

2ND

## ANNUAL CONFERENCE ON THE **FRONTIERS OF STATISTICS**

# DATA SCIENCES ACROSS DISCIPLINES



May 11th - 12th University of South Florida (CWY 108)

For Additional Information Call (813) 974-9734

## AGENDA OVERVIEW

### Friday, May 11, 2018

9.00 am – 9.30 am	Registration & Reception
9.30 am – 9.50 am	Opening Remarks
9.50 am - 10.40 am	Cyber-Security, Invited speaker
10.40 am – 10.55 am	Break
10.55 am – 11.55 am	Cyber-Security, Presentations
12.00 pm – 1.00 pm	Lunch Break
1.00 pm – 2.00 pm	Cyber-Security, Presentations
2.00 pm – 2.10 pm	Break
2.10 pm – 2.55 pm	Health Science, Invited Speaker
2.55 pm – 3.05 pm	Break
3.05 pm – 3.45 pm	Health Science, Presentations
3.45 pm – 3.55 pm	Break
3.55 pm - 5.00 pm	Health Science, Invited Speaker / Presentations

### Saturday, May 12, 2018

Saturday, May 12, 2018			
9.30 am – 10.00 am	Registration		
10.00 am – 11.00 am	Business Analytics, Presentations		
11.00 am – 11.15 am	Break		
11.15 am – 12.15 pm	Environmental Sciences, Presentations		
12.15 pm – 1.15 pm	Lunch Break		
1.15 pm – 1.55 pm	Environmental Sciences, Presentations		
1.55 pm – 2. 10 pm	Break		
2.10 pm - 3. 10 pm	Advances in Statistics, Presentations		
3.10 pm – 3.30 pm	Closing Remarks (Prof. Tsokos)		

## Friday, May 11- CWY 108

#### 9:00 am - 9:30 am Registration and Reception

9.30 am – 9.50 am Opening Remarks: Dr. Chris P. Tsokos

Distinguished University Professor, USF, Department of Mathematics and Statistics

IFNA President

Executive Director of USOP

Chief-Editor, Probability & Statistics, Atlantis Press, Springer

## Cyber-Security

**Chair: K. Ruwani M. Fernando** Department of Mathematics & Statistics, University of South Florida, Tampa, FL



9.50 am – 10.40 am



#### **KEYNOTE SPEAKER**

The Magnitude of the Cyber-Security Problem, Trends and Best Practices.

Sri Sridharan, Director, Florida Center for Cybersecurity

10.40am – 10.55 am

Break

#### 10.55 am – 11.15 am

#### Game Theory Methods for Cyber-Security

**Prof. Kandethody Ramachandran**, Department of Mathematics and Statistics, University of South Florida, Tampa, Florida. Director, Interdisciplinary Data Sciences Consortium.

**Abstract:** Internet is an integral way of conducting daily business for government agencies to entertainers. With the increase of mobile users, digital applications and data networks, it is necessary to protect these from attack, damage or unauthorized access. A cyber-security problem can be viewed as a conflict-resolution scenario that typically consists of a security system and at least two decision makers (e.g. attacker and defender) that can each have competing objectives. For instance, the defender may be interested to ensure that the system performs at or above a certain acceptable level, and the attacker's objective may be to disrupt the system and degrade it. Game theory is an appropriate tool that can be used to analyze such problems. In this work, we present how game theory can be used to find strategies for both an attacker and the administrator. We model the interactions between them as a stochastic game. Various formulations of game theory will be presented to deal with different cyber security situations. We will develop mathematical models of security systems to analyze the system's performance and to predict the likely behavior of key decision makers that influence/control the system.

## 11.15 am – 11.35An Introduction and a Comprehensive Analysis on Vulnerability Space, a Probability Space.amDr. Sasith Rajasooriya, Miami University Oxford, Ohio

**Abstract:** Even though there are many important contributions in the modeling of the concept of Vulnerability Life Cycle, those models are not comprehensive enough to explain most of the real world aspects regarding vulnerabilities. Several important states of vulnerabilities are yet to be discussed and included in relevant analyses. Zero Day Vulnerabilities are not very well explained in many models developed even though it is well known that zero-day vulnerabilities represent a major threat to Cybersecurity. Therefore, it is extremely important that we have a proper and comprehensive analytical model for the vulnerability life cycle that would present all the probable states of any vulnerability. However, before developing a comprehensive Vulnerability Life Cycle Model it is mandatory that we understand and list all the possibilities that a vulnerability could face. In other words, we need to identify all possible states of a vulnerability. To achieve this objective, in this study, we introduce the concept of Vulnerability Space, the probability space where all possible incidents that would occur are included.

11.35 am – 11.55 am	Cybersecurity: A Stochastic Predictive Model to Determine Overall Network Security Risk Using Markovian Process
	Nawa Raj Pokhrel, Department of Mathematics and Statistics, University of South Florida, Tampa, FL
	<b>Abstract</b> : There are several security metrics developed to protect the computer networks. In general, common security metrics focus on qualitative and subjective aspects of networks lacking formal statistical models. In the present study, we propose a stochastic model to quantify the risk associated with the overall network using Markovian process in conjunction with Common Vulnerability Scoring System (CVSS) framework. The model we developed uses host access graph to represent the network environment. Utilizing the developed model, one can filter the large amount of information available by making a priority list of vulnerable nodes existing in the network. Once a priority list is prepared, network administrators can make software patch decisions. Gaining in depth understanding of the risk and priority level of each host helps individuals to implement decisions like deployment of security products and to design network topologies.
12.00 pm – 1.00 pm	Lunch Break
1.00 pm – 1.20 pm	<ul> <li>Cybersecurity: A Predictive Analytical Model for Software Vulnerability</li> <li>Dr. Netra Khanal, The University of Tampa, Tampa, FL</li> <li>Abstract: A software vulnerability is defined as a flaw that exists in computer resources or control that can be exploited by one or more threats. Vulnerabilities are discovered throughout the entire life cycle of the software. In this presentation, we examine the existing models on the subject area and propose a new time-based nonlinear differential equation model. Our proposed model is based on the fact that the vulnerability saturation is a local phenomenon, and it possesses an increasing cyclic behavior within the software vulnerability life cycle. The daily vulnerability data is extracted from National Vulnerability Database (NVD) and is designed to obtain cumulative guarterly dataset.</li> </ul>
	We apply the proposed model in cumulative quarterly vulnerability data for three Operating Systems: Mac OS X, Windows 7, and Linux Kernel. Our model performs significantly better when compared with the existing models in terms of fitting and prediction capabilities.

1.20 pm – 1.40 pm

#### Cyber-Security: A Stochastic Modeling Approach for Security Quantification

K. Ruwani M. Fernando, Department of Mathematics and Statistics, University of South Florida, Tampa, FL

**Abstract:** Losses due to Cyber-attacks are immense, varying from loss of money and confidential information to the spoilage of reputation. It is extremely important for corporations to have security metrics in order to mitigate the security threats arising from Cyber-attacks. This paper proposes several measures for network security quantification through application of Stochastic Modeling Techniques. The security quantification measures include: identification of the most critical attack state, the most probable goal among many goal states, the number of steps the attacker is expected to take to compromise the security of the system and the Mean Time to Security Failure (MTTSF). The study is primarily based on software vulnerabilities and we concentrate on attacks launched remotely through internet. We mainly consider four states which are common to any attack scenario: breach/penetrate, strike integrity, strike confidentiality and strike availability. The analytic methods we are using are developed for a generalized framework but can be altered to represent specific network security frameworks that a particular company may require.

#### 1.40 pm – 2.00 pm

#### Nonhomogeneous Risk Rank Analysis Method for Security Network System

Dr. Pubudu Hitigala Kaluarachchilage, Miami University, Oxford, Ohio.

Abstract: A Network system could have numerous vulnerabilities. We understand the process of generating vulnerabilities is highly stochastic and outcomes are hard to predict. Similarly, the behavior of attacks and attackers also have higher level unpredictability. When considering a particular system based on the discovered vulnerabilities the analysis must consider the dynamic nature of the effect of vulnerabilities over time. It is observed that the effect of vulnerabilities vary with the time over their life cycle. Therefore, for a particular system, the most threatening vulnerability at time t1 might not be the same at time t2. Hence, it would be very useful to have analytical models to observe the behavior of the rank of vulnerabilities based on the magnitude of the threat with respect to time for a given network system. Such ranking distribution over time would empower the defenders by giving the priority directions to attend on fixing vulnerabilities. In this study we have attempted to address this need.

2.00 pm – 2.10 pm

## Health Science

#### **Chair: Jordan Creed**

Moffit Cancer Center, Tampa, Florida



#### 2.10 pm – 2.55 pm



#### **KEYNOTE SPEAKER**

Statistical Classification System for "Web of Science" Articles

Dr. Michael J. Schell, Moffit Cancer Center

**Abstract:** Applied statisticians have an ethical responsibility to not be "lax in seeking out the most appropriate statistical tools", as stated in the 1999 ASA Ethical Guidelines. However, it is difficult to

keep up-to-date with the statistical literature, especially if one has a wide-ranging statistical practice. To facilitate these desiderata, a corpus of 3642 articles that have demonstrated very high overall and recent citation rates in the Web of Science, including all 124 journals identified as having a key topic focus of "Statistics Probability" from 1895 to the present have been identified. In order to make this list of articles more easily used as a reference for statisticians, a hierarchical 7-category classification system, paralleling the Linnaean system in biology was developed. It will be described in this presentation.

2.55 pm – 3.05 pm

Break

#### 3.05 pm – 3.25 pm

#### Introduction of FDA Adverse Event Reporting System and data analysis

**Dr. Feng Cheng**, Department of Pharmaceutical Sciences, College of Pharmacy. Department of Biostatistics, College of Public Health, University of South Florida

**Abstract:** FDA Adverse Event Reporting System (FAERS) is the largest post-marketing surveillance database in the world. More than 8 million records from patients have been collected in FAERS. The database contains seven tables including patient demographic information (DEMO), drug information (DRUG), drug adverse reaction (REAC), patient outcome for the event (OUTC), report sources for the event (RPSR), drug therapy start dates and end dates (THER), and the indications for use (INDI). The availability of real-world drug adverse events data from FAERS provides an excellent resource for the discovery of unexpected drug adverse events and possible drug-drug interactions. Several algorithms have been developed for identifying possible associations between adverse reactions and drug (or drug combinations). This talk will review FAERS system and the data analysis.

#### 3.25 pm – 3.45 pm Clustering Analysis of Nonlinear EEG and ECG Signals Using Different Distances and Data Reduction Methods

Mahdi Goudarzi, Department of Mathematics and Statistics, University of South Florida, Tampa, FL Abstract: EEG (Electroencephalography) and ECG (Electrocardiography) are monitoring methods to record the electrical activity of the brain and heart, respectively. It has extensive application to diagnose abnormalities related to the brain and heart such as epilepsy, sleep disorders, depth of anesthesia, coma, encephalopathies, brain death and heart abnormalities. The time of analysis is one of the most important factors. The classification of these types of signals has been investigated from several perspectives. Instance-based and frame-based are two main approaches to detect abnormalities, but the time of analysis is the subject area with less attention than correct classification and abnormality detection. The dimensionality and noisy data are two big obstacles during the analysis with respect to time. Symbolic Aggregate Approximation known as SAX is a methodology for reducing a time series window to a symbol. This method allows a time series of arbitrary length n to be reduced to a string of arbitrary length w (w<<n). In this study, we utilize nonparametric SAX method in conjunction with Mahalanobis distance to decrease the number of signals. Then with respect to classification of signal recurrent boosting technique, Adaboost, is implemented. Since the nature of the signals is time series, recurrent boosting can consider the correlation in time by adding previous labels as a new feature. Our result shows that the time factor of the analysis will decrease drastically when a combination of the aforementioned methods is applied.

3.55 pm – 4.40 pm



#### **KEYNOTE SPEAKER**

#### *Prevalence of Chronic Hepatitis B and C in the United States*

**Dr. Henry Roberts**, Center for Disease Control and Prevention (CDC), Division of Viral Hepatitis.

#### Abstract:

Background: The number of persons with chronic hepatitis B virus

(HBV) infection in the United States is affected by diminishing numbers of young persons who are susceptible because of universal infant vaccination since 1991, offset by numbers of HBV-infected persons migrating to the United States from endemic countries.

Methods: The prevalence of HBV infection was determined by serologic testing and analysis among non-institutionalized persons aged 6 years and older for: antibody to hepatitis B core antigen (anti-HBc), indicative of prior HBV infection; hepatitis B surface antigen (HBsAg), indicative of chronic (current) infection; and antibody to hepatitis B surface antigen(anti-HBs), indicative of immunity from vaccination. These prevalence estimates were analyzed in three periods of the National Health and Nutrition Examination Survey (NHANES): 1988-1994 (21,260 persons); 1999-2008 (29,828); and 2007-2012 (22,358). In 2011-2012, for the first time, non-Hispanic Asians were oversampled sampled in NHANES.

Results: For the most recent period (2007-2012), 3.9% had anti-HBc, indicating about 10.8 (95% CI 9.4-12.2) million non-institutionalized US residents having ever been infected with HBV. The overall prevalence of chronic HBV infection has remained constant since 1999: 0.3% (95% confidence intervals, 0.2% - 0.4%), and since 1999, prevalence of chronic HBV infection among non-Hispanic blacks has been 2-3 fold greater than the general population. An estimated 3.1% (1.8% - 5.2%) of non-Hispanic Asians were chronically infected with HBV during 2011-2012; which reflects a 10-fold greater prevalence than the general population. Adjusted prevalence of vaccine induced immunity increased 16% since 1999, and the number of persons (mainly young) with serologic evidence of vaccine-protection from HBV infection rose from 57.8 (95% CI 55.4-60.1) million to 68.5 (95% CI 65.4-71.2) million.

Conclusion: Despite increasing immune protection in young persons vaccinated in infancy, an analysis of chronic hepatitis B prevalence in racial and ethnic populations indicates that during 2011-2012 there were 847,000 HBV infections (which included ~400,000 non-Hispanic Asians) in the non-institutionalized US population.

#### 4.40 pm – 5.00 pm

#### Association Study of Drugs and Adverse Event

Minh Pham, Department of Mathematics and Statistics, University of South Florida, Tampa, FL

**Abstract:** We consider the problem of using data mining techniques on electronic health records to discover associations between drugs and adverse event. Using data mining algorithms is the most efficient and inexpensive way to inspect post-marketing drug safety, which is one of the objectives of the U.S. Food and Drug Administration to protect the public health. This is, however, a challenging task due to the high dimensionality and sparsity of the FDA's data. In the first half of this presentation, we will give a brief overview of the statistical methods and algorithms proposed in the literature, which could be divided into frequentist approaches, Bayesian approaches, and multivariate approaches. In the second half, we applied these methods to the FDA's Adverse Event Reporting System and compare their performances on a known set of drugs and adverse events provided by the Observational Medical Outcomes Partnership. In addition, several methods that had been successfully applied in the Genome-Wide Association Studies were adapted to our problem and included in the comparison study.



#### 10.20 am – 10.40 am Artificial Intelligence Systems and Investor Suitability for Hedge Funds

Michael Kotarinos, Department of Mathematics and Statistics, University of South Florida, Tampa, FL

**Abstract:** Artificial intelligence algorithms have generated significant interest by fund managers seeking to generate favorable performances relative to benchmarks. While these algorithms have found success in other areas, AI returns have lagged behind fundamentals and benchmark algorithms. This presentation first discusses both computational and legal issues with traditional machine learning algorithms being used by hedge funds and then presents an alternative method using a combination of MLTS clustering with a set of preferences over risk and asset structure. By combining this clustering algorithm with a set of preferences, the algorithm overcomes some of the computational issues plaguing traditional algorithms in this space while addressing some of the legal issues that have been raised by the SEC. The presentation concludes with some closing remarks on the challenges of automated trading in the future and possible regulatory issues that may arise.

#### 10.40 am – 11.00 am A Statistical Model that Identifies the Democracy Ranking for Countries in the World

**AKMR Bashar,** Department of Mathematics and Statistics, University of South Florida, Tampa, FL

**Abstract:** The Economist Intelligence Unit (EIU) annually collect information on the contributable Variables in 167 countries that they believe drives DEMOCRACY. The survey consists of a questionnaire in five different attributes: Electoral Process and Pluralism (EPP), functioning of a government (FG), Political Participation (PP), Political Culture (PC) and Civil Liberties (CL). Using this data EIU identifies a score for each of the country and classifies the country as Full Democracy, Hybrid Regime, Flawed Democracy and Authoritarian Regime. In the present study, we have utilized the database of EIU and developed a statistical model that identifies significant variables and interactions to predict the democracy scores of the countries. In addition, we rank the attributable variables with respect to their percent contribution to the democracy score. Our results are compared to the descriptive findings of EIU.

*11.00 am – 11.15 am* Br

### **Environmental Sciences**

**Chair: Mohamed Abusheha** Department of Mathematics & Statistics, University of South Florida, Tampa, FL



#### 11.15 am – 11.35 am

Statistical Analysis and Forecasting of The Atmospheric Carbon Dioxide in The Middle East as The Main Factor of The Global Warming

Maryam Habadi, Department of Mathematics and Statistics, University of South Florida, Tampa, FL

Abstract: Saudi Arabia has been ranked as the 8<sup>th</sup> largest carbon dioxide emitter among all the countries in the worlds and some of the Middle Eastern countries at the top 50 based on the record of total  $CO_2$  emissions in from fossil fuel burning and cement production in 2014. As it is known that the excessive rise of carbon dioxide from the normal level as the side effect of the industrial revolution has the significant effect in blocking the heat and increase the temperature within the Earth's atmosphere. First in this study, we have developed a data-driven nonlinear statistical model to identify and rank the significant types of fossil fuel burning (gas fuel, liquid fuel, and solid fuel), cement manufacture, and gas flaring and their interactions based on their percentage of contribution to the atmospheric  $CO_2$  concentrations in the Middle East. Then, the results of the finding are compared to the findings of the United States, European Union and South Korea. Second, the multiplicative seasonal autoregressive integrated moving average (seasonal ARIMA) model is used to develop statistical time series forecasting models to predict carbon dioxide in the atmosphere in the Middle East and atmospheric temperature in Saudi Arabia. Thus, the resulting statistical predictive model is useful in forecasting and monitoring the future level of carbon dioxide emission and extract meaningful statistics and characteristics about the emission of carbon dioxide in the Middle East.

## 11.35 am – 11.55 amSpatiotemporal trends in daily precipitation extremes and their connection with NorthAtlantic tropical cyclones for the Southeastern United States

Dr. Bhikhari Tharu, Mathematics Department, Spelman College, Atlanta, GA

**Abstract:** Changes in extreme precipitation are associated with changes in their probability distributions and the characteristics of quantiles derived from fitted distributions. In this study, the linear quantile regression method is employed to analyze spatio-temporal trends of extreme precipitation and to study the impact of North Atlantic Tropical Cyclones in the distribution of extreme precipitation for the Southeastern United States. Daily annual maximum precipitation over the period of 64 years (1950 -2013) for 107 sites was used for the analysis. Our results show that changes in upper quantiles of the distributions of the extreme precipitation have occurred in the Southeastern United States. Analysis of the potential changes in the distribution of the extreme precipitation by separating the historical record into two periods, i.e., before and after 1981, reveals that upper-quantile trends have increasing magnitude in most of the sites for the latest time period. Analysis of the impact of tropical cyclones in the extreme precipitation distribution shows that overall, the heavy rainfall events in the recent decades may have been caused by tropical cyclones. Such results are particularly useful for water managers who are more concerned with extreme values rather than the averaged one. Hence, our study has significant implication in environmental and infrastructural assessment as well as disaster risk management.

#### 11.55 am –12.15 pm A Statistical Approach to Underground Object Detection

Sajad S. Jazayeri, School of Geosciences, University of South Florida

**Abstract:** A novel statistical monitoring scheme is introduced that measures dissimilarity of Ground Penetrating Radar signals to signals from target-free locations using Dynamic Time Warping to detect multiple hidden buried object prior to reaching them. Further, investigating the potential burial site lead to estimating their locations as well as depth with high accuracy. The analytics run fast and could be performed in real-time without user's interference.

12.15 pm – 1.15 pm

Lunch Break

1.15 pm – 1.35 pm	Real-time and Automatic Underground Object Detection by using Ground Penetrating Radar SignalsDr. Abolfazi Saghafi, Department of Mathematics, Physics, and Statistics, University of the SciencesAbstract: Maximum energy of Ground Penetrating Radar signals is estimated via Power Spectral Density method and tightly monitored using state-of- the-art Sequential Confidence Intervals to detect potential burial sites. Upon detecting a highly probable burial site, automatic analytics estimate the location and depth of multiple buried objects with high accuracy in real-time. Testing the performance of the proposed analytics in numerous synthetic and real scenarios even in noisy media proved its efficiency for real-life exploration.
1.35 pm – 1.55 pm	Differential Equation Model of Carbon Dioxide Emission Using Functional Linear Regression Dr. Ram C. Kaffle, Sam Houston State University, Texas
	Abstract: Carbon dioxide is one of the major contributors to Global Warming. In the present study, we develop a differential equation to model the carbon dioxide emission data in the atmosphere using functional linear regression approach. In the proposed method, a differential operator is defined as data smoother and used the penalized least square fitting criteria to smooth the data. The profile error sum of squares is optimized to estimate the differential operator using functional regression. The solution of the developed differential equation estimates and predicts the rate of change of carbon dioxide in the atmosphere at a particular time. We apply the proposed model to fit the emission of carbon dioxide data in the continental United States.
1.55 pm – 2.10 pm	Break

## Advances in Statistics

**Chair: Michael Kotarinos** Department of Mathematics & Statistics, University of South Florida, Tampa, FL



#### 2.10 pm – 2.30 pm Optimal Reliability Demonstration Test Plan for Multiple Objectives

Dr. Lu Lu, Department of Mathematics and Statistics, University of South Florida, Tampa, FL

**Abstract:** Reliability demonstration tests are commonly performed in product development or a validation process to demonstrate whether a product meets specified requirements on reliability. Among binomial demonstration tests, zero-failure tests have been most commonly used for their simplicity and the use of minimum sample size to achieve an acceptable consumer's risk level. However, these often result in unacceptably high risks for producers as well as low probabilities of passing the tests even when products have adequate reliability. In this talk, we explicitly explore the interrelationship between multiple objectives that are commonly of interest when planning a demonstration test, and propose structured decision-making strategies using a Pareto front based approach for selecting optimal test plans that simultaneously balancing multiple criteria. A variety of strategies are suggested for scenarios with different user priorities, and graphical tools are developed to effectively quantify the trade-offs between choices and to facilitate informed decision-making. Potential impacts on the final decision of some subjective user inputs are studied to offer insights and useful guidance for general applications.

#### 2.30 pm – 2.50 pm

#### Modern applications of Regression

Dr. Rebecca D Wooten, Florida Southern College

**Abstract:** This presentation introduces modern techniques of regression used to fit co-dependent measures: conic sections, indirect relationships and implicit equations; and how to fit bivariate probability distributions to co-dependent variables using these methods.

#### 2.50 pm – 3.10 pm

#### Generalized Family of Distributions with Applications

Dr. Gokarna Aryal, Department of Mathematics, Statistics and CS, Purdue University Northwest

**Abstract:** Choice of a probability distribution which should be adopted to model the data under study has an undeniable impact on the quality of the results. Therefore, there has been a growing interest in the construction of parametric classes of generalized probability distributions to model data under study. In this talk we will discuss some generalizations of Pareto distribution. In particular, the construction of so-called beta exponential Pareto (BEP) distribution will be discussed. Several lifetime distributions including the beta-Weibull, beta-exponential, beta-Rayleigh, generalized Weibull, Weibull among others are embedded in the proposed distribution. Various mathematical properties along with parameter estimation and simulation issues will be discussed. The importance and flexibility of the proposed distribution will be illustrated by means of real data analysis.

3.10 pm – 3.30 pm

Closing Remarks by Professor Chris P. Tsokos, Distinguished Professor, USF, Tampa. Fl

#### **Organizing Committee**

Michael Kotarinos Ruwani Fernando Freeh Alenezi Nawa Raj Pokhrel Tianbo Zhang Shao Huang

# 2ND ANNUAL CONFERENCE ON THE **FRONTIERS OF STATISTICS**

