INTERNATIONAL FEDERATION OF NONLINEAR ANALYSTS (IFNA)

ASA USF STUDENT CHAPTER

Analytics and Big Data across the Disciplines
Frontiers of Statistics

University of South Florida, 2016

CWY 108, April 1st and 2nd

International Federation
IFNA
DIVERSITY • COOPERATION • COLLABORATION
of Nonlinear Analysts
American Statistical Association
University of South Florida Student Chapter
IDSC
BIG DATA
UNIVERSITY of SOUTH FLORIDA
TRUTH
WISDOM
1956

Overview Agenda

Friday, April 01, 2016

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

Friday, April 01, 2016

9:30 am-10:00 am	Welcome Speech and Opening Remarks
10:00 am-10:50 am	Data Science across Disciplines, Invited Speaker
10:50 am-11:00 am	Small Break
11:00 am-12:00 pm	Data Science across Disciplines, Student Presentations
12:00 pm-1:00 pm	Lunch Break
1:00 pm-1:50 pm	Data Science Across Disciplines Closing remarks, Invited Speaker
1:50 pm-2:00 pm	Small Break
2:00 pm-2:40pm	Biostatistics Session, Student Presentations
2:40 pm-2:50pm	American Statistical Association at USF Student Chapter
2:50pm-3:00pm	Small Break
3:00pm-4:00pm	Biostatistics Session, Invited Speaker
4:00pm-5:00pm	Panel Discussion with the Invited Speakers

Saturday, April 02, 2016

8:40am- 9:00am	Registration
9:00am-10:50am	Environment Session, Invited Speakers
10:50am-11:00am	Small Break
11:00 am-11:45 am	Cybersecurity
11:45 am-12:00 pm	Closing Remarks (Tsokos)

Friday, April 1- CWY 108

9:00am – 9:30am Registration

9:30am - 10:00am Welcome Speech: Chris Tsokos

Distinguished Professor at USF Statistics and Mathematics Department

IFNA President

==============Data Science across Disciplines=================

10:00am - 10:50am **Statistical Analysis for Device Usage in Electronic Classrooms using SAS Macro.**

Invited Speaker: Young Xu, PhD, Department of Mathematics and Statistics, Radford University.

With modern industrialization and high technology, people tend to consume more and more energy. With world population increase very fast and living standard increasing rapidly, the challenge for energy saving becomes more and more obvious. We plan to start with a statistical analysis to one campus electronic devices usage to understand the pattern of the energy cost in e-classroom. We developed a SAS macro for people to use for similar research purpose. We wish with better understand of how people consume energy to eventually save significant amount of energy and have a better future. The technology and multimedia devices becoming ever more present in classroom environments from Elementary Schools to Universities, understanding which devices and how often they are used is essential for this study. We are using data collected from sixty eight multimedia classrooms at Radford University from Crestron Room View to develop time series models. Having the statistical models of classroom technology/multimedia will aid IT Professionals, Administrators, and Teachers by allowing for more efficient classroom design and device selection. It provides one direction to help people to conduct big data analysis for energy saving purpose.

10:50am - 11:00am Small Break

11:00am – 11:20am **Support, Resistance, and Applications of Technical Analysis to Time Series Modeling of Financial Assets**

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

Support and Resistance levels are commonly used by financial technicians to model upward and downward pressures in financial markets across asset classes. These levels can be used to incorporate some aspects of market fundamentals into the analysis of a stock's momentum, thus incorporating important market characteristics and features to improve decision making. In this paper the authors began with a review of support and resistance levels and basic concepts from technical analysis. Using stochastic calculus allowed for the demonstration of the theoretical applications of these levels and how they can naturally occur across markets. The authors then proposed a new time series model for financial data that extends basic ARIMA models to incorporate support and resistance levels. The report continues with an example using the S&P 500 showing how incorporating support and resistance leads to improved decision making and more control over the modeling process. The authors concluded with some closing remarks regarding the relationship between technical analysis and statistics and implications for future research.

11:20am - 11:40am

Active and Dynamic Approaches for Clustering Time Dependent Information: Lag Target Time Series Clustering and Multi-Factor Time Series Clustering.

Doo Young Kim, Department of Mathematics and Statistics, University of South Florida, Tampa, FL.

One of data mining schemes in statistics is clustering panel data such as longitudinal data and time series data. Classical approaches to cluster such time dependent information do not properly count time dependencies among objects we are interested to analyze. In the present study, we propose an approach which takes time dependencies into our consideration by introducing appropriate weight factors with an add-on approach which allows us to measure pairwise distances in multi-dimensional space not just in two dimension. We refer to these approaches LTTC (Lag Target Time Series Clustering) and MFTC (Multi-Factor Time Series Clustering), respectively. These proposed methods in the present study are applicable to any time dependent information from various research areas, and we have applied these methods to state level brain cancer mortality rates in the United States that illustrates the importance of subject methods.

11:40am - 12:00pm

Time Dependent Kernel Density Estimation-based Classification of Time Series Data

Xing Wang, Department of Mathematics and Statistics, University of South Florida, Tampa, FL.

To improve the performance of the existing statistical feature-based time series classification algorithm, the Time Dependent Kernel Density Estimation (TDKDE) based Random Forest classification algorithm is developed in this study. The performance is examined using twenty datasets from the UCR Time Series Classification Archive, which has been widely used as a benchmark for evaluating the performance of time series classification algorithms. The experimental results verify that, TDKDE-based Random Forest approach is significantly superior over the statistical feature-based Random Forest approach in terms of the out-of-bag (OOB) errors.

12:00pm - 1:00pm Lunch break

1:00pm - 1:50pm **An Overview of Functional Data Analysis**

Invited Speaker: Keshav P. Pokhrel, Department of Mathematics and Statistics, University of Michigan-Dearborn Dearborn, Michigan

Functional data analysis (FDA) is a reasonably new development in the literature of Statistics. I will discuss briefly about current state of developments in FDA. As most of the existing inferential techniques are inadequate for functional data, there is a strong need to develop robust inferential framework for functional data analysis techniques. In particular, we will discuss about some inferential techniques on functional time series and layout possible areas to explore in functional autocorrelation. Major dataset in the discussion is extracted from National Cancer Institutes population based database: Surveillance, Epidemiology and End Results (SEER).

1:50pm - 2:00pm Small Break

=================Biostatistics========================

2:00pm - 2:20pm **Bayesian Age-Period-Cohort Model of Lung Cancer Mortality**

Bhikhari P. Tharu, Chris P. Tsokos, Department of Mathematics and Statistics, University of South Florida, Tampa, FL.

Ram Kafle, Sam Houston State University- Huntsville, TX.

The objective of this study was to analyze the time trend for lung cancer mortality in the population of the USA by 5 years based on most recent available data namely to 2010. The knowledge of the mortality rates in the temporal trends is necessary to understand cancer burden. Bayesian Age-Period-Cohort model was fitted using Poisson regression with histogram smoothing prior to decompose mortality rates based on age at death, period at death, and birth-cohort. Mortality rates from lung cancer increased more rapidly from age 52 years. It ended up to 325 deaths annually for 82 years on average. The mortality of younger cohorts was lower than older cohorts. The risk of lung cancer was lowered from period 1993 to recent periods. The reduction in carcinogens in cigarettes and increase in smoking cessation from around 1960 might led to decreasing trend of lung cancer mortality after calendar period 1993.

2:20pm - 2:40pm

On Heredity Factors of Parkinson's disease: A Parametric and Bayesian Analysis

Abolfazl Saghafi, Abolfazl Saghafi, Chris P. Tsokos, Rebecca Wooten, Department of Mathematics and Statistics, University of South Florida, Tampa, FL.

The prevalent chance of having Parkinson's disease (PD) in families whose none of the parents, one of the parents, and both of the parents carry the disease is estimated and compared in present study. Maximum likelihood and Bayesian approach have been used. Historical data on the number of people in grandparents' family who had the PD has been incorporated with the data in the form of prior information to draw Bayesian estimations. For the families with negative history of PD the prevalent chance is estimated to be 20% meaning that a child in this family has 20% chance of developing Parkinson. If there is positive history of PD in the family the prevalent chance increases to 33% when none of the parents had PD and to 44% when both of the parents had the disease. The chance of developing PD in a family whose solely mother is diagnosed with the disease is estimated to be 26% in comparison to 31% when only father is diagnosed with Parkinson's.

2:40pm – 2:50pm

American Statistical Association USF Chapter - Presentation

Zheni Stefanova, President, Department Mathematics and Statistics, USF

2:50pm – 3:00pm

Small Break

3:00pm – 3:30pm

Medicaid eligible but not enrolled, who are they? A case study of the cohort of breast cancer patients in North Carolina.

Ke Meng, PhD and Bong- Jin Choi, PhD, University of North Carolina at Capel Hill.

Presented by Invited Speaker Bong –Jin Choi

There are approximately 9.2 million people in the United States who are dually eligible beneficiaries, which refer to the population who are qualified for both Medicare and Medicaid benefits. Medicare benefits are easier to access and everyone who is 65 or plus is automatically enrolled. However, not all the dual eligible population take up their Medicaid benefits. Using a unique linked data resource created by the University of North Carolina Lineberger Comprehensive Cancer Center, this study demonstrates the characteristics of Medicaid non-take up population and explores the covariates that may affect Medicaid take up. 37,057 Medicare enrollees are linked with North Carolina Cancer Registry, and 7,653 (21%) are eligible for Medicaid coverage for at least one month during the year 2006-2011. 2,103 Medicare enrollees are eligible for Medicaid benefits for the entire study window of 2003 to 2006 and 1,596 ages 65+. These 1,596 became the final analytic cohort for this study. 88% of Medicare enrollees 65-75 take up the Medicaid benefits, compared to 92% of 76+ ($p<0.05$), and 89% of the White take up the benefits, compared to 93% of non-white populations ($p<0.05$).We also estimated a logistic regression using Medicaid non-take up as an outcome and age, race and census tract poverty level as covariates and no significant results were found.

3:30pm – 4:00pm

Probabilistic Big Data Linkage Methodology for Record Linkage Using Cancer Registry Data and Insurance Claim Data

Bong-Jin Choi PhD, University of North Carolina at Capel Hill

There are two main types of linkage algorithms (deterministic and probabilistic.) Both have been successfully implemented in previous research studies. Choosing the best algorithm to use in a given situation depends on many factors including time, resources, research question, and the quality of available identifiers. To link two datasets, we need to clean and standardize identifiers in both datasets. We then concatenate all personal identification variables such as SSN, Names, address, etc. and hash them using MD5. The MD5 can then split into multiple unsigned integers. This significantly reduces computing time and enables us to assign a Unique Personal Identification Number (UPID) which enables matching across data sets. Using machine learning, Bayesian, and optimizing sensitivity analysis, we linked the multiple big data sets using probabilistic linkage based on UPID. We compared this approach against the SEER-MDCR algorithm as the gold standard.

4:00pm – 5:00pm

Discussion Panel with the Invited Speakers

Saturday, April 2- CWY 108

8:40am – 9:00am

Registration

================Environment Session ==================

9:00am - 9:45am

Bayesian Quantitative Microbial Risk Assessment Model for Assessing Wastewater Microbial Risks

Invited Speaker: Ram C. Kafle, Assistant Professor of Statistics, Department of Mathematics and Statistics, Sam Houston State University, TX, USA.

Direct and indirect reuse of wastewater is increasing recently because of limited clean water supply. The reuse of such untreated or minimally treated wastewater present a serious public health problems. The current wastewater reuse guidelines to minimize the risk of population illness are based on Quantitative Microbial Risk Assessment (QMRA) . Model. In this study, we develop the Bayesian approach to QMRA model incorporating ethnographic data and pathogen measurement to assess the impact of riverbank filtration system on consumer health burdens for different pathogens infections resulting from indirect wastewater reuse with lettuce irrigation. The data for this study were collected from Bolivia. In this presentation, we also talk about the implementation of Bayesian Network to QMRA to determine the interventions that can minimize the disease burden associated with the reuse of wastewater.

9:45am - 10:00am

Assessing and Adjusting Non-proportional Hazards in Cox Proportional Hazard Model for Survival Data

Muditha Perera and Chris Tsokos, Department of Mathematics and Statistics, University of South Florida, Tampa, FL.

Assessing and Adjusting Non-proportional Hazards in Cox Proportional Hazard Model for Survival Data Cox proportional hazards model is one of the most common method used to analyze survival data to estimate the effects of prognostic factors on time to an event. It assumes that the covariate effects are constant over time. In other words, hazard ratios do not change with time. But, in practical applications of the Cox model there may be covariates that violate the proportional hazards assumption. Underlying assumptions need to be checked before making any interpretations by the Cox proportional hazards model otherwise it would result in misleading conclusions. In recent years, methods have been developed to check the validity of the proportional hazards assumption and to adjust the Cox model if there are any non-constant hazard ratios. In the current study, we compare few of those methods along with applications to real life survival time data.

10:00am - 10:15am **Exploratory Factor Analysis and Modeling of Hurricanes in the Atlantic Basin.**

Joy D'Andrea , University of South Florida Sarasota – Manatee College of Arts and Sciences, Department of Mathematics & Statistics

Exploratory factor analysis (EFA) is used to determine the number of latent variables that are needed to explain the correlations among a set of observed variables. In this study, the latent variables are the meteorological measures such as the location of a storm, wind speed and pressure. Further analysis investigates the probability of storms being present in the Atlantic Basin using logistic regression and atmospheric conditions recorded at a fixed location. Using a time delay, we can address the probability of a storm formation over a time differential.

10:15am - 10:30am **Statistical Analysis And Modeling Of The Atmospheric Carbon Dioxide In The Middle East And Comparisons With USA, EU And South Korea.**

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

Global warming is considered as one of the important issues facing our planet. It refers to an increase in average global temperatures that caused mostly by increases in Carbon Dioxide . During the last two decades, the emission in the middle eastern countries increased by over 200% based on The Energy Information Administration (EIA). Thus, the aim of this study is to structure a good statistical model for atmospheric in the Middle East to identify significant attributable variables that produce the emissions. Fossil fuel burning (gas fuel, liquid fuel, and solid fuel) cement manufacture, and gas flaring and their interactions have been identified and ranked based on their percentage of contribution to CO2 in the atmosphere. Finally, the results of this modeling are compared to the findings of the United States, European Union and South Korea.

10:30am - 10:50am **Modeling Carbon Dioxide Emission Data using Differential Equation**

Invited Speaker : Netra Khanal,PhD, Department of Statistics, University of Tampa

Carbon dioxide (CO2) is one of the major contributors in Global Warming. This study focuses on developing a system of differential equations using time series data of significant contributable variables of carbon dioxide in the atmosphere in the continental United States. We define the differential operator as data smoother and use the penalized least square fitting criteria to smooth the data. The proposed model gives an estimate of the rate of change of carbon dioxide in the atmosphere. The data set is obtained from the Carbon Dioxide Information Analysis Center (CDIAC), the primary climate-change data and information analysis center of the United States Department of Energy.

10:50am - 11:00am Small Break

==============Cyber Security Session=====================

11:00am - 11:15am **Cybersecurity: A Statistical Predictive Model for the Expected Path Length**

Pubudu Kalpani Kaluarachchi, Department of Mathematics and Statistics, University of South Florida, Tampa, FL.

The object of this study is to propose a statistical model for predicting the Expected Path Length (expected number of steps the attacker will take, starting from the initial state to compromise the security goal-EPL) in a cyber-attack. The model we developed is based on utilizing vulnerability information along with having host centric attack graph. Utilizing the developed model, one can identify the interaction among the vulnerabilities and individual variables (risk factors) that drives the Expected Path Length. Gaining a better understanding of the relationship between vulnerabilities and their interactions can provide security administrators a better view and an understanding of their security status. In addition we have also ranked the attributable variables and there contribution in estimating the subject length. Thus, one can utilize the ranking process to take precautions and actions to minimize Expected Path Length.

11:15am - 11:30am **Markov Model Approach to Vulnerability Life Cycle and Security Risk Evaluation**.

Sasith Rajasooriya, Department of Mathematics and Statistics, University of South Florida, Tampa, FL.

The objective of this study is to proposing of a risk evaluation model for a vulnerability by examine the vulnerability Life cycle and the CVSS score. Having a better understanding of the behavior of a vulnerability with respect to time will give great advantage to avoid exploitations and introduce patches for the vulnerability before attacker take the advantage of that particular vulnerability. Utilizing the proposed model one can identify the risk factor of vulnerability being exploited with time. Measuring of the risk factor of a vulnerability will help to improve the security level of software and take appropriate decision to patch the vulnerability before exploiting.

11:30am - 11:45am **Non Parametric Analysis on Software Vulnerabilities**

Nawa Raj Pokhrel, Department of Mathematics and Statistics, University of South Florida, Tampa, FL.

Cyberattack is the direct exploitation of the computer resources such as hardware, software, operating system and network. Cyberattacks normally uses the malicious code to change the computer code, logic of the data, as a result compromise the resources lead to the cybercrime, such as information and identity theft. All this happens when weakness exist on computer resources on any form, technically flaw or weakness on the resources is called vulnerability. Vulnerability evaluation plays the measure role for the security position and risk management. However, common security metrics are often qualitative, subjective in fact lacking formal statistical model. This papers mainly discussed the non-parametric analysis of the base score. It is divided into 3 categories as low medium and high ranges from (0-3), (4, 7) and (7-10) respectively. Kruskal Wallis tests is applied on three categories to check whether the median values of those categories are equal or not. This result provides insight of the data and helps us to develop the proper software vulnerability model in future.

11:45am - 12:00am **Closing Remarks – Chris Tsokos**

==