tolerance level.

It has been observed that highly correlated variables tend to benefit the most from this iterative refinement procedure.

### **4. Choice of Cumulative Distribution Function to Support LHS**

The cumulative distribution function (cdf) required to generate synthetic variables (step 10 above ) could be constructed using either a fitted theoretical distribution or by using the observed empirical cumulative distribution function. The empirical cdf has the advantage of making no assumptions. However, it provides little information for the extreme tails of the distribution. Fitted theoretical cdfs are dependent on the assumed distributional family, but if the assumption is relatively correct, the resulting cdf might produce better values in the tails of the distribution.

To better control the univariate statistical properties of the synthetic data, we propose using the empirical cdf for each of the population variables. Constructing the empirical cdf is straight-forward. We wish to be flexible in the algorithm to produce synthetic data sets of a different size than the original data set. Therefore, if we have  $n$  rows for the original data set, we may be interested in producing a synthetic data set with  $N$  rows. Particularly if  $N > n$ , some extrapolation will be needed in using the empirical cdf.

We make one other modification to the empirical cdf. To protect the true values for extreme outliers in the database, we smooth the tails of the distribution as follows:

- Choose  $p$ ,  $0 < p < 1$ , and choose some tolerance  $T > 0$ .
- If  $|Y_{(j)} - Y_{(j-1)}| / Y_{(j)} > T$ , replace any order statistic  $Y_{(j)}$  with  $pY_{(j)} + (1-p) Y_{(j-1)}$

## 5. Treatment of Identifiable Subpopulations

When categorical variables identify subgroups that have clearly distinct distributions, either distinct rank correlations or univariate characteristics that distinguish these groups, a synthetic data set that can retain this structure will be more useful to the analyst. To accomplish this, a separate LHS sample can be drawn for subgroups that have specific values for combinations of these variables. In our example described below, the classification variables that might be examined for this include industry type, geological descriptors, income categories and many other categorical variables. By separately applying the procedure described here to subgroups, each subgroup in the synthetic data set can have its own statistical characteristics, since the LHS-based synthetic data will reproduce the statistical characteristics of each one of the subgroups identified in the data.

To be able to measure the rank correlations of each subgroup of the population, it is essential that each subgroup contain enough observations. When this basic condition is not satisfied, the subgroup must be merged into a larger unit by combining it with other closely related subgroup(s) until the rank correlation structure of the resultant subgroup can be determined to a needed degree of precision. Depending upon the number of observations, the subgroup collapsing procedure may need to be repeated over different combinations of classification variables until there are enough observations to determine the correlations.

When the subgroups are not identifiable, a synthetic data set can be generated by drawing a single LHS-based sample to represent the entire population. In this case, the categorical

variables are simply additional variables which have a discrete univariate distribution where the cdf is a step function.

## **6. Size of the Synthetic Data Set**

An additional useful feature of the LHS procedure proposed here is that it permits flexibility in the size of the synthetic data set that is generated. This is an important advantage, since it can be used to provide additional protection against the disclosure of outliers. These so-called “population uniques” generally appear at the tail end of the distributions, and can be used to identify specific members of the data set. By generating a synthetic data set several times larger than the original, and given the smoothing mentioned previously, the individual outlier becomes one of several observations that have moderately different values. The larger sample size also permits the distribution of the ‘unique’ original observation over different combinations of classification variables included in a subset of the data. This feature would be especially helpful for protection for establishment surveys. Such an approach prevents one to one comparisons, and at the same time helps retain the closeness to the overall statistical characteristics of the original data.

The challenge with simulating a larger sample size than we have observed is that in theory there should be an increased chance of selecting an observation larger than the maximum of the observed sample. When the empirical sample cdf is used as the basis of the LHS sample, the maximum in the observed sample will determine the maximum in the simulated sample. How best to simulate observations in the tails of the distribution is a topic for future research.

## **7. An Illustrative Example<sup>2</sup>**

To demonstrate the LHS-based synthetic data generation procedure, we have used 13 variables from the Commercial Building Energy Consumption Survey (CBECS) carried out by the Energy Information Administration (EIA) of the U. S. Department of Energy.<sup>3</sup> Our test data set consists of two categorical variables (PBA and YRCON) and 11 continuous variables. The variables selected for the analysis are: (1) electricity consumption – ELBTU, (2) electricity expenditures – ELEXP, (3) natural gas consumption – NGBTU, (4) natural gas expenditures – NGEXP, (5) major fuel consumption – MFBTU, (6) major fuel expenditures – MFEXP, (7) principal building activity -- PBA, (8) year constructed – YRCON, (9) total floor space – SQFT, (10) number

---

<sup>2</sup> The following example involves data from a sample survey. While these typically have sample weights determined by a sample design that might involve stratification or clustering, for the current paper we are restricting our attention to surveys in which each data case has identical sampling weight. We hope to address this complication in future work.

<sup>3</sup> The Commercial Buildings Energy Consumption Survey is a national-level sample survey of commercial buildings and their energy suppliers conducted quadrennially by the Energy Information Administration.

of employees – NWKER, (11) percent of floor space cooled – COOLP, (12) percent of floor space heated – HEATP, and (13) percent of floor space lit – LTOHRP.

Our test population consists of 5655 observations. The rank correlation matrix is as shown:

Rank Correlation Matrix:

1.0000	.9782	.3658	.3715	.9273	.9627	-.0930	.2313	.8143	.7650	.3341	.2236	.2535
.9782	1.0000	.3652	.3755	.9113	.9797	-.0968	.2252	.8118	.7640	.3169	.2076	.2399
.3658	.3652	1.0000	.9956	.5159	.4295	.0817	-.0232	.3675	.3153	.0929	.2439	.0797
.3715	.3755	.9956	1.0000	.5173	.4395	.0782	-.0183	.3735	.3224	.0926	.2374	.0805
.9273	.9113	.5159	.5173	1.0000	.9640	-.0584	.1263	.8341	.7532	.2447	.2891	.2218
.9627	.9797	.4295	.4395	.9640	1.0000	-.0846	.1847	.8379	.7728	.2837	.2411	.2335
-.0930	-.0968	.0817	.0782	-.0584	-.0846	1.0000	-.0333	-.1409	-.2159	-.0330	.0801	.0769
.2313	.2252	-.0232	-.0183	.1263	.1847	-.0333	1.0000	.1199	.1597	.2364	.0543	.1583
.8143	.8118	.3675	.3735	.8341	.8379	-.1409	.1199	1.0000	.7605	.1075	.0942	.1340
.7650	.7640	.3153	.3224	.7532	.7728	-.2159	.1597	.7605	1.0000	.2416	.1782	.2065
.3341	.3169	.0929	.0926	.2447	.2837	-.0330	.2364	.1075	.2416	1.0000	.4213	.2140
.2236	.2076	.2439	.2374	.2891	.2411	.0801	.0543	.0942	.1782	.4213	1.0000	.2553
.2535	.2399	.0797	.0805	.2218	.2335	.0769	.1583	.1340	.2065	.2140	.2553	1.0000

By following the procedure outlined here, a total of 5,000 synthetic observations, consisting of 13 synthetic variables, were generated. The original data set was used to construct the cdfs. A smoothing constant  $p = 0.5$  was used to smooth the ends of the cdfs. Each synthetic observation carries a sampling weight of 1.13 (5,655/5,000). The sampling weight is required to match totals from the simulated data with totals from the original data.

A total of nine iterative refinement steps were required to bring the synthetic data rank correlation matrix to within 0.01% of the target rank correlation matrix. The rank correlation structure of the synthetic data matrix at the end of the initial step and at the end of the last refinement step, along with the sum absolute deviation of the upper (or lower) off-diagonal elements of the rank correlation matrix at the end of each iterative step are as follows:

Iteration: 0 Abs Diff: .7188911E+00

Rank Correlation Matrix:

1.0000	.9794	.3482	.3534	.9319	.9649	-.0891	.2190	.8203	.7681	.3158	.2119	.2387
.9794	1.0000	.3477	.3570	.9174	.9809	-.0924	.2138	.8180	.7669	.3004	.1975	.2266
.3482	.3477	1.0000	.9959	.4928	.4097	.0818	-.0258	.3520	.3011	.0844	.2294	.0707
.3534	.3570	.9959	1.0000	.4943	.4187	.0784	-.0222	.3577	.3075	.0833	.2235	.0716
.9319	.9174	.4928	.4943	1.0000	.9663	-.0582	.1260	.8386	.7577	.2365	.2713	.2113
.9649	.9809	.4097	.4187	.9663	1.0000	-.0821	.1776	.8419	.7753	.2708	.2282	.2216
-.0891	-.0924	.0818	.0784	-.0582	-.0821	1.0000	-.0309	-.1340	-.2059	-.0285	.0754	.0737
.2190	.2138	-.0258	-.0222	.1260	.1776	-.0309	1.0000	.1183	.1515	.2179	.0500	.1460
.8203	.8180	.3520	.3577	.8386	.8419	-.1340	.1183	1.0000	.7608	.1125	.0956	.1332
.7681	.7669	.3011	.3075	.7577	.7753	-.2059	.1515	.7608	1.0000	.2318	.1685	.1939
.3158	.3004	.0844	.0833	.2365	.2708	-.0285	.2179	.1125	.2318	1.0000	.3981	.1955
.2119	.1975	.2294	.2235	.2713	.2282	.0754	.0500	.0956	.1685	.3981	1.0000	.2366
.2387	.2266	.0707	.0716	.2113	.2216	.0737	.1460	.1332	.1939	.1955	.2366	1.0000

Iteration: 1 Abs Diff: .6919879E-01

Iteration: 2 Abs Diff: .9697238E-02

Iteration: 3 Abs Diff: .3696279E-02

Iteration: 4 Abs Diff: .3167076E-02

Iteration: 5 Abs Diff: .3095646E-02  
 Iteration: 6 Abs Diff: .3083975E-02  
 Iteration: 7 Abs Diff: .3081255E-02  
 Iteration: 8 Abs Diff: .3080559E-02  
 Iteration: 9 Abs Diff: .3080209E-02

Rank Correlation Matrix:

```

1.0000 .9782 .3657 .3714 .9273 .9627 -.0930 .2313 .8143 .7650 .3340 .2236 .2534
.9782 1.0000 .3651 .3754 .9113 .9797 -.0967 .2252 .8118 .7640 .3169 .2076 .2399
.3657 .3651 1.0000 .9956 .5158 .4294 .0817 -.0232 .3674 .3152 .0929 .2439 .0797
.3714 .3754 .9956 1.0000 .5172 .4394 .0782 -.0183 .3734 .3223 .0926 .2374 .0805
.9273 .9113 .5158 .5172 1.0000 .9640 -.0584 .1263 .8341 .7532 .2447 .2891 .2218
.9627 .9797 .4294 .4394 .9640 1.0000 -.0846 .1847 .8379 .7728 .2837 .2411 .2335
-.0930 -.0967 .0817 .0782 -.0584 -.0846 1.0000 -.0333 -.1409 -.2159 -.0330 .0801 .0769
.2313 .2252 -.0232 -.0183 .1263 .1847 -.0333 1.0000 .1199 .1596 .2363 .0543 .1583
.8143 .8118 .3674 .3734 .8341 .8379 -.1409 .1199 1.0000 .7605 .1075 .0942 .1340
.7650 .7640 .3152 .3223 .7532 .7728 -.2159 .1596 .7605 1.0000 .2416 .1782 .2065
.3340 .3169 .0929 .0926 .2447 .2837 -.0330 .2363 .1075 .2416 1.0000 .4212 .2139
.2236 .2076 .2439 .2374 .2891 .2411 .0801 .0543 .0942 .1782 .4212 1.0000 .2552
.2534 .2399 .0797 .0805 .2218 .2335 .0769 .1583 .1340 .2065 .2139 .2552 1.0000
    
```

The resulting rank correlation matrix is now almost identical to the target. The Pearson correlation coefficient, unlike the rank correlation coefficient, is extremely sensitive to the actual values of the variables and therefore is difficult to reproduce exactly. The Pearson rank correlation coefficient matrix for the original population and the synthetic data are as follows:

Pearson Correlation Coefficient - Actual Data

```

1.0000 .9292 .4212 .4436 .8074 .9218 -.0412 .1086 .7241 .0277 .1816 .0926 .1131
.9292 1.0000 .3624 .3917 .7457 .9824 -.0452 .1008 .7393 .0383 .1844 .0928 .1118
.4212 .3624 1.0000 .9500 .7459 .4601 .0244 .0105 .3373 -.0037 .1020 .0913 .0641
.4436 .3917 .9500 1.0000 .7395 .4949 .0194 .0208 .3680 -.0062 .1079 .1045 .0728
.8074 .7457 .7459 .7395 1.0000 .8203 .0152 .0601 .6330 .0130 .1669 .1165 .1052
.9218 .9824 .4601 .4949 .8203 1.0000 -.0413 .0903 .7518 .0338 .1885 .1074 .1158
-.0412 -.0452 .0244 .0194 .0152 -.0413 1.0000 -.0183 -.0635 -.1084 -.0310 .0659 .0587
.1086 .1008 .0105 .0208 .0601 .0903 -.0183 1.0000 .1090 -.0357 .2251 .0513 .1589
.7241 .7393 .3373 .3680 .6330 .7518 -.0635 .1090 1.0000 .0479 .1392 .0798 .0968
.0277 .0383 -.0037 -.0062 .0130 .0338 -.1084 -.0357 .0479 1.0000 -.0974 -.1345 -.3586
.1816 .1844 .1020 .1079 .1669 .1885 -.0310 .2251 .1392 -.0974 1.0000 .4494 .2635
.0926 .0928 .0913 .1045 .1165 .1074 .0659 .0513 .0798 -.1345 .4494 1.0000 .3288
.1131 .1118 .0641 .0728 .1052 .1158 .0587 .1589 .0968 -.3586 .2635 .3288 1.0000
    
```

Pearson Correlation Coefficient - Synthetic Data

```

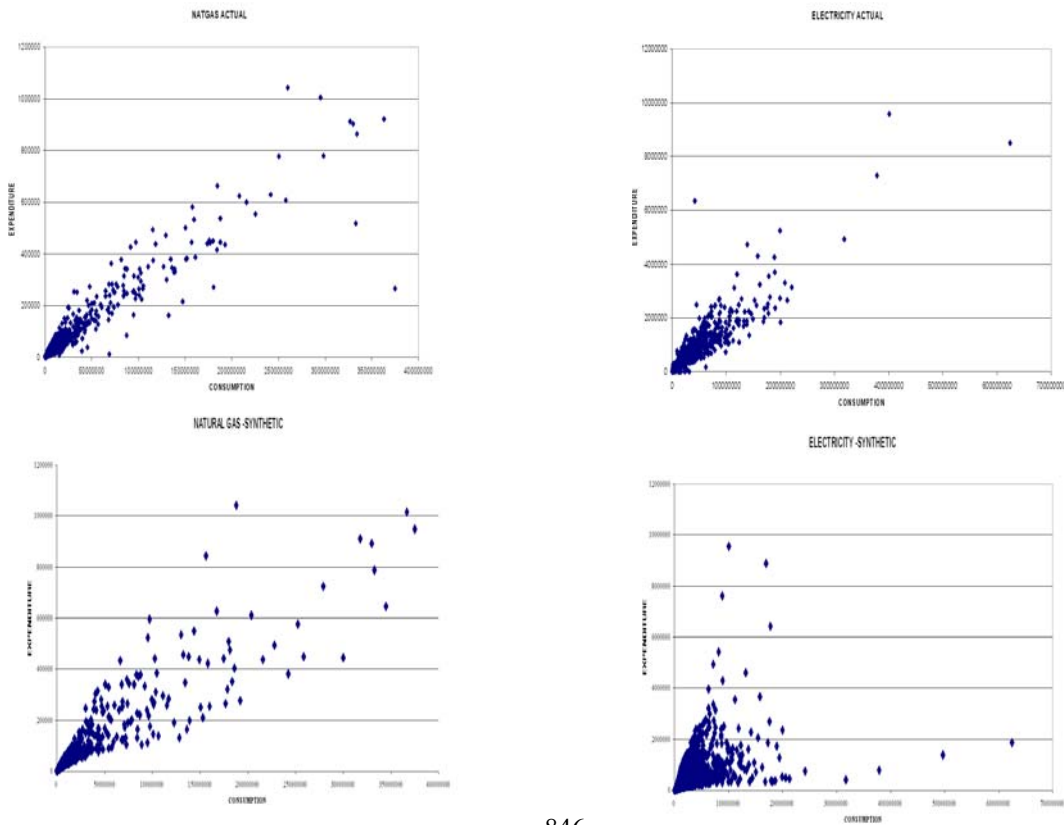
1.0000 .6883 .2485 .2366 .4731 .5835 -.0468 .1224 .4140 .1965 .1317 .0793 .1023
.6883 1.0000 .2737 .2860 .4223 .7481 -.0573 .1175 .4293 .2592 .1219 .0716 .1008
.2485 .2737 1.0000 .9335 .4557 .3638 .0135 .0192 .2461 .1047 .0529 .0544 .0460
.2366 .2860 .9335 1.0000 .4077 .3683 .0143 .0259 .2435 .1312 .0629 .0689 .0500
.4731 .4223 .4557 .4077 1.0000 .6804 -.0181 .0276 .5273 .2282 .0661 .1064 .0830
.5835 .7481 .3638 .3683 .6804 1.0000 -.0465 .0796 .5440 .2988 .1028 .0934 .1014
-.0468 -.0573 .0135 .0143 -.0181 -.0465 1.0000 -.0483 -.1010 -.0834 -.0402 .0551 .0494
.1224 .1175 .0192 .0259 .0276 .0796 -.0483 1.0000 .0558 .0335 .2053 .0417 .1557
.4140 .4293 .2461 .2435 .5273 .5440 -.1010 .0558 1.0000 .3128 .0271 .0472 .0699
.1965 .2592 .1047 .1312 .2282 .2988 -.0834 .0335 .3128 1.0000 .0328 .0339 .0384
.1317 .1219 .0529 .0629 .0661 .1028 -.0402 .2053 .0271 .0328 1.0000 .3348 .1854
    
```

.0793 .0716 .0544 .0689 .1064 .0934 .0551 .0417 .0472 .0339 .3348 1.0000 .2125  
 .1023 .1008 .0460 .0500 .0830 .1014 .0494 .1557 .0699 .0384 .1854 .2125 1.0000

For highly correlated variables, such as variables 3 and 4 (natural gas consumption and expenditure), the procedure could successfully reproduce the Pearson rank correlation coefficient. But the same is not true for highly correlated variables 1 and 2 (electricity consumption and expenditure).

By separately plotting the consumption values against the corresponding expenditure values for the original population, as well as for the synthetic data, it is easy to see that the electricity data has a more highly skewed distribution than the natural gas. Additionally it contains quite a few outliers in the upper tail which could be looked at as population uniques. These two characteristics are responsible for these effects. The natural gas consumption and expenditure values are more evenly distributed over the range of the population with minor exceptions. By fine tuning the empirical cdfs, or by replacing the empirical cdfs by appropriate theoretical cdfs, one should improve the quality of the bivariate characteristics of synthetic electricity data.

Original Population & Synthetic Data  
 Consumption VS Expenditure



So far as the univariate statistical properties are concerned, the LHS-based procedure is well known to reproduce them almost exactly. This is apparent from the table below:

Univariate Statistics

Var	Population		Synthetic Data	
	Average	Stnd Dev	Average	Stnd Dev
1	.7929E+07	.2312E+08	.7961E+07	.2345E+08
2	.1539E+06	.4348E+06	.1544E+06	.4394E+06
3	.4498E+07	.2109E+08	.4516E+07	.2126E+08
4	.1546E+05	.5900E+05	.1551E+05	.5943E+05
5	.1523E+08	.4570E+08	.1528E+08	.4622E+08
6	.1848E+06	.4956E+06	.1853E+06	.4999E+06
7	.1238E+02	.1089E+02	.1238E+02	.1089E+02
8	.5398E+01	.1799E+01	.5399E+01	.1798E+01
9	.1293E+06	.2671E+06	.1294E+06	.2674E+06
10	.1385E+04	.1068E+05	.1390E+04	.1068E+05
11	.6492E+02	.4055E+02	.6493E+02	.4055E+02
12	.8412E+02	.3123E+02	.8412E+02	.3123E+02
13	.8828E+02	.2504E+02	.8828E+02	.2502E+02

## 8. Conclusion

The Latin Hypercube Sampling technique offers a viable alternative to other methods of protection of sensitive data in public use data files.

### References

- [1] Dandekar, Ramesh A. (1993), "Performance Improvement of Restricted Pairing Algorithm for Latin Hypercube Sampling", ASA Summer conference (unpublished).
- [2] Iman R.L. and Conover W. J. (1982), "A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables", *Commun. Stat.*, B11(3): pp. 311-334.
- [3] McKay M.D., Conover W. J. and Beckman, R. J. (1979), "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code", *Technometrics* 21(2): pp. 239-245.
- [4] Stein M. (1987), "Large Sample Properties of Simulations Using Latin Hypercube Sampling", *Technometrics* (29)2: pp. 143-151.