Analysis and Synthesis of Correlated Data[1]

Introduction

Monte-Carlo (MC) simulation of multiple instances of devices, testers, use conditions, etc. requires generation of many instances of sets of numbers. For example one might need the elevation and daily average temperature at each of many locations. Since, for example, higher elevations are associated with lower temperatures, these numbers are likely to be correlated. Once the data has been acquired, the generation of MC samples can be a random sampling of the original data. For example, for each MC instance, one would randomly select a location, and use the mean temperature and elevation for that location. One could extend this idea by using weighted sampling.

Monte-Carlo sampling from original data is warranted if the original data is not a sample of the population, but in fact represents the entire population. However, if the data is actually a sample of the population, there are disadvantages to MC sampling from original data:

1. Original data may have outliers which are not representative of the design targets.

2. Original data is not easily adjustable to do "what-if" scenarios for cases in which data does not yet exist.

3. If the number of MC instances sampled exceeds the original dataset size, the sampling will not be true, since many instances will be repetitively sampled.

So, we develop in this paper a methodology by which a parametric statistical model may be extracted, and then MC samples can be generated from the parametric model, rather than directly from data. MC samples based on a parametric statistical model will not generate statistically significant outliers, may be adjusted by manipulation of underlying statistical parameters (to explore "what-if" scenarios), and may be used to generate unlimited MC samples. The method described here preserves correlations, and preserves certain constraints on statistical variables which commonly occur. The method can handle variables with the following types of constraint:

1. Unconstrained: $-\infty < x < \infty$

2. Positive constraint: $x \ge 0$ (variances, such as the daily temperature variance, have this property)

3. Fractional constraint: $0 \le x \le 1$ (fraction of time in a state)

4. Time-in-states constraint: $0 \le x_1 \le 1$, $0 \le x_2 \le 1$, $0 \le x_3 \le 1$, where $x_1 + x_2 + x_3 + ... = 1$ (fraction of time in multiple states).

[1] C. Glenn Shirley, May 7, 2004

The methodology described here will find application in synthesizing use conditions by MC methods.

The plan of the paper is to show the method by which statistical parameters for correlated multinormal data can be extracted, and then used to synthesize correlated data with the same parameters as the original data. Then, we extend the method by showing how to handle statistical variables which are constrained various ways.

Analysis and Synthesis of Correlated, Unconstrained Data

Analysis

Consider the following dataset consisting of 1367 associated numbers shown in Table I. Each instance corresponds to a row in this table. There is no particular constraint on the values. The objective is to fit this data to a parametric statistical model, and then synthesize data from the model. Distributions and correlations of this data are shown in the Jmp analysis in Figs. 1 and 2.

Table I Example dataset. Columns 1 and 2 are correlated to each other, but Column 3 is uncorrelated to the others.

T_mean1	T_mean2	Tmean_3
11.128	17.236	30.920
12.154	17.286	27.497
15.561	15.561	28.182
10.667	15.627	28.261
11.849	15.552	32.524
10.939	19.079	24.159
15.615	15.615	28.358
12.950	16.642	27.327
12.282	20.762	31.387
8.747	14.628	29.233
10.568	19.341	28.046
11.525	19.214	27.126
10.129	17.698	25.932
12.986	20.993	19.861
12.365	20.286	35.752
13.357	21.213	33.168
13.953	20.863	32.545
11.957	17.602	30.010
13.523	19.684	32.959
11.996	18.736	24.430
13.599	20.784	25.789
14.267	20.255	24.692
14.286	21.897	36.028
6.931	18.187	32.267

and 1343 more rows.

Tmean_1	
Mean	16.279808
Std Dev	5.8725802
Std Err Mean	0.1588344
upper 95% Mean	16.591394
lower 95% Mean	15.968222
N	1367

Tmean_2	
Mean	24.373974
Std Dev	3.6992313
Std Err Mean	0.1000523
upper 95% Mean	24.570247
lower 95% Mean	24.177701
N	1367

Tmean_3	
Mean	30.048748
Std Dev	4.0786681
Std Err Mean	0.1103149
upper 95% Mean	30.265153
lower 95% Mean	29.832343
N	1367

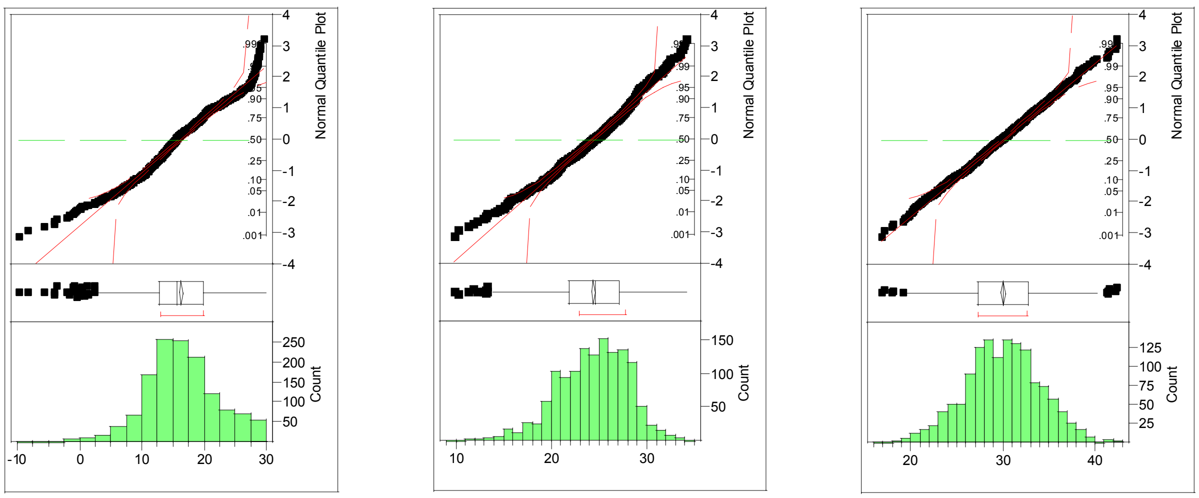

Fig. 1 Distributions of data in Table I

	T_mean1	T_mean2	Tmean_3
T_mean1	1.0000	0.7484	0.0445
T_mean2	0.7484	1.0000	0.0412
Tmean_3	0.0445	0.0412	1.0000

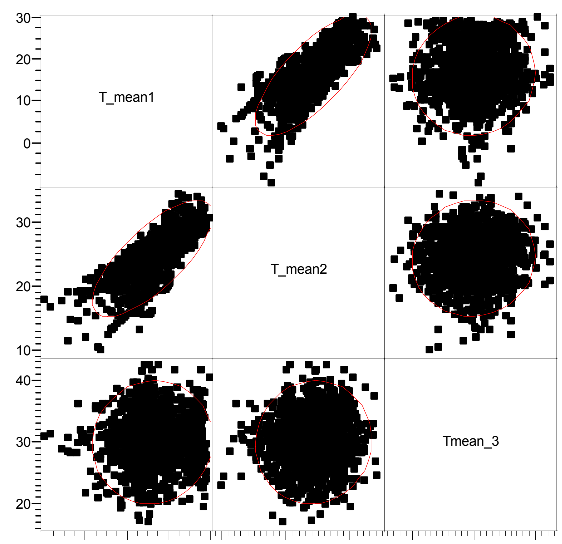

Fig. 2 Correlations of data in Table I.

Specific statistics of this data are of interest. The mean of the m^{th} variable is:

$$\langle x \rangle_m = \frac{1}{N} \sum_i x_m^i \quad . \tag{1}$$

The covariance between the m^{th} and n^{th} variable is:

$$V_{mn} = \frac{1}{N} \sum_i \left(x_m^i - \langle x \rangle_m \right)\left(x_n^i - \langle x \rangle_n \right) \quad . \tag{2}$$

The variance of the m^{th} variable is given by Eq. (2) with $m = n$. The correlation coefficient between the m^{th} and n^{th} variable (that is, the correlation matrix) is:

$$\rho_{mn} = \frac{V_{mn}}{\sqrt{V_{mm}}\sqrt{V_{nn}}} \quad . \tag{3}$$

It is possible for the covariance matrix to exist, but for computational difficulty to occur if a variable has zero variance. For this reason, it is usually more convenient to compute the covariance matrix, and then test for vanishing variances before computing Eq. (3). Some software (in particular Jmp) provide the correlation matrix (if it is not singular). In this case, the covariance matrix can be computed from the correlation matrix and the variance for each variable using a rearrangement of Eq. (3):

$$V_{mn} = \sqrt{V_{mm}}\sqrt{V_{nn}}\,\rho_{mn}$$

Synthesis of a Single Variable, Normally Distributed

If just one variable is of interest, then synthesis of a random variable with mean $\langle x \rangle$ and variance V is done using

$$x^i = \langle x \rangle + \sqrt{V} \times z^i \tag{4}$$

where z^i is the i-th instance of a normally distributed variable with zero mean and unit variance which can be generated via the inverse normal cdf according to

$$z^i = \Phi^{-1}\left(ran^i \right) \tag{5}$$

In the case of *Tmean*_1 this becomes:

$$x^i = 16.279808 + 5.8725802 \times \Phi^{-1}\left(ran^i \right) \tag{6}$$

where the values were obtained from the Jmp statistics in Fig. 1. 2000 data points generated from this are compared with the original data in Fig. 3. Notice that all the statistical indicators are well matched, but of necessity, the detailed shape (especially in the tails) is not reproduced in the synthesized data.

Tmean_1

Mean	16.279808
Std Dev	5.8725802
Std Err Mean	0.1588344
upper 95% Mean	16.591394
lower 95% Mean	15.968222
N	1367

Tmean_1 (synthesized)

Mean	16.182618
Std Dev	5.88826
Std Err Mean	0.1316655
upper 95% Mean	16.440834
lower 95% Mean	15.924402
N	2000

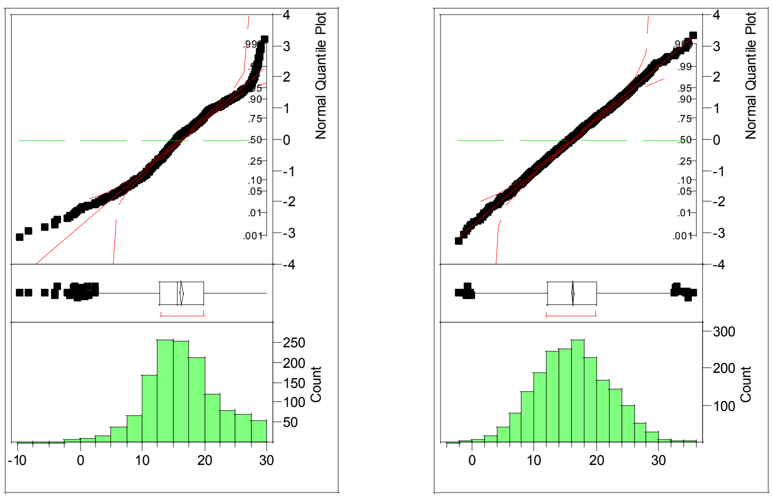

Fig 3. *Tmean*_1 raw data (left) versus synthesized (right).

Synthesis of Multiple Correlated Variables, Normally Distributed

Each individual variable in Table I can be separately analyzed and synthesized in the manner of the previous section. However, if it is important to generate samples of the 3 variables which preserve the correlation of variables, then we use an extension of the approach described in previous section.

A generalized version of Eq. (4) is the matrix equation

$$\vec{x} = \overrightarrow{\langle x \rangle} + \vec{C}'\,\vec{z} \qquad (7a)$$

or, explicitly,

$$\begin{bmatrix} x_1^i \\ x_2^i \\ x_3^i \\ .. \\ .. \end{bmatrix} = \begin{bmatrix} \langle x \rangle_1 \\ \langle x \rangle_2 \\ \langle x \rangle_3 \\ .. \\ .. \end{bmatrix} + \begin{bmatrix} C_{11} & 0 & 0 & .. & .. \\ C_{21} & C_{22} & 0 & & \\ C_{31} & C_{32} & C_{33} & & \\ .. & & & & \\ .. & & & & \end{bmatrix} \begin{bmatrix} z_1^i \\ z_2^i \\ z_3^i \\ .. \\ .. \end{bmatrix} \tag{7b}$$

where the column vector on the right-hand side consists of n independently sampled normally distributed variables (mean 0, and variance unity), and where the matrix is the lower-triangular Choelesky root of the covariance matrix:

$$\overleftrightarrow{V} = \overleftrightarrow{C}'\overleftrightarrow{C} \tag{8}$$

or

$$\begin{bmatrix} V_{11} & V_{12} & V_{13} & .. & .. \\ V_{21} & V_{22} & V_{23} & & \\ V_{31} & V_{32} & V_{33} & & \\ .. & & & .. & \\ .. & & & & .. \end{bmatrix} = \begin{bmatrix} C_{11} & 0 & 0 & .. & .. \\ C_{21} & C_{22} & 0 & & \\ C_{31} & C_{32} & C_{33} & & \\ .. & & & .. & \\ .. & & & & .. \end{bmatrix} \begin{bmatrix} C_{11} & C_{12} & C_{13} & .. & .. \\ 0 & C_{22} & C_{23} & & \\ 0 & 0 & C_{33} & & \\ .. & & & .. & \\ .. & & & & .. \end{bmatrix} \tag{9}$$

where the upper and lower Cholesky roots are transposes of each other ($C_{nm} = C_{mn}$), and where the covariance matrix V is real symmetric ($V_{nm} = V_{mn}$). In addition, the covariance matrix must be positive definite if the Choelsky root is to be extracted. Positive definiteness means that, for any vector $\vec{x}$, $\vec{x}'\overleftrightarrow{V}\vec{x} > 0$. It can be shown that this will be true if and only if

1. $V_{ii} > 0$ for all i
2. $V_{ii} \times V_{jj} > V_{ij}^2$ for $i \neq j$
3. Element with largest modulus lies on main diagonal.
4. $\det(\overleftrightarrow{V}) > 0$

Notice that if one of the variables in Table I has zero variance, that is, has the same value for all instances, then at least conditions 1 and 2 will be violated, and it will not be possible to extract the Cholesky root. If this is the case, however, it is not necessary to consider the variable as a statistical variable. Although this therefore presents no mathematical difficulty, it is a case which must be taken into account when writing software code to extract statistical synthesis models.

An algorithm to extract the Cholesky root of a real-symmetric positive definite matrix is readily available[2], and has been implemented in Excel Visual Basic.

It is useful to point out that in the case of a single variable, Eq. (9) reduces to

$$V_{11} = (C_{11})^2 \tag{10}$$

so that Eq. (7a) becomes Eq. (4). In a sense, the Cholesky root is the "square root" of the covariance matrix.

The individual variable statistics, correlation matrix, and lower-triangular Cholesky root may be extracted from the data in Table I, and are shown in Table II.

Table II Mean and standard deviation (square root of variance) for the 3 variables in Table I. Also tabulated are the correlation matrix (Rho), and the lower-triangular Cholesky root of the covariance matrix. This is an application of an Excel tool "UCPET".

Statistics	Mean	SD	
T_mean1	*16.2798*	5.8704	
T_mean2	*24.3740*	3.6979	
Tmean_3	*30.0487*	4.0772	

Rho	T_mean1	T_mean2	Tmean_3
T_mean1	1.0000	0.7484	0.0445
T_mean2	0.7484	1.0000	0.0412
Tmean_3	0.0445	0.0412	1.0000

Cholesky	T_mean1	T_mean2	Tmean_3
T_mean1	*5.8704*	*0*	*0*
T_mean2	*2.7675*	*2.4526*	*0*
Tmean_3	*0.1814*	*0.0483*	*4.0729*

Notice in Table II that *T_mean*1 and *T_mean*2 have a significant correlation, whereas *Tmean*_3 has a very low (effectively zero) correlation with *T_mean*1 and *T_mean*2.

The values shown in italics in Table I are the parameters needed in Eq. (7b) to synthesize variables with the required correlation.

So Eq. (7b) will be written

$$\begin{bmatrix} x_1^i \\ x_2^i \\ x_3^i \end{bmatrix} = \begin{bmatrix} 16.2798 \\ 24.3740 \\ 30.0487 \end{bmatrix} + \begin{bmatrix} 5.8704 & 0 & 0 \\ 2.7675 & 2.4526 & 0 \\ 0.1814 & 0.0483 & 4.0729 \end{bmatrix} \begin{bmatrix} z_1^i \\ z_2^i \\ z_3^i \end{bmatrix} \tag{11}$$

---

[2] "Numerical Recipes..", W.H. Press, B.P. Flannery, S. A. Teukolsky, W. T. Vetterling, Cambridge Univ. Press (1992)

where the elements of the “z” vector are independent (uncorrelated) normal deviates samples from a normal distribution with a mean of zero, and a variance of unity. Notice that the first equation in Eq. (11) is identical with Eq. (4). Any number of triples of correlated numbers may be generated by repeated use of Eq. (11) with triples of uncorrelated normal deviates. When this is done, we obtain the sample shown in Table III.

Table III Synthetic data generated by using Eq. (11).

T_mean1	T_mean2	Tmean_3
6.908	14.681	30.052
11.192	20.304	37.431
13.842	19.550	23.273
16.696	30.713	28.610
19.559	24.448	33.982
23.137	26.160	28.866
15.073	25.369	23.661
11.280	23.093	21.982
13.619	21.542	25.599
22.802	25.896	28.223
18.356	26.505	26.362
20.306	20.933	32.166
25.493	27.225	27.881
19.685	27.315	32.662
23.614	25.213	34.447
13.311	22.289	26.791
10.885	24.881	25.881
17.354	24.166	32.518
17.060	29.140	32.881
12.900	20.744	27.849
8.239	20.076	30.855
25.225	32.867	33.289
25.138	25.664	25.641

.. 2000 rows in total.

The synthetic data in Table III may be analyzed in exactly the same way as the original data was analyzed above. Figs. 4 and 5 show the results of the analysis and may be compared with Figs. 1 and 2.

Tmean_1	
Mean	16.461433
Std Dev	5.7884461
Std Err Mean	0.1294336
upper 95% Mean	16.715272
lower 95% Mean	16.207594
N	2000

Tmean_2	
Mean	24.465198
Std Dev	3.6643954
Std Err Mean	0.0819384
upper 95% Mean	24.625892
lower 95% Mean	24.304504
N	2000

Tmean_3	
Mean	30.085861
Std Dev	4.0826019
Std Err Mean	0.0912898
upper 95% Mean	30.264894
lower 95% Mean	29.906828
N	2000

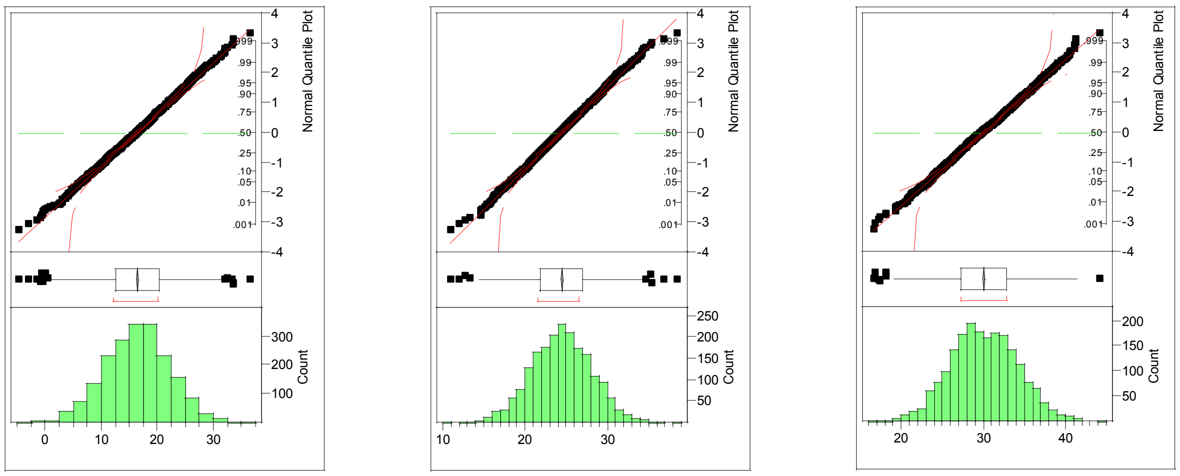

Fig. 4. Synthesized data generated using Eq. (11)

	T_mean1	T_mean2	Tmean_3
T_mean1	1.0000	0.7434	0.0021
T_mean2	0.7434	1.0000	0.0165
Tmean_3	0.0021	0.0165	1.0000

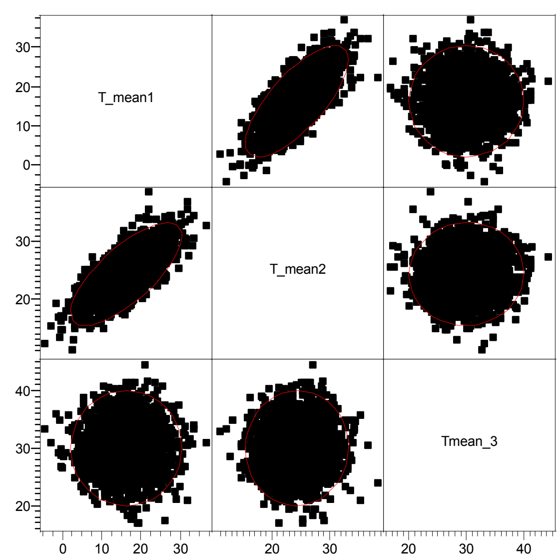

Fig. 5. Correlations of synthesized data. This is to be compared with Fig. 2.

Analysis and Synthesis of Multiple Correlated Constrained Variables

Sometimes statistical variables (or data) are constrained. Table IV is an example of data for which two of the statistical variable values for each unit are constrained:

$$0 \le f1 \le 1 \tag{12a}$$

$$0 \le T\_\mathrm{var}1 < \infty \tag{12b}$$

whereas one is not constrained

$$-\infty < T\_mean1 < \infty . \quad \text{(unconstrained, assumed normal)} \tag{12c}$$

The constraint on $f1$ is typical of statistical variables which represent a fraction. The constraint on T_var1 is typical of variables which are positive definite, such as variances of a variable for a specific unit.

Table IV Example data with fraction of time in ambient f1, average temperature T_mean1, and variance T_var1.

f1	T_mean1	T_var1
0.928	11.1283306	11.1532073
0.898	12.1539885	12.5669417
0.773	15.5611623	6.86547489
0.891	10.6667886	22.8168552
0.982	11.8492834	15.2243792
0.911	10.9386929	25.4351889
0.714	15.615437	9.33561594
0.934	12.9496939	14.2044274
0.942	12.2815545	20.8951555
0.776	8.74737591	29.1459802
0.912	10.5676292	27.5987067
0.872	11.5250515	22.8812167
0.909	10.1287694	26.0174409
0.959	12.9862047	23.5971688
0.757	12.3647639	23.934025
0.913	13.3573995	23.2383428
0.757	13.9532134	18.141637
0.948	11.9565801	21.4910688
0.853	13.5228469	21.9822903

.. a total of 1367 data points.

f1

Mean	0.8528425
Std Dev	0.0996552
Std Err Mean	0.0026954
upper 95% Mean	0.85813
lower 95% Mean	0.8475551
N	1367

T_mean1

Mean	16.279808
Std Dev	5.8725802
Std Err Mean	0.1588344
upper 95% Mean	16.591394
lower 95% Mean	15.968222
N	1367

T_var1

Mean	14.956333
Std Dev	10.784559
Std Err Mean	0.2916877
upper 95% Mean	15.528537
lower 95% Mean	14.384128
N	1367

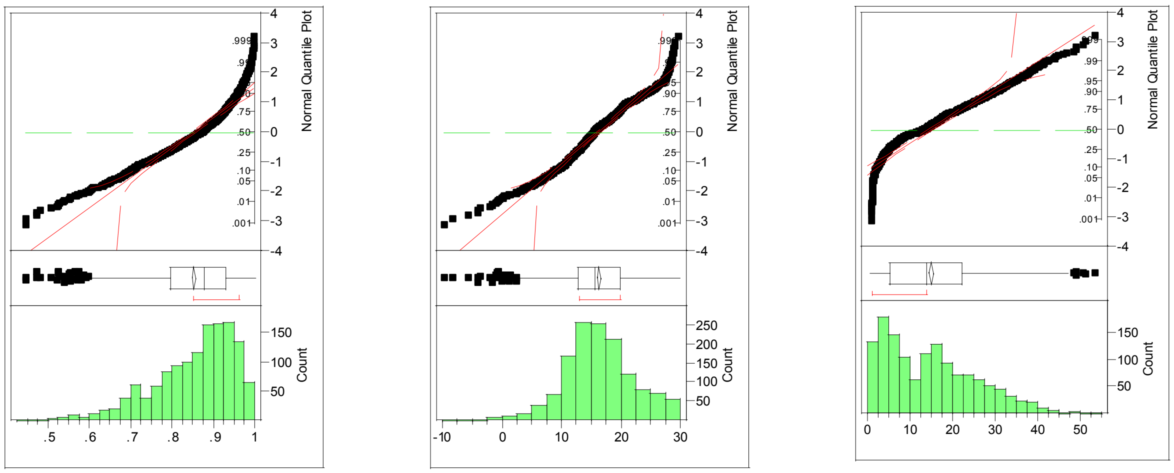

Fig. 6 Distributions of data in Table IV, showing that $f1$ and T_var1 are constrained by Eqs. (12a) and (12b).

	f1	T_mean1	T_var1
f1	1.0000	-0.0072	0.0129
T_mean1	-0.0072	1.0000	-0.6762
T_var1	0.0129	-0.6762	1.0000

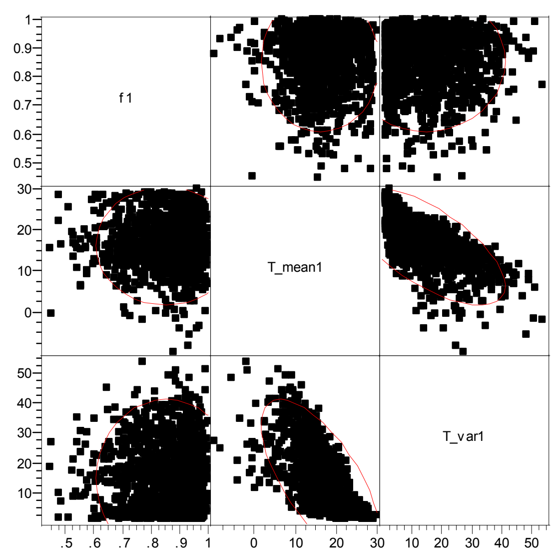

Fig. 7 Correlations of data in Table IV. The "flattened" sides of the correlations involving one or both of $f1$ and T_var1 are due to the constraints of Eqs. (12a) and (12b)

We determined the parameters of normal, correlated, distributions which fit the data in Table IV. Those parameters were used to synthesize 2000 data points (*f*1, *Tmean*_1, *T_var*1). Then the synthesized data were analyzed in the same way as the original data, and the results are shown in Figs. 8 and 9. It is apparent that sampled values of *f*1 and *T_var*1 will violate the constraints of Eqs. (12a) and (12b). Sampled values such as these will cause computational problems. It is a bad practice to do a "normal" MC simulation and then discard samples which violate the constraints of Eqs. (12a) and (12b), because, for one thing, this will not preserve the average and variance of the original data.

f1

Mean	0.8517617
Std Dev	0.0985768
Std Err Mean	0.0022042
upper 95% Mean	0.8560846
lower 95% Mean	0.8474388
N	2000

T_mean1

Mean	16.564037
Std Dev	5.9199803
Std Err Mean	0.1323748
upper 95% Mean	16.823644
lower 95% Mean	16.30443
N	2000

T_var1

Mean	14.511564
Std Dev	11.067986
Std Err Mean	0.2474877
upper 95% Mean	14.996925
lower 95% Mean	14.026203
N	2000

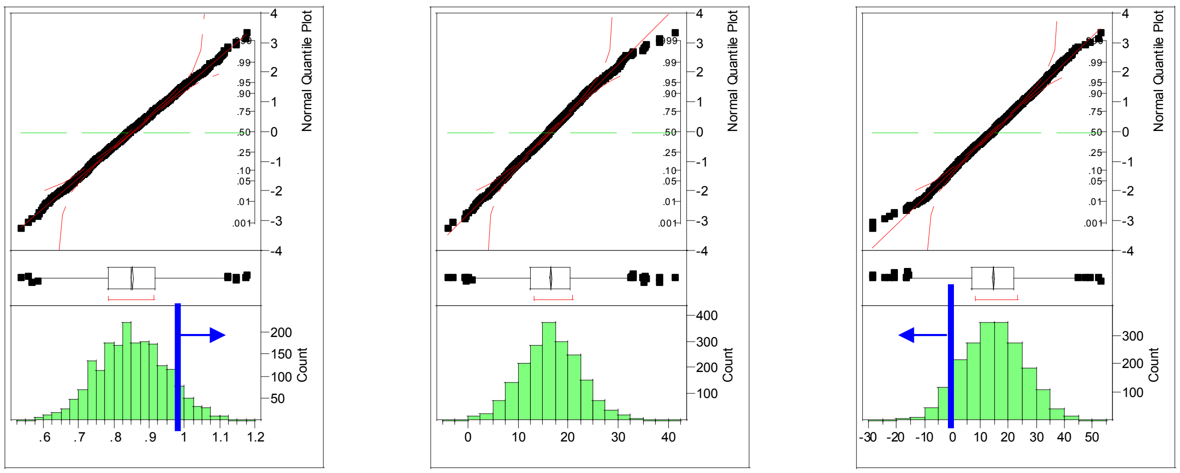

Fig. 8. Synthesis of from "unconstrained" model (assuming that variables are normally distributed) derived from data in Table IV. This produces unphysical values of parameters (arrows, and blue lines). That is, negative variances, and values of f1 which exceed unity are generated.

	f1	T_mean1	T_var1
f1	1.0000	0.0080	0.0009
T_mean1	0.0080	1.0000	-0.6910
T_var1	0.0009	-0.6910	1.0000

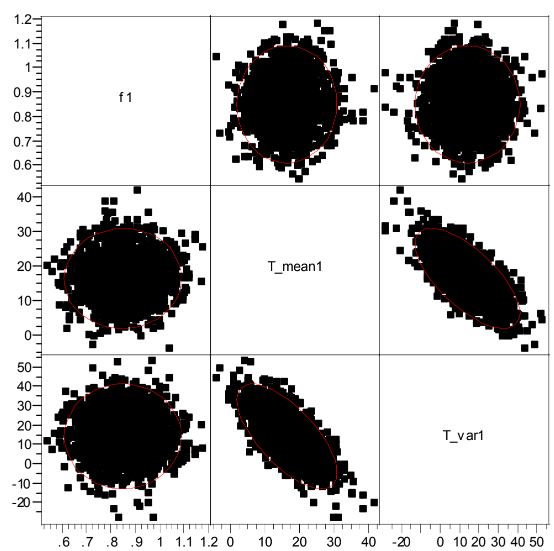

Fig. 9. Correlation of variables synthesized from fits to the data in Table IV assuming normal distributions for the statistical variables.

To synthesize statistical variables which will satisfy the constraints of Eqs. (12a) and (12b) it is necessary to transform the original data in each column of Table IV from a "constrained" variable into one which is not constrained, then fit a correlated normal model to the transformed variables. Synthesis from the model requires sampling of correlated normal variables (some of them transformed), and then application of the inverse transformations into the constrained variables. We show three, not necessarily unique, transformations which cover all scenarios yet encountered in the study of use conditions. The transformation functions have the properties 1) The mean and variance of original and transformed variables are preserved, and 2) in the limit of the constraint disappearing (mean and variance such that little of the data is actually constrained), the transformation *automatically* reverts to an identity transformation.

Positive Constraint: Variables Constrained by $x_n \geq 0$

For variables which are variances, and $x_n \geq 0$ a useful transformation into the variables y_n may be derived from a condition on two cumulative distribution functions (cdfs):

$$\Phi\left(\frac{y_n - E}{\sqrt{V}}\right) = 1 - ChiDist\left(\frac{df}{E}x_n, df\right) \tag{14a}$$

where "ChiDist" is Excel terminology for the Chisquare distribution, and where the degrees of freedom is the integer (> 0) closest to

$$df = \frac{2E^2}{V} \tag{14b}$$

The variable x_n on the rhs of Eq. (14) has mean E and variance V because it is a property of the Chisquare distribution that for $ChiDist(z, df)$, $E(z) = <z> = df$, and $Var(z) = <z^2> - <z>^2 = 2df$, so for the variable x in Eq. (14)

$$\left\langle \frac{df}{E}x \right\rangle = \frac{df}{E}\langle x \rangle = df \quad \text{or} \quad \langle x \rangle = E$$

and

$$Var\left(\frac{df}{E}x\right) = \left(\frac{df}{E}\right)^2 V = 2df \quad \text{or} \quad df = \frac{2E^2}{V}$$

The Chisquare distribution has the property

$$ChiDist(x, df) \xrightarrow[df \to \infty]{} 1 - \Phi\left(\frac{x - df}{\sqrt{2 \times df}}\right) \tag{15}$$

so, using Eq. (15) we may write the limiting form for the rhs of Eq. (14a) as

$$1 - ChiDist\left(\frac{df}{E}x_n, df\right) \xrightarrow[df \to \infty]{}$$

$$\Phi\left(\frac{\frac{df}{E}x - df}{\sqrt{2 \times df}}\right) = \Phi\left(\sqrt{\frac{df}{2}}\frac{x - E}{E}\right) = \Phi\left(\sqrt{\frac{df}{2E^2}}(x - E)\right) = \Phi\left(\frac{x - E}{\sqrt{V}}\right)$$

where we have used the relationsips shown earlier for relating *E* and *V* to *df*. Or, simply

$$y_n \xrightarrow[df \to \infty]{} x_n .$$

To derive the unconstrained transformed data y_n (which will be fitted to a normal distribution) from the constrained data x_n, we invert the lhs of Eq. (14a) to get

$$y_n = E + \sqrt{V} \times \Phi^{-1}\left\{1 - ChiDist\left(\frac{df}{E}x_n, df\right)\right\} \tag{16}$$

where $df \approx 2E^2 / V$ (nearest integer > 0).

On the other hand, for synthesis, once the unconstrained variable y_n is generated from a multivariate MC synthesis, the corresponding constrained variable is computed by the inverse of Eq. (16):

$$x_n = \frac{E}{df} ChiInv\left\{1 - \Phi\left(\frac{y_n - E}{\sqrt{V}}\right), df\right\} \tag{17}$$

where again, $df \approx 2E^2 / V$ (nearest integer > 0).

Fractional Constraint: Variables Constrained by $0 \le x_n \le 1$

Variables which are constrained by $0 \le x_n \le 1$ are plausibly fitted by a Beta distribution. We can define a transformation by a condition on the cdfs

$$\Phi\left(\frac{y_n - E}{\sqrt{V}}\right) = BetaDist(x_n, a, b) \tag{18}$$

where

$$a = E\left\{\frac{E(1 - E)}{V} - 1\right\} \tag{19a}$$

and

$$b = (1-E)\left\{\frac{E(1-E)}{V} - 1\right\}. \tag{19b}$$

That is, both $\{x_n\}$ and $\{y_n\}$ have the same mean and variances. Note that, for a Beta distribution $V \le E(1-E)$, and it can be shown that

$$y_n \xrightarrow[V/[E(1-E)]\to 0]{} x_n \, .$$

When determining parameters of the model, one first uses original data to determine E and V, and thence, via Eqs. (19), a and b. This is then used to transform all of the original data points (each value in a column such as f1 in Table Y) according to

$$y_n = E + \sqrt{V}\Phi^{-1}\left[BetaDist(x_n, a, b)\right] \tag{20}$$

which follows from Eq. (18).

When synthesizing data, y_n will be generated, and it is transformed into the synthetic values of x_n by the inverse of Eq. (20):

$$x_n = BetaInv\left\{\Phi\left(\frac{y_n - E}{\sqrt{V}}\right), a, b\right\} \tag{21}$$

where a and b are determined from Eqs. (19).

We build a model by transforming T_var1 in Table IV according to Eq. (16) and $f1$ according to Eq. (20), before determining means and covariances of the transformed variables. Data was then synthesized by MC generation of transformed variable deviates which, for T_var1 and $f1$ were transformed back into the original variables using the inverse transformations Eqs. (17) and (21), respectively.

The synthesized data (1000 points) are analyzed in Fig. 9 and Fig 10. These figures are to be compared with the corresponding analysis of the original data in Figs. 6 and 7.

f1

Mean	0.8602552
Std Dev	0.0975378
Std Err Mean	0.0030844
upper 95% Mean	0.8663079
lower 95% Mean	0.8542025
N	1000

T_mean1

Mean	16.136787
Std Dev	5.8933379
Std Err Mean	0.131779
upper 95% Mean	16.395226
lower 95% Mean	15.878348
N	2000

T_var1

Mean	15.246895
Std Dev	12.085739
Std Err Mean	0.2702453
upper 95% Mean	15.776887
lower 95% Mean	14.716903
N	2000

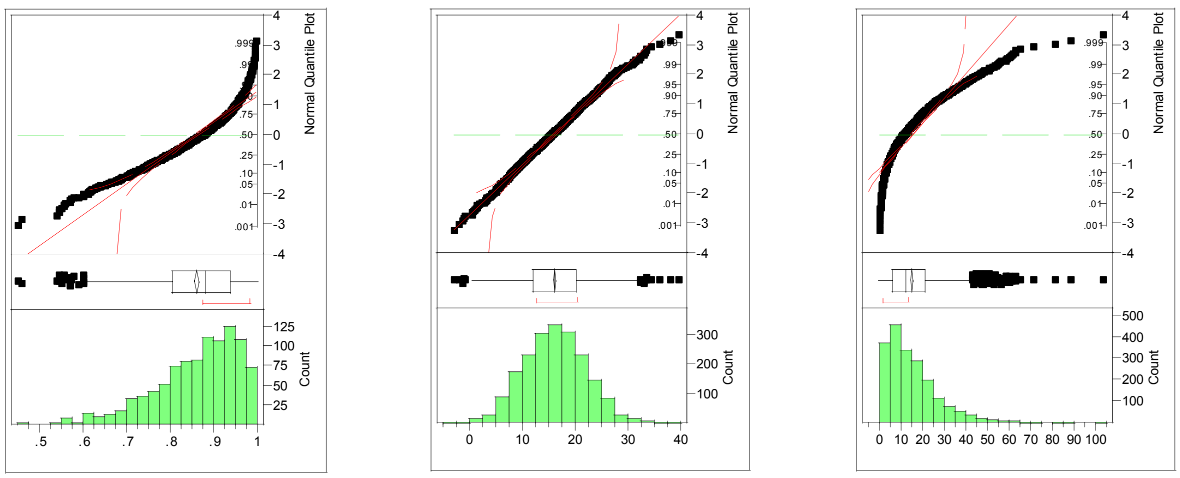

Fig. 9. Distributions of $f1$, T_mean1, and T_var1 synthesized using a Beta (for $f1$) and Chisquare (for T_var1) transformation of the data. The synthesized data automatically satisfies definitional constraints.

	f1	T_mean1	T_var1
f1	1.0000	0.0309	-0.0020
T_mean1	0.0309	1.0000	-0.6416
T_var1	-0.0020	-0.6416	1.0000

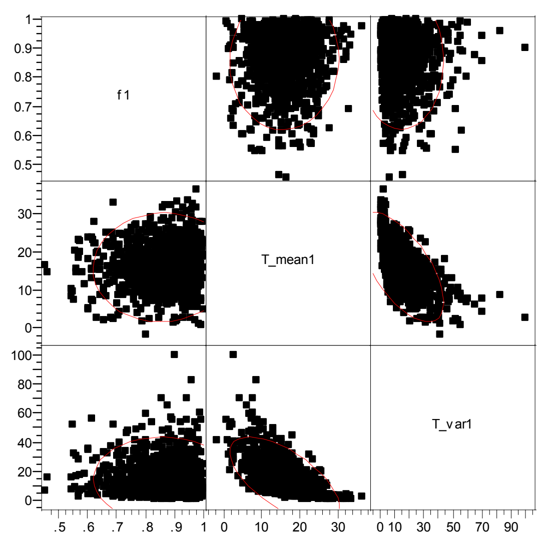

Fig. 10. Correlation of $f1$, T_mean1, and T_var1 synthesized using a Beta (for $f1$) and Chisquare (for T_var1) transformation of the data. This simulation reproduces the "flat-sidedness" of the correlations appropriate to the constraints.

Time-In-States Constraint

Models of time in state are often given in the form of the example in Table V where each variable is constrained to lie in the range $0 \le s_n \le 1$ (n = 1, N), but with the additional constraint $\sum_{n=1,N} s_n = 1$. Analysis of the data in Table V are shown in Figs. 11 and 12.

Table V. 160 points of time-in-state data (table has been folded 4-fold).

Total	Run	Idle	Off	Total	Run	Idle	Off	Total	Run	Idle	Off	Total	Run	Idle	Off
1.00	0.02	0.01	0.97	1.00	0.19	0.02	0.79	1.00	0.34	0.03	0.63	1.00	0.61	0.26	0.14
1.00	0.03	0.02	0.95	1.00	0.19	0.28	0.53	1.00	0.35	0.29	0.36	1.00	0.62	0.31	0.07
1.00	0.04	0.27	0.69	1.00	0.19	0.66	0.15	1.00	0.35	0.03	0.61	1.00	0.64	0.14	0.22
1.00	0.05	0.11	0.84	1.00	0.20	0.34	0.45	1.00	0.37	0.08	0.55	1.00	0.64	0.02	0.34
1.00	0.05	0.31	0.64	1.00	0.21	0.79	0.00	1.00	0.37	0.13	0.50	1.00	0.64	0.01	0.35
1.00	0.06	0.02	0.93	1.00	0.21	0.79	0.00	1.00	0.37	0.54	0.09	1.00	0.64	0.00	0.36
1.00	0.07	0.17	0.77	1.00	0.22	0.01	0.78	1.00	0.38	0.34	0.28	1.00	0.65	0.02	0.34
1.00	0.07	0.29	0.64	1.00	0.22	0.78	0.00	1.00	0.38	0.58	0.04	1.00	0.65	0.22	0.13
1.00	0.08	0.00	0.92	1.00	0.22	0.78	0.00	1.00	0.39	0.02	0.59	1.00	0.66	0.07	0.27
1.00	0.08	0.24	0.68	1.00	0.22	0.78	0.00	1.00	0.39	0.49	0.12	1.00	0.67	0.17	0.16
1.00	0.08	0.15	0.77	1.00	0.22	0.03	0.74	1.00	0.39	0.57	0.03	1.00	0.68	0.32	0.00
1.00	0.08	0.27	0.65	1.00	0.23	0.71	0.06	1.00	0.40	0.31	0.29	1.00	0.68	0.31	0.01
1.00	0.08	0.01	0.90	1.00	0.23	0.31	0.47	1.00	0.43	0.56	0.01	1.00	0.74	0.15	0.11
1.00	0.09	0.11	0.80	1.00	0.23	0.31	0.46	1.00	0.43	0.29	0.28	1.00	0.75	0.10	0.14
1.00	0.10	0.29	0.62	1.00	0.23	0.07	0.70	1.00	0.43	0.01	0.56	1.00	0.76	0.02	0.22
1.00	0.10	0.87	0.03	1.00	0.23	0.45	0.31	1.00	0.44	0.13	0.42	1.00	0.78	0.06	0.16
1.00	0.10	0.55	0.34	1.00	0.23	0.04	0.72	1.00	0.44	0.55	0.01	1.00	0.78	0.00	0.21
1.00	0.11	0.04	0.86	1.00	0.23	0.27	0.50	1.00	0.45	0.45	0.10	1.00	0.78	0.00	0.21
1.00	0.11	0.89	0.00	1.00	0.24	0.64	0.12	1.00	0.45	0.02	0.52	1.00	0.80	0.19	0.01
1.00	0.12	0.19	0.69	1.00	0.24	0.41	0.35	1.00	0.46	0.00	0.54	1.00	0.82	0.15	0.03
1.00	0.12	0.46	0.42	1.00	0.24	0.01	0.75	1.00	0.46	0.42	0.12	1.00	0.83	0.03	0.14
1.00	0.12	0.12	0.76	1.00	0.25	0.03	0.72	1.00	0.47	0.02	0.50	1.00	0.83	0.02	0.15
1.00	0.12	0.12	0.75	1.00	0.27	0.25	0.48	1.00	0.50	0.43	0.06	1.00	0.84	0.15	0.01
1.00	0.13	0.10	0.78	1.00	0.27	0.66	0.07	1.00	0.51	0.08	0.41	1.00	0.87	0.02	0.11
1.00	0.14	0.37	0.49	1.00	0.28	0.07	0.65	1.00	0.51	0.05	0.44	1.00	0.89	0.06	0.05
1.00	0.14	0.13	0.73	1.00	0.28	0.28	0.44	1.00	0.51	0.46	0.03	1.00	0.89	0.03	0.07
1.00	0.14	0.44	0.42	1.00	0.28	0.00	0.72	1.00	0.53	0.42	0.05	1.00	0.90	0.09	0.00
1.00	0.14	0.38	0.48	1.00	0.28	0.31	0.40	1.00	0.53	0.10	0.37	1.00	0.93	0.05	0.02
1.00	0.15	0.10	0.75	1.00	0.29	0.36	0.35	1.00	0.53	0.40	0.07	1.00	0.93	0.00	0.07
1.00	0.15	0.41	0.43	1.00	0.29	0.37	0.34	1.00	0.54	0.25	0.20	1.00	0.93	0.03	0.04
1.00	0.16	0.02	0.82	1.00	0.29	0.70	0.01	1.00	0.55	0.09	0.37	1.00	0.94	0.01	0.05
1.00	0.16	0.09	0.74	1.00	0.30	0.38	0.32	1.00	0.55	0.02	0.43	1.00	0.95	0.05	0.00
1.00	0.16	0.15	0.69	1.00	0.30	0.19	0.51	1.00	0.56	0.03	0.41	1.00	0.95	0.05	0.00
1.00	0.17	0.18	0.65	1.00	0.31	0.02	0.66	1.00	0.56	0.29	0.15	1.00	0.96	0.03	0.01
1.00	0.17	0.14	0.69	1.00	0.32	0.57	0.11	1.00	0.56	0.06	0.38	1.00	0.97	0.03	0.00
1.00	0.17	0.30	0.53	1.00	0.32	0.20	0.48	1.00	0.56	0.07	0.37	1.00	0.98	0.01	0.01
1.00	0.18	0.11	0.71	1.00	0.33	0.14	0.53	1.00	0.57	0.35	0.09	1.00	0.98	0.01	0.01
1.00	0.18	0.18	0.64	1.00	0.33	0.39	0.28	1.00	0.57	0.14	0.29	1.00	0.98	0.01	0.01
1.00	0.18	0.28	0.54	1.00	0.34	0.51	0.16	1.00	0.58	0.11	0.31	1.00	0.98	0.01	0.01
1.00	0.19	0.05	0.76	1.00	0.34	0.21	0.45	1.00	0.58	0.42	0.00	1.00	0.99	0.01	0.00

When

It is possible to convert N variables $\{s_n, n = 1, N\}$ into N-1 variables $\{x_n, 1, N\text{-}1\}$ each of which independently satisfy the "fractional constraint" of the previous section, and may therefore be fitted by the Beta distribution, and synthesized there from. We next show the required transformation, and its inverse.

<u>Constrained to Unconstrained Transformation</u>

For 3 variables, with example values $s_1 = 0.1$, $s_2 = 0.6$, $s_3 = 0.3$ we can write

$$x_1 = \frac{s_1}{s_1 + s_2 + s_3} = \frac{s_1}{1} = 0.1 \tag{22a}$$

$$x_2 = \frac{s_2}{s_2 + s_3} = \frac{0.6}{0.6 + 0.3} = 0.66667 \tag{22b}$$

$$x_3 = \frac{s_3}{s_3} = 1 \tag{22c}$$

In general the transformation to unconstrained fractional variables is

$$x_n = \frac{s_n}{\sum_{j=n,N} s_j} . \tag{23}$$

The variables $\{x_n\}$ are individually constrained according to $0 \le x_n \le 1$ but have no other constraint. Therefore the statistical parameter extraction and synthesis of unconstrained (but physically hard to interpret) variables will follow the methods of the fractional constraint above. After synthesis, the unconstrained variables may be converted (back) into the physically more meaningful constrained fractional variables by the following transformation.

<u>Unconstrained to Constrained Transformation</u>

The inverse of Eqs. (22), with example values, is

$$s_1 = x_1 = 0.1 \tag{24a}$$

$$s_2 = x_2 \times (1 - s_1) = 0.66667 \times (1 - 0.1) = 0.6 \tag{24b}$$

$$s_3 = x_3 \times (1 - s_1 - s_2) = 1 \times (1 - 0.1 - 0.6) = 0.3 \tag{24c}$$

In general, the inverse transformation is

$$s_1 = x_1 \tag{25a}$$

and then

$$s_n = x_n \times \left(1 - \sum_{j=1}^{n-1} s_j \right) \quad n = 2, N \tag{25b}$$

The transformed variables will have the constraint that $0 \le x_n \le 1$, n = 1, N-1, but there is no constraint on the sum of these variables.

When the data in Table V is transformed according to the transformation of Eq. (22), fitted to independent time-in-state distributions, and synthesized as in the "Fractional Constraint" section above, the synthesized data (1000 points) is as shown in Figs. 13 and 14. These figures are to be compared with Figs. 12, and 13.

Fig. 15 is a 3D plot which shows an interesting comparison between the original Run/Idle/Off data (160 points), and 1000 points of synthesized data. Points constrained by Run + Idle + Off fractions lie on the 111 plane in the all positive octant of the 3D plot.

Run	
Mean	0.4131043
Std Dev	0.276089
Std Err Mean	0.0218268
upper 95% Mean	0.4562121
lower 95% Mean	0.3699966
N	160

Idle	
Mean	0.2249357
Std Dev	0.2206493
Std Err Mean	0.0174439
upper 95% Mean	0.2593872
lower 95% Mean	0.1904841
N	160

Off	
Mean	0.36196
Std Dev	0.2851428
Std Err Mean	0.0225425
upper 95% Mean	0.4064814
lower 95% Mean	0.3174386
N	160

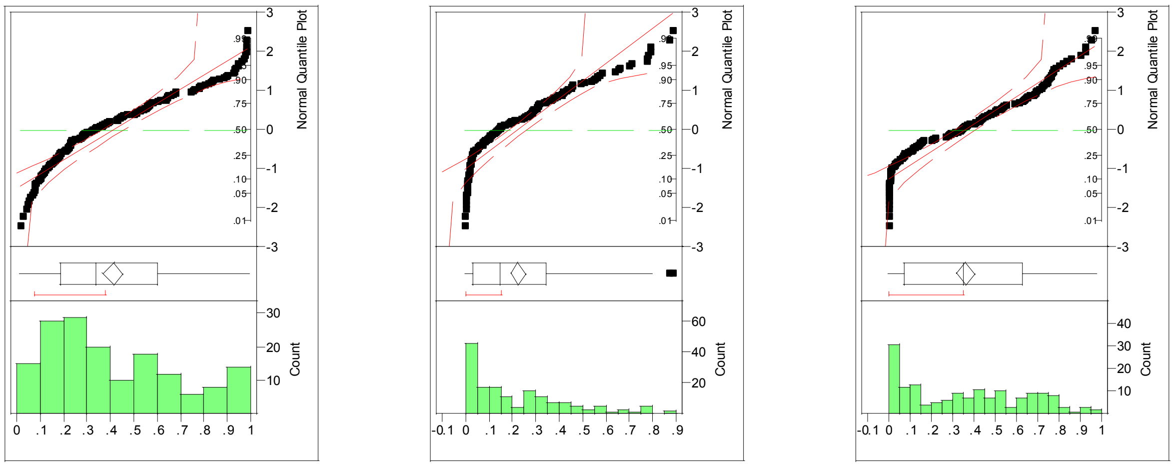

Fig. 11. Distributions of times in Run, Idle and Off state data in Table V.

	Run	Idle	Off
Run	1.0000	-0.3579	-0.6913
Idle	-0.3579	1.0000	-0.4273
Off	-0.6913	-0.4273	1.0000

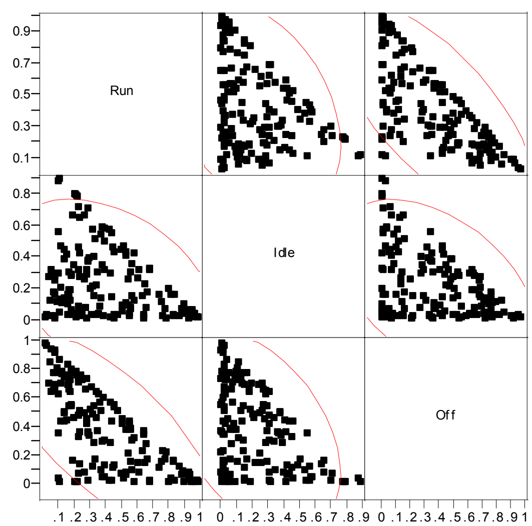

Fig. 12 Correlations of Run, Idle and Off state data in Table V. The triangular shapes reflect the constraint Run + Idle + Off = 1

	Run	Idle	Off
Mean	0.4079639	0.2242344	0.3678017
Std Dev	0.2697937	0.2220097	0.2856076
Std Err Mean	0.0085316	0.0070206	0.0090317
upper 95% Mean	0.4247059	0.2380112	0.385525
lower 95% Mean	0.391222	0.2104577	0.3500784
N	1000	1000	1000

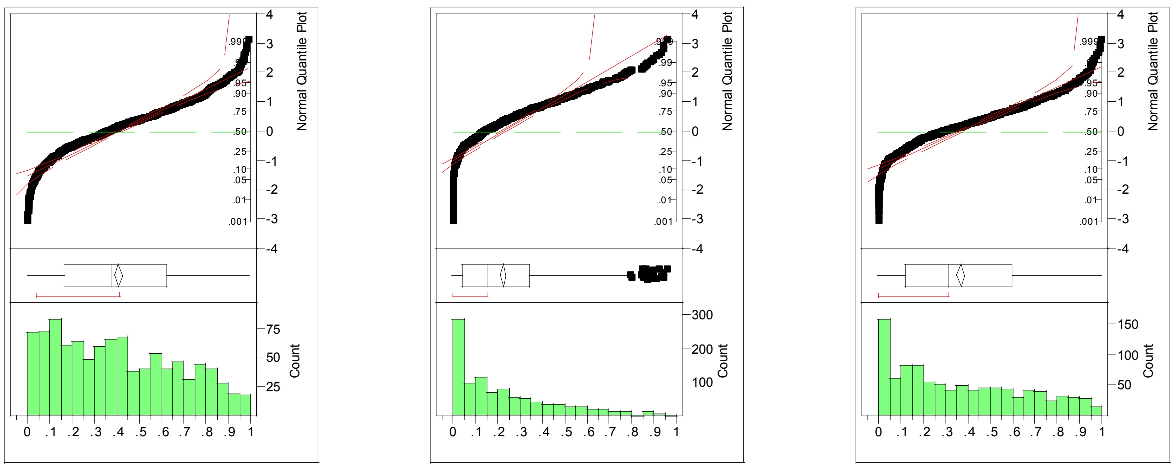

Fig. 13 Distributions of synthesized of Run/Idle/Off data. This is to be compared with Fig. 11.

	Run	Idle	Off
Run	1.0000	-0.3381	-0.6818
Idle	-0.3381	1.0000	-0.4579
Off	-0.6818	-0.4579	1.0000

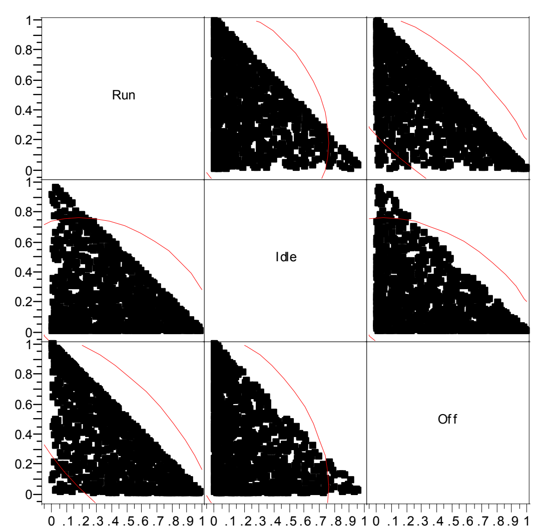

Fig. 14. Correlations of synthesized Run/Idle/Off data. This is to be compared with Fig. 12. The diagonal boundaries reflect the Run + Idle + Off = 1 constraint.

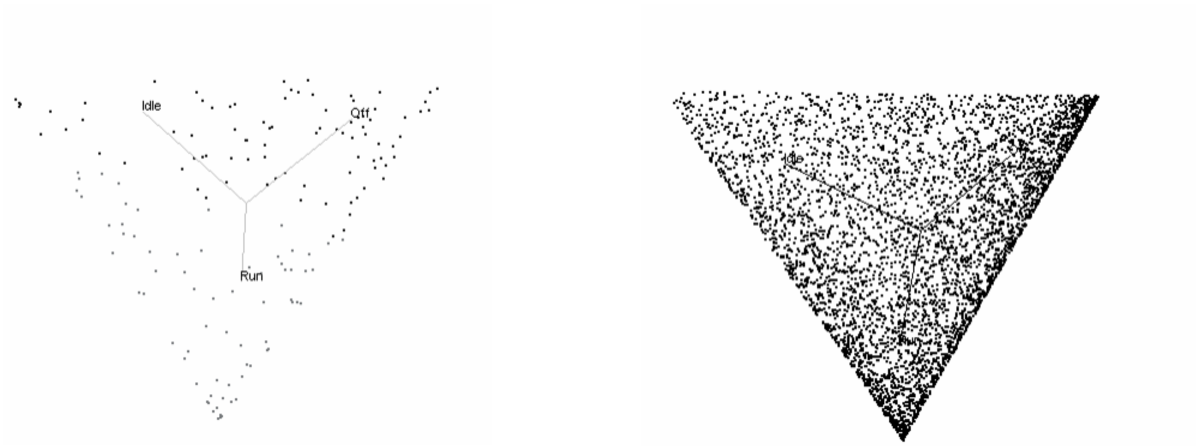

Fig 15. 3D plot of Run/Idle/Off original data, and synthesized data. Notice the concentration along the Run/Off axis.