Seventh Annual Red River Valley Statistical Conference

North Dakota State University
Department of Statistics

Friday, April 7, 2017

Seventh Annual Red River Valley Statistical Conference

Session 1: Chair: Gang Shen Location: Prairie Rose Room

10:00 am *Analysis and Prediction of Beijing's Air Quality,* Di Gao
10:15 am *Modeling the Number of Reported Motor Vehicle Accidents in the Red River Valley Area,* Brian Mullen
10:30 am *Time Series Analysis of Fatal Crash in Michigan 1994-2015,* Liming Xie
10:45 am *Predicting PM2.5 in Shanghai Using ARMA Model,* Xiaoyi Zhou

Session 2: Chair: Seung Won Hyun Location: Prairie Rose Room

11:15 am *Forecasting Crude Oil Prices—Time Series Approach,* Jingjun Zhao
11:30 am *Adaptive Bayesian C-optimal Design of Estimating Multiple ED_ps Toxicological Studies,* Anqing Zhang
11:45 am *Particle Swarm Optimization Algorithm for Searching the D-optimal Design for the 5-Parameter Logistic Model,* Qiang Li
12:00 pm *Modifying SAMseq to Account for Asymmetry in the Distribution of Effect Sizes when Identifying Differentially Expressed Genes,* Ekua Kotoka

REFRESHMENTS—12:15-12:55 p.m. Hidatsa Room

Session 3: Chair: Ron Degges Location: Prairie Rose Room

1:00 pm *Bracketing 2016-2017 Season NCAA March Madness,* Di Gao
1:15 pm *Detecting the Relationship between Satisfaction of Students and Gender Instructors at NDSU,* Yue Zhou
1:30 pm *A Geostatistical Analysis of Housing Prices in Fargo,* Tika Lamitare
1:45 pm *Introduction to Discriminative Probabilistic Models of Relational Data,* Yue Ming & Xiyuan Liu

ABSTRACTS FOR SPEAKERS

(in alphabetical order by first author's last name)

Author: Di Gao

Title: Analysis and Prediction of Beijing's Air Quality

Beijing is the world's third most populous city proper. Other than its population, "smog" is now another symbol for Beijing. The hazardous smog in Beijing is a very serious environmental problem. Understanding the pattern or causal factors of smog weather is now becoming very important. In this talk, I will introduce a times series model to locate the key factors that may highly relate to air quality. Also, the model will provide a forecast of the air quality in the future.

Author: Di Gao

Title: Bracketing 2016-2017 Season NCAA March Madness

NCAA's March Madness is becoming more and more popular. Based on AGA's (American Gaming Association) report (2015), about 40 million people filled out 70 million March Madness brackets (Moyer, 2015). This talk will cover several statistical and machine learning approaches for bracketing, including Bayesian LASSO with logit link which possess the desirable win-lose probability self-consistency. The new LASSO Bayesian procedure allows us to pick a few significant explanatory variables from a large pool of candidate team statistics and then borrow the information from the recent winning rate among different seeds. The brackets of NCAA for the current season will be provided at the end of this talk, and the accuracy of bracketing will also be discussed.

Author: Ekua Kotoka

Title: Modifying SAMseq to Account for Asymmetry in the Distribution of Effect Sizes when Identifying Differentially Expressed Genes

RNA-Seq is a developing technology for generating gene expression data by directly sequencing mRNA molecules in a sample. RNA-Seq data consist of counts of reads recorded to a particular gene that are often used to identify differentially expressed (DE) genes. A common statistical method used to analyze RNA-Seq data is Significance Analysis of Microarray with emphasis on RNA-Seq data (SAMseq). SAMseq is a nonparametric method that uses a resampling technique to account for differences in sequencing depths when identifying DE genes. We propose modifications of this method that take into account asymmetry in the distribution of the effect sizes by taking into account the sign

of the test statistics. Through simulation studies, we show that the proposed method, compared with the traditional SAMseq method and other existing methods provide better power for identifying truly DE genes while sufficiently controlling FDR in most settings. We illustrate the use of the proposed method by analyzing a C57BL/6J (B6) and DBA/2J (D2) mouse strains samples.

Author: Tika Lamitare

Title: A Geostatistical Analysis of Housing Prices in Fargo

As the City of Fargo grows in population and its size is expanding, the demand for residential housing market is increasing. Econometricians have usually considered age, size of the house, number of bedrooms, number of bathrooms, and proximity to shopping malls as the core factors that affect prices of residential houses. There are two primary regression techniques to model housing prices: hedonic regression modeling and geostatistical modeling. Hedonic regression models place an equal importance on the core structural and neighborhood's characteristics of a property. Geostatistical models also emphasize those characteristics. The major difference between these two techniques lies in the fact that geostatistical models effectively incorporate the underlying correlation structure between properties based on their separation distance. The first step in this study will be to identify the correlation structure by plotting a variogram of the residuals. The initial estimates from the plotted variogram will be used to generate new covariance parameter estimates. Furthermore, new estimates will be used to create two different regression models: geostatistical model with exponential covariance function and geostatistical model with spherical covariance function. After model selection steps, interpolation will be performed based on a technique known as regression kriging.

Author: Qiang Li

Title: Particle Swarm Optimization Algorithm for Searching the D-optimal Design for the 5-Parameter Logistic Model

Particle Swarm Optimization(PSO) is a population based stochastic optimization method inspired by social behavior of bird flocking or fish schooling (Eberhart and Kennedy, 1995). This technique is widely used in engineering and computer science for solving very high dimensional optimization problem. Recently, PSO was applied to complex optimal design problems due to its simplicity, efficiency, and effectiveness. In this project, we apply PSO algorithm to search the D-optimal design for the 5-parameter logistic(5PL) model. The 5PL model dramatically improves the accuracy of assays when the dose-response has an asymmetrical sigmoidal curve and the D-optimal provides the best

experimental design for estimating the model parameter accurately and efficiently. In the future, we will apply PSO algorithm to search optimal designs for the 5PL model for studying more complex real application problems such as multi-objective problems and robustness properties to misspecified model forms and model parameter values.

Eberhart, R. C., and Kennedy, J.. (1995). *A new optimizer using particle swarm theory.* Proceedings of the Sixth International Symposium on Micro Machine and Human Science, Nagoya, Japan, 39-43. Piscataway, NJ: IEEE Service Center.

Authors: Yue Ming & Xiyuan Liu

Title: Introduction to Discriminative Probabilistic Models for Relational Data

Most statistical classification methods have focused on independent and identically distributed (iid) data. However, such case is very rare in reality. For example, hyperlinked webpages, social networks and cross-citation in patents and scientific papers. Hence the paper: Discriminative Probabilistic Models for Relational Data(Taskar 2002) introduced a framework of relational Markov network (RMN) that can significantly improve the accuracy of classification by modeling the relational dependencies.

Reference:

Taskar, B., Abbeel, P., Koller, D.: Discriminative Probabilistic Models for Relational Classification. Proc. of Uncertainty on Artificial Intelligence, Edmonton, Canada (2002) 485–492

Author: Brian Mullen

Title: Modeling the Number of Reported Motor Vehicle Accidents in the Red River Valley Area

Motor vehicle accidents are an unfortunate reality of everyday life. They cost society time and tax dollars, the involved individuals substantial amounts money, and sometimes even more than just money. Therefore, there is interest in studying the frequency of the accidents. Data of all reported motor vehicle accidents was extracted from the Red River Regional Dispatch Center from their publicly posted dispatch logs from March 6th 2011 until February 25th 2017. Daily weather observations from the Hector International Airport for the same dates was also collected. Then a seasonal ARIMA model (SARIMA) was fitted to the motor vehicle accident time series as well as exogeneous variables; accumulated amounts of snowfall, and precipitation in freezing temperatures (assumed to be freezing rain) to create a SARIMAX model. Given daily estimated amounts of accumulated snowfall and precipitation, this model will forecast the number of reported motor vehicle accidents.

Author: Liming Xie

Title: Time Series Analysis on Fatal Crash in Michigan 1994-2015

The traffic crash deaths caused by many kinds of vehicles have been serious in the world. The traffic crash deaths from Michigan state is better than other states in the United States. The statistical data in Michigan showed that the rate of death (Death Rate = Persons killed per 100 million MVMT) decreased from 1.6 in 1994 to 1.3 in 2015. The accident fatal death reasons would be various, but these are mainly by driver's drinking, health issues, drug use, and weather condition. Majority of fatal crash chosen for this project are listed as the main resources of Michigan road crash facts. This project is to apply statistical theories to analyze the traffic crash reasons so that we could find some effective ways to avoid or reduce traffic crash deaths for traffic safety office to schedule their police from relative industries such as insurance will also find it helpful in predictive analysis.

Author: Anqing Zhang

Title: Adaptive Bayesian C-optimal Design for Estimating Multiple ED_ps in Toxicological studies

In the previous research, maximum likelihood (ML) estimation approach was adopted when calculated locally c-optimal design for estimating multiple ED_ps. The problem of the ML approach is that by using only a point estimate, the parameter uncertainty is ignored and it may cause misleading plausible parameter values. If the estimated values of parameters are not close to true values, locally optimal designs are far from optimums. To address this parameter uncertainty problem, two-stage c-optimal design for estimating multiple ED_ps is studied. Another approach to reduce the model parameter dependency is to adopt Bayesian optimal design. Bayesian optimal design use the informative prior distribution of the unknown parameters to derive a better design. One challenge is that it needs heavy computation in the numerical calculation procedure when search Bayesian optimal design. To overcome this problem, clustering methods can be applied. We choose three common clustering methods: k-means, kernel k-means, and fuzzy c-means to compare their performances. Finally, compare the two-stage c-optimal design and the adaptive Bayesian c-optimal design with some other traditional designs like Uniform design to see how well they perform in estimating multiple ED_ps.

Author: Jingjun Zhao

Title: Forecasting Crude Oil Prices—Time Series Approach

Forecasting crude oil prices has been a hot research topic in both academic and industrial areas. Accurate predictions of future trends of crude oil prices are of particular importance to energy analysts. Many methods and approaches have been developed for predicting oil prices. However, due to the high volatility of oil prices, it remains one of the most challenging forecasting problems. Among these methods and approaches for predicting oil prices, time series analysis is relatively easier to create and implement a statistical model and easier to identify changes in trends. An autoregressive integrate moving average (ARIMA) based time series model for crude oil prices prediction is proposed in this project, and the experiment results shows that the proposed model is capable of effectively forecasting the future trends of crude oil prices from selected real historical data. No single type of forecasting model is sufficient to accurately predict the trends of crude oil prices. Combing the forecasting performance of different specifications is future work of this project.

Author: Xiaoyi Zhou

Title: Predicting PM2.5 in Shanghai Using ARMA Model

Background: In recent years, air pollution in China is increasingly serious, which significantly affects human life and health [1]. Particulate matter (PM) represents a mixture of solid particles and liquid droplets in the air. In particularly, fine particles with diameter less than 2.5 μm are called PM2.5, which are reported as major pollutant of air pollution [2]. Prediction of PM2.5 plays an important role in control and reduction of pollutant in air [2]. A research [2] has been done to predict the concentration of PM2.5 in a couple of Chinese cities. However, the research just designed models based on concentration of PM2.5 time series data. In this project, two more covariates—humidity and temperature were considered.

Objective: Air quality index (AQI) in Shanghai 2016 data were downloaded from U.S. Department of State. An ARMA model was constructed to reach the goals. The aims of this project were as follows: 1. Determine the relationships between concentration of PM2.5 and covariates (humidity and temperature). 2. Predict concentration of PM2.5 based on historical data.

Expected results: The temperature was deleted from the model due to it was not correlated with concentration of PM2.5; The higher humidity result in higher concentration of PM2.5. Predicted concentration of PM2.5 followed the trend and seasonality in historical concentration data.

Author: Yue Zhou

Title: Detecting the Relationship between Satisfaction of Students and Gender of Instructors at NDSU

It is well known that the students' rating of instructor (SROI) is biased in general. This study was aimed at answering whether there exists gender bias against female instructor at NDSU. Approximately 30,000 feedback of SROI from 8 institutions during 2013-2014 were studied in this study. A proportional odds model for the students' ordinal categorical ratings was proposed towards this end.