Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A.A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

A NON-PARAMETRIC APPROACH TO SIMULATE PANEL DATA

Bahram Yousefi
Mohanad Ajina
Muhammad Imran
Kathryn Laskey

George Mason University
4400 University Drive
Fairfax, VA 22030, USA

ABSTRACT

Real world multivariate data mostly contains correlation structure because generally some variables tend to have a similar behavior or some dependency structure. This is due to the nature of data generation process. The variables can contain cross-sectional correlation, correlation between variables at a given time stamp, and temporal correlation, correlation between various observations of a given variable at various timestamps. Also, there are datasets which contain both cross-sectional and temporal correlations. Modeling of such correlation structure in the data is important because it provides us the predictive power and affects the Machine Learning algorithms in various ways, due to relationships within variables. We propose a methodology to simulate a cyber activity dataset, containing both cross-sectional and temporal correlations, also called Panel dataset. The proposed methodology uses non-parametric approach to induce correlations among feature vectors and across time without disturbing the marginal distributions of features.

1 INTRODUCTION

Simulating data is a powerful method to synthesize a natural phenomenon or to estimate the future state of a data generation process. For example, data simulation is widely used in economics or finance studies for risk mitigation (Chan and Wong 2015). Environmental studies is also a resort of data simulation to measure the effect of various factors involved in most of the natural process (Villanueva et al. 2012). Based on certain assumptions and factors analysts can setup a data simulation process to conduct multiple experiments and study the outcomes of those experiments without actually waiting for an event to happen in real-world. Today, it is feasible to conduct such studies because of the computation power to simulate a future event multiple times and quantify its effect.

The presented case study contains multiple variables that are dependent, or are correlated, among each other. This dependency between the variables is due to the nature of the problem and definition of these variables. For example, a variable also called feature, representing number of visits to job search websites has a correlation of 0.39 with another variable representing number of visits to professional networking websites. For instance, in real world, this dependency can be also observed between two variables when a person who is looking for a job is also likely to have professional networking behavior. This nature of dependency between variables can have an impact on the performance of machine learning algorithms (Nicodemus and Malley 2009) through which this data is passed in later stage of our project. The discussion about why and how we use machine learning algorithms on the correlated features set achieved from this data simulation process is beyond the scope of this paper. This work focuses on constructing a dataset using provided marginal distributions, cross-sectional and temporal Spearman rank correlations for Inference Enterprises Modeling (IEM). One important component for the IEM is the input data, or population synthesis (Huang et al. 2018). The ideal case for IEM is when real and complete data is used. However,

978-1-5386-6572-5/18/$31.00 ©2018 IEEE

this is not the case all the time. Due to privacy preserving constraints or classified information, the input data for IEM is redacted and no longer a direct representation of ground truth data. In case of the redacted data, we are given summary data in form of Means, Standard Deviation (STD), and Cross-Sectional (CSC) and Temporal Correlation (TC). This paper proposes a panel data reconstruction methodology. The panel data in this context means a dataset which contains both cross-sectional and temporal correlations.

To this end, the proposed methodology has a two-step process. The first step is to generate multivariate standard Gaussian correlated data set that satisfies the cross-sectional and temporal correlations, the second step is to transfer the correlation from this simulated correlated dummy data to uncorrelated feature vectors of interest. The key factor here is that marginal distributions of features are preserved while achieving the target rank correlation structure. Also, the approach generalizes for any size of dataset because it can be modularized as we have explained in the following sections.

2 LITERATURE REVIEW

Since early 70s, there is interest in generating random correlated data points, variables, for different purposes (Park et al. 1996). For example, in finance or economics most of the time analysts are interested in knowing about the future of stocks. Risk analysts are keen about knowing whether the set of stocks are highly correlated and what are the chances that all the stocks they have in their portfolio will crash together (Chan and Wong 2015). To simulate such a phenomenon it is important to simulate data which contains the dependency structure, crucial to answer the risk analyst’s questions.

Population synthesis methods such as synthetic reconstruction, combinatory optimization, sample free fitting, Markov Chain Monte Carlo simulation, and sample-free methods are all probability based approaches. These methods are computationally expensive even for generating a small disruption (Ye et al. 2017). Also, a nonlinear optimization method is proposed to generate disruption given redacted data such as mean, standard deviation, histogram and correlation (Yousefi et al. 2018). Another approaches partially use the conventional method of using Cholesky Factorization technique to correlate the data but one limitation of these approaches is that they disturb the means and variances of the individual features because this approach inherently performs matrix multiplication. This change in distributions is undesirable in many cases.

In addition, a Copulas approach is used to generate correlated variates. Copulas are better at inducing dependency structure at the tails mostly (Embrechts et al. 2003). Out of the box, Copulas assume the variables to be continuous type. In our case we have mix of continuous and discrete distributions to be correlated based on the given correlation matrices. One problem we faced is when using Copula that it has the capability to produce cross-sectional correlated data or temporal correlated data at a given instance. It does not generate a dataset which contains both cross-sectional and temporal correlations simultaneously in the dataset, as shown in the methodology proposed by Mathworks (“Generate Correlated Data Using Rank Correlation” 2017). This methodology can be used to achieve both type of correlation simultaneously if we also have cross-correlation function of the different variables at different time lags, in addition to available data such as cross-sectional and temporal correlations. In our problem, we are not provided with cross-correlation function which can be used to reconstruct the complete correlation matrix that captures the dependency within variables (cross-sectional correlation), dependency between observations of a particular feature at different times (temporal correlation) and how the observations of various features relate with each other at different instances of time (cross-correlation function). Due to the limitation of the provided data we did not continue development with Copula based approaches to induce two dimensional correlations, i.e., across features and across time.

3 METHODOLOGY

This methodology is applicable to the situation when the available data is marginal distributions of features which neither have cross-sectional correlation nor have temporal correlation. From various previous works, (Yoo et al. 2015) it is proven that to construct a model which produces simulated data representative of the real-world data, it is important to match statistics like cross-sectional correlation and temporal correlation. Inducing these correlations is a step forward towards generating the actual data from summarized data, in our case study. Cholesky Decomposition is utilized as a part of the process to simulate data which contains cross-sectional correlation and temporal correlations. The whole process of generating correlated multivariate distributions, step by step is explained in the following steps of the algorithm:

- **Step 1: Generating Multivariate Normally Distributed Random Vectors**
  The first step is to generate X = (X_1, . . ., X_n). Where X ~ MN_n(0, Σ). Let Z = (Z_1, . . ., Z_n)^T , where the Z_i's are I.I.D N(0, 1) for i = 1, . . ., n. If C is a (n × m) matrix then it follows that C^TZ ~ MN(0, C^TC). Therefore, the problem reduces to finding C such that C^TC = Σ. The Cholesky decomposition of Σ can be used to find such a matrix, C.
- **Step 2: The Cholesky Decomposition of a Symmetric Positive-Definite Matrix**
  One of the applications of Cholesky Decomposition is to correlate identically independent random variables given a correlation structure (Haugh 2017). From Linear Algebra it is known that any symmetric positive-definite matrix, M, can be written as M = U^TDU, where U is an upper triangular matrix and D is a diagonal matrix with positive diagonal elements. Since the assumption is that Σ is symmetric positive-definite, thus Σ can be written as Σ = U^TDU = (U^T√D)(√DU) = (√DU)^T(√DU). Therefore, the matrix C = √DU satisfies C^TC = Σ. It is called the Cholesky Decomposition of Σ.
- **Step 3: Correlating Normal Random Variables**
  Now that the Cholesky Decomposition of Σ is computed, call it matrix C. The matrix multiplication of C^T with Z (from the previous step) can be performed as X = C^TZ. Now, X will contain the required correlation structure.Step 2 and step 3 are repeated on the same data because first pass induces the temporal correlations and second pass induces the cross-sectional correlation on the same temporally correlated. Once the dataset has both type of correlations, this dataset is used to capture its correlations and transfer it to the actual uncorrelated feature distributions.
- **Step 4: Transferring the rank correlation**
  Correlated standard Gaussian samples are obtained from step 3 which hold the required correlation structure provided. The uncorrelated feature vectors retrieved using summary statistics and the given histograms, is the actual data of interest which has to be correlated. The aim of this step is to transfer the Spearman rank correlation from data generated in step 3 to these uncorrelated feature vectors. The Spearman correlation is defined as "In statistics, Spearman's rank correlation coefficient or Spearman's rho is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables) ("Spearman's Rank-Order Correlation" n.d.)." The Spearman correlation is calculated using the Equation 1.

$$Spearman\,Correlation = 1 - \frac{6\sum d^2}{n^3 - n} \quad (1)$$

  where "d" is the difference between the rank of a data pair, and "n" is the number of data points in the data set. In order to achieve similar Spearman correlation, the uncorrelated data generated should be reordered, such that it matches the rank order of data from step 3. The following example portrays the mechanism of the approach to transfer rank correlations from correlated dataset to uncorrelated dataset in three sub-steps. For this example, suppose that data from step 3 carries Spearman rank correlation of 0.7 between two dummy variables, say feature 1 and feature 2.
- **Step 4.a.)** This step starts by sorting the data from step 3, which is a panel data (the dataset which has cross-sectional and temporal correlations) containing the required rank correlation structure, in an ascending order and retain the indices of the data points. Figure 1 includes Table 1, Table 2 and Table 3 all of which show a small example of the sub-step.

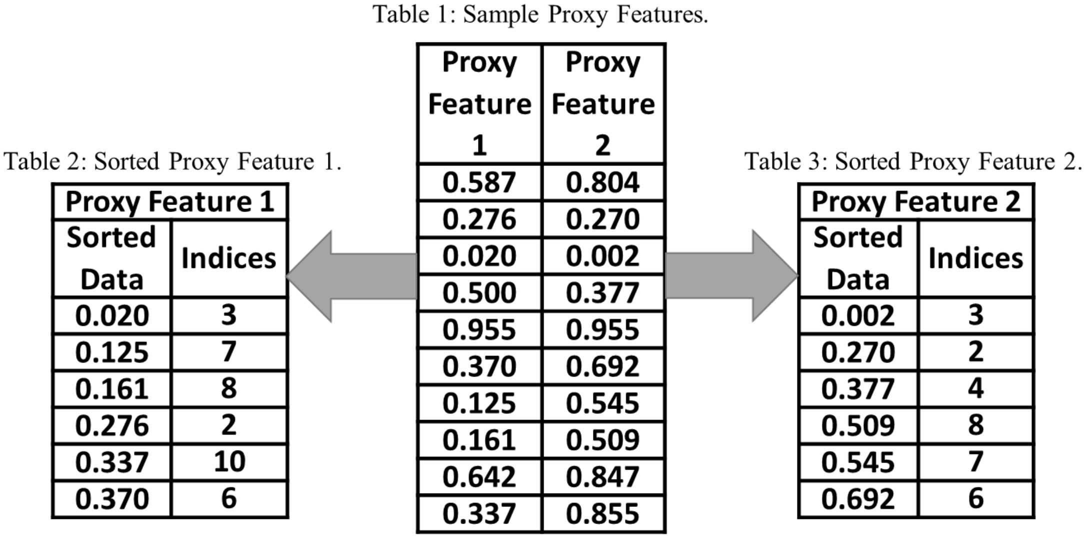


Table 1: Sample Proxy Features.

Proxy Feature 1	Proxy Feature 2
0.587	0.804
0.276	0.270
0.020	0.002
0.500	0.377
0.955	0.955
0.370	0.692
0.125	0.545
0.161	0.509
0.642	0.847
0.337	0.855

Table 2: Sorted Proxy Feature 1.

Proxy Feature 1	
Sorted Data	Indices
0.020	3
0.125	7
0.161	8
0.276	2
0.337	10
0.370	6

Table 3: Sorted Proxy Feature 2.

Proxy Feature 2	
Sorted Data	Indices
0.002	3
0.270	2
0.377	4
0.509	8
0.545	7
0.692	6

Figure 1: A sample rank correlation structure.

- **Step 4.b.)** Now, the actual features are sorted in ascending order. Figure 2 shows Table 4 and Table 5 to demonstrate an example of two features and the sorted features.

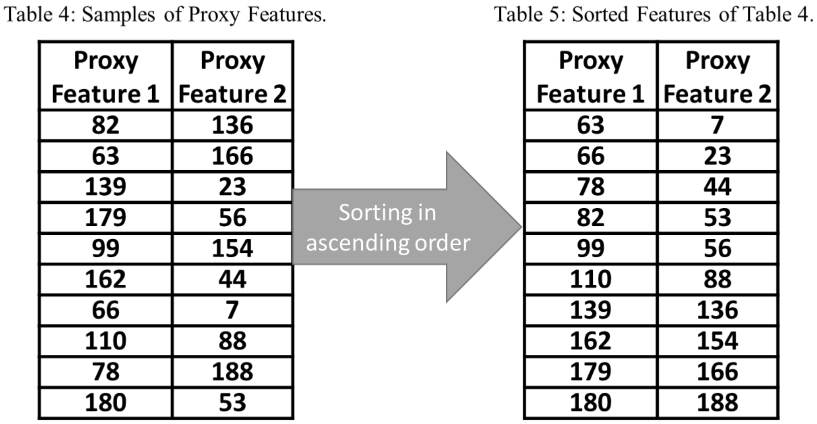


Table 4: Samples of Proxy Features.

Proxy Feature 1	Proxy Feature 2
82	136
63	166
139	23
179	56
99	154
162	44
66	7
110	88
78	188
180	53

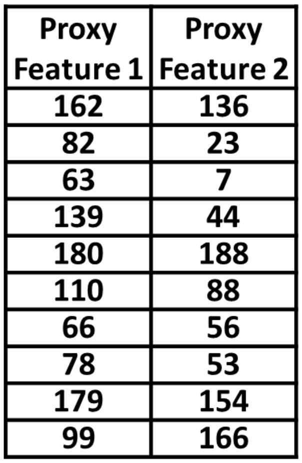
Sorting in ascending order

Table 5: Sorted Features of Table 4.

Proxy Feature 1	Proxy Feature 2
63	7
66	23
78	44
82	53
99	56
110	88
139	136
162	154
179	166
180	188

Figure 2: Example of sorting ranks.

- **Step 4.c.)** The last sub-step is reordering the sorted actual features according to retained indices from step 4.a as shown in Figure 3.



Proxy Feature 1	Proxy Feature 2
162	136
82	23
63	7
139	44
180	188
110	88
66	56
78	53
179	154
99	166

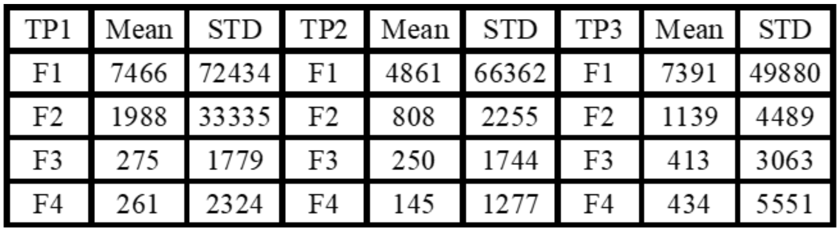
Figure 3: A result of Step 4.c.

Finally, to illustrate that executing the above process induces the required Spearman rank correlation. We calculate the correlation based on the data obtained at step 4.c (see Figure 3) which is 0.709.

4 CASE STUDY

In this section, a simplified problem will be demonstrated numerically and graphically given three-time periods and four features measured at each time period to generate 1000 observations. Also later in the section, the results are compared to the desired correlations to demonstrate the viability of the proposed method.

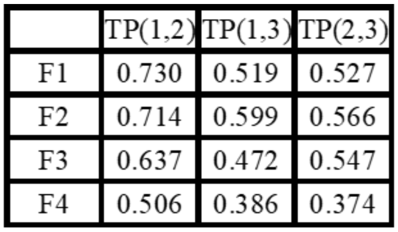
The means and standard deviations (SD) of feature distributions before any treatment are shown in Figure 4, where "TP" stands for Time Period and "F" stands for a feature number.



TP1	Mean	STD	TP2	Mean	STD	TP3	Mean	STD
F1	7466	72434	F1	4861	66362	F1	7391	49880
F2	1988	33335	F2	808	2255	F2	1139	4489
F3	275	1779	F3	250	1744	F3	413	3063
F4	261	2324	F4	145	1277	F4	434	5551

Figure 4: Features' Mean and SD used in the Case Study.

The Temporal Correlation (TC) and the Cross-Sectional Correlation (CSC) for all features and for each time period is also provided as part of the problem and are shown in Figure 5 and Figure 6.



	TP(1,2)	TP(1,3)	TP(2,3)
F1	0.730	0.519	0.527
F2	0.714	0.599	0.566
F3	0.637	0.472	0.547
F4	0.506	0.386	0.374

Figure 5: Features' Temporal Correlation (TC) used in the Case Study.

For example, the aim is to generate 1000 observations. We start with simulating a multivariate standard Gaussian distribution of the same dimension as shown in Figure 11, acting as the proxy distributions for each feature in each time period.

Next step is to multiply the proxy distributions by Cholesky decomposition of temporal correlation matrix to obtain temporally correlated distributions. Then these temporally correlated distributions are multiplied with Cholesky decomposition of Cross-correlation matrix to obtain a dataset which contains both temporal and cross-sectional correlation. Once this kind of data is obtained, the rank correlations of these proxy distributions are transferred to the actual feature distributions which are the output of primary interest in this case study.

TP1	F1	F2	F3	F4	TP2	F1	F2	F3	F4
F1	1	0.698	0.569	-0.015	F1	1	0.816	0.508	-0.016
F2	0.698	1	0.626	-0.06	F2	0.816	1	0.499	-0.054
F3	0.569	0.626	1	-0.037	F3	0.508	0.499	1	-0.027
F4	-0.015	-0.06	-0.037	1	F4	-0.016	-0.054	-0.027	1
TP3	F1	F2	F3	F4	TP4	F1	F2	F3	F4
F1	1	0.598	0.271	-0.019	F1	1	0.815	0.683	-0.013
F2	0.598	1	0.19	-0.066	F2	0.815	1	0.773	-0.042
F3	0.271	0.19	1	-0.045	F3	0.683	0.773	1	-0.054
F4	-0.019	-0.066	-0.045	1	F4	-0.013	-0.042	-0.054	1

Figure 6: Features' Cross-Sectional Correlation (CSC) used in the Case Study.

After applying the described methodology, the Spearman rank correlations within the distributions are calculated and shown below. It can be seen from Figure 7 and Figure 8 that the relationship between given temporal correlation and the temporal correlation calculated from data after applying the methodology is linear that means the treated dataset has the temporal correlation close to the given one. This is desired result of the proposed process.

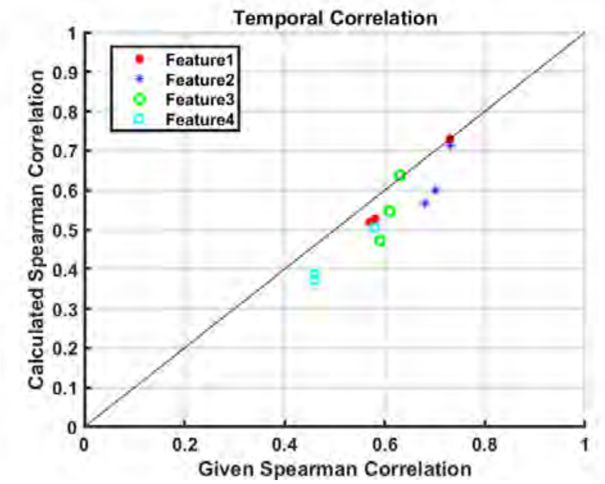


Figure 7: Correlation among four features shown, in each of the three time periods.

Furthermore the treated data can be tested for Cross-sectional correlations. The following graph shows a close to linear relationship between the Cross-sectional correlation calculated from the treated data and the given Cross-sectional correlations. The linear nature of relationship proves that the methodology produces the desired result which is to achieve Cross-sectional correlations which closely match the given Cross-sectional correlations.

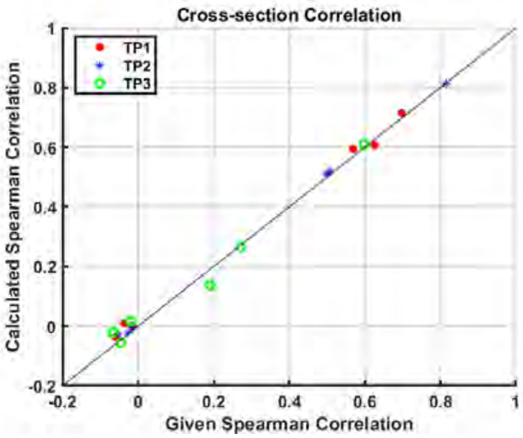


Figure 8: Correlation among three time periods shown, for each of the four features.

After treating the data with correlation inducing methodology presented in previous sections, the means and standard deviations of the feature distributions can be recalculated to make sure that there is no distortion caused in the data due to correlation induction. In this case the means and standard deviations can be retrieved from the treated as given in Figure 10, this implies that data distributions did not get shifted throughout the process.

5 DISCUSSION

A methodology is presented to generate observations for datasets with both temporal and feature set correlations. For example, considering the problem of Inference Enterprises Modeling (IEM) for a particular organization (e.g., organization A in the IARPA SCITE project). An IEM represents the automated portions of existing and proposed insider threat detection enterprises and forecasts the performance of those enterprises in detecting a threat ("Scientific Advances to Continuous Insider Threat Evaluation (SCITE)" n.d.). Due to privacy concerns, individual private information cannot be provided (e.g., email activity data of individuals), but instead summary statistics such as mean, SD, CSC and ST for 3000 employees (observations), 8 months (time periods) and 24 detectors (features) are provided. The dataset has 576000 activity data points in total. Reconstruing the dataset given the temporal and cross-feature correlation data using optimization-based approaches are accurate yet extremely computationally expensive. However, using the proposed methodology, the dataset is reconstructed in a more efficient manner with CSC and ST information that are close to the desired characteristics (i.e., provided summary statistics) as shown in Figure 9 and Figure 10. Consequently, The main significance of the methodology is its computational efficiency while producing accurate results. Therefore, this methodology is recommended for synthesizing large-scale datasets with both temporal and cross-feature correlation information.

6 CONCLUSIONS

Based on the results presented in the case study, the advantage of this approach can be seen, which is it does not distort or shift the marginal data distributions. Due to non-parametric nature of the methodology, it does not make any assumptions about the actual distributions of features and the objective of data possessing both temporal and cross-sectional correlations simultaneously is also achieved. This is the primary objective of the whole methodology. Also, the approach does not involve intense mathematical computations which can be a limitation to scalability. Furthermore, it is easy to infer from discussion of the methodology that it is not domain specific and can be used to address any kind of data simulation which requires the generation of panel data while strictly preserving the data distributions.

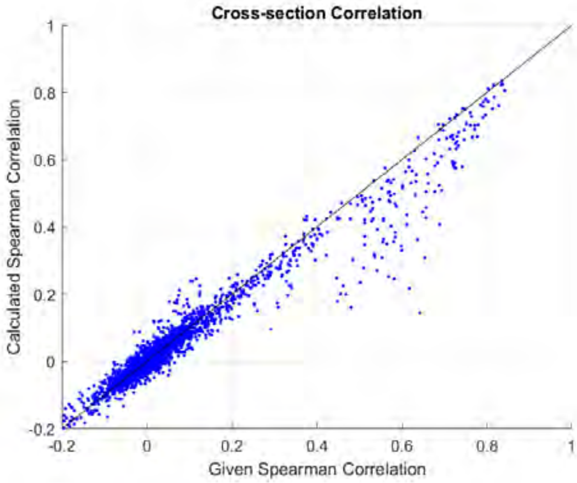


Figure 9: Correlation among all features shown for all time periods.

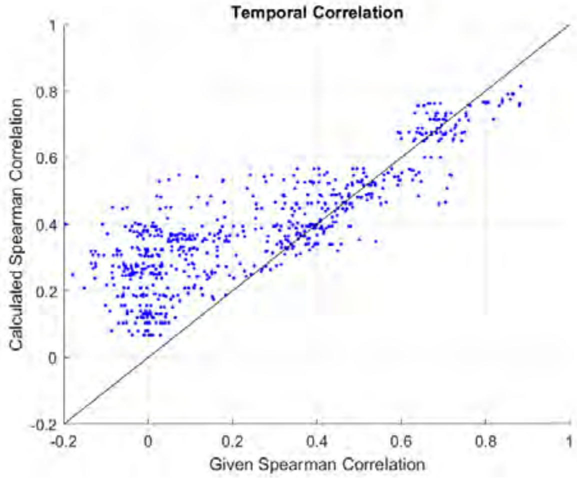


Figure 10: Correlation among all time periods shown for all features.

ACKNOWLEDGMENTS

Research project used in this paper was supported under IARPA contract 2016 16031400006. The content is solely the responsibility of the authors and does not necessarily represent the official views of the U.S. Government.

REFERENCES

Chan, N. H., and H. Y. Wong. 2015. “Simulation Techniques in Financial Risk Management, 2nd Edition”. *Journal of Risk and Insurance* 76:455–58.

Embrechts, P., F. Lindskog, and A. Mcneil. 2003. “Modelling Dependence with Copulas and Applications to Risk Management”. In *Handbook of Heavy Tailed Distributions in Finance*, 329–84. Elsevier.

MATLAB and Simulink. 2017. “Generate Correlated Data Using Rank Correlation.” Accessed August 4, 2018. https://www.mathworks.com/help/stats/generate-correlated-data-using-rank-correlation.html.

Haugh, M. 2017. “Generating Random Variables and Stochastic Processes”. Accessed August 11, 2018. http://www.columbia.edu/~mh2078/MonteCarlo/MCS_Generate_RVars.pdf.

Huang, E., A. K. Zaidi, and K. B. Laskey. 2018. “Inference Enterprise Multimodeling for Insider Threat Detection Systems”. In *Disciplinary Convergence in Systems Engineering Research*, edited by Azad M. Madni, Barry Boehm, Roger G. Ghanem, Daniel Erwin, and Marilee J. Wheaton, 175–86. Cham: Springer International Publishing.

Nicodemus, K. K., and J. D. Malley. 2009. “Predictor Correlation Impacts Machine Learning Algorithms: Implications for Genomic Studies”. *Bioinformatics* 25(15):1884–90.

Park, C. G., T. Park, and D. W. Shin. 1996. “A Simple Method for Generating Correlated Binary Variates”. *The American Statistician* 50(4):306–310.

IARPA. “Scientific Advances to Continuous Insider Threat Evaluation (SCITE)”. https://www.iarpa.gov/index.php/working-with-iarpa/index.php?option=com_content&view=article&id=517&Itemid=324. Accessed August 5, 2018.

Laerd Statistics. “Spearman’s Rank-Order Correlation”. Accessed August 4, 2018. https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php.

Villanueva, D., A. Feijoo, and J. L. Pazos. 2012. “Simulation of Correlated Wind Speed Data for Economic Dispatch Evaluation”. *IEEE Transactions on Sustainable Energy* 3(1):142–49.

Ye, P., X. Hu, Y. Yuan, and F. Wang. 2017. “Population Synthesis Based on Joint Distribution Inference Without Disaggregate Samples”. *Journal of Artificial Societies and Social Simulation* 20(4):1-16.

Yoo, W., R. Mayberry, S. Bae, and K. Singh. 2015. “A Study of Effects of Multi-Collinearity in the Multivariable Analysis”. *International journal of applied science and technology* 4(5):9.

Yousefi, B., M. Imran, and K. B. Laskey. 2018. “Synthesizing Inference Enterprises from Redacted Data”. In *Conference on Systems Engineering Research*, May 8-9, Charlottesville, Virginia, USA.

AUTHOR BIOGRAPHIES

BAHRAM YOUSEFI is Research Assistant Professor at the C4I and Cyber Center at George Mason University (GMU). He has experience working on various architecture design and evaluation projects at System Architectures Laboratory (at GMU). Moreover, he studied extensively the Insider Threat detections mechanisms in the Inference Enterprise project in both Sensor Fusion Laboratory and C4I and Cyber Center. His research areas are Resilience of Enterprises and Inference Enterprise Modeling through Multi-modeling. His email address is byousefi@gmu.edu.

MOHANAD AJINA is PhD candidate in Electrical and Computer Engineering at George Mason University. He received his undergrad from Southern Illinois University Carbondale in 2014. In the same year, he joined George Mason University to do his Master in Electrical Engineering and he graduated in 2016. His email address is majina@gmu.edu.

MUHAMMAD IMRAN is a Faculty Research Associate for the C4I & Cyber Center at George Mason University (GMU) and leads the statistical modeling & machine learning team for the effort described in this paper. He completed his Masters in Data Analytics Engineering with outstanding academic achievement award in 2017 from GMU. His research interests are advanced data analytics, statistical modeling, machine learning, deep learning, and human augmented intelligence. His email address is mimran4@gmu.edu.

KATHRYN BLACKMOND LASKEY is Professor of Systems Engineering and Operations Research and Associate Director of the C4I Center at George Mason University. She is Secretary of the Washington Metropolitan Chapter of INCOSE, faculty advisor of the GMU student chapter of INCOSE, member of the Board of the International Society of Information Fusion, and Board chair of the Association for Uncertainty in Artificial Intelligence. Her primary research areas are information fusion and semantic interoperability in uncertainty rich domains. Her PhD in Statistics and Public Policy is from Carnegie Mellon University. Her email address is klaskey@gmu.edu.

#	F1 TP1	F2 TP1	F3 TP1	F4 TP1
1	12005.56	3568.926	399.058	129.096
2	5373.867	997.793	640.667	566.678
3	22331.69	5160.127	304.294	165.484
4	6791.730	3193.064	324.875	236.008
5	156.620	3289.484	236.910	746.927
⋮	⋮	⋮	⋮	⋮
999	10792.790	610.702	100.272	633.620
1000	19807.300	5153.101	205.731	388.239
#	F1 TP2	F2 TP2	F3 TP2	F4 TP2
1	69.566	8613.841	582.662	307.366
2	85.780	3210.636	1433.105	41.184
3	51.233	3647.968	1269.162	86.081
4	78.016	7948.733	1330.789	362.545
5	21.093	9309.837	1086.799	3.728
⋮	⋮	⋮	⋮	⋮
999	12.056	2714.418	2018.157	548.151
1000	18.025	9217.595	1913.166	310.618
#	F1 TP3	F2 TP3	F3 TP3	F4 TP3
1	244.071	53.320	10357.750	2135.989
2	419.443	10.890	20952.050	807.276
3	259.714	42.803	21559.770	1396.667
4	27.486	29.668	6526.631	2032.048
5	337.248	117.577	2128.528	2150.078
⋮	⋮	⋮	⋮	⋮
999	268.376	10.137	15514.76	829.427
1000	71.296	132.321	16743.67	2380.318

Figure 11: 1000 Observations simulated using proposing methodology.