DEPARTMENT OF

STATISTICAL & ACTUARIAL SCIENCES

July 25, 2019

15th Annual MSc Day

Schedule of Events

9:00 am — **Poster Session I Setup**

9:25 am — **Opening Remarks, Department Chair**

9:30 am — **Poster Session I**
Nazanin Afghan, Hadi Vafadar-Moradi, Golara Zafari, Johanna de Haan Ward, Jun Fang, Madhusoodan Gunasingam, Utsa Karan, Shanshan Liu, Bowei Zhang

11:00 am — **Poster Session I Cleanup**
Post Session II Setup

11:25 am — **Poster Session II**
Shima Ahmandi, Mojtaba Dargahi, Sahba Salarian, Shirzad Mohajerani, Azar Eftekhari Targhi, Sean Farquharson, Xinyu Jiang, Jiliang Li, Yulin Wang

12:55 pm — **Poster Session II Cleanup**

1:00pm — **Lunch in the Atrium**
(Physics & Astronomy 1st floor Atrium)

2:00pm — **Closing Remarks & Awards, Graduate Chair**

Abstracts

Poster Session I

Nazanin Afghan, Financial Modelling, supervised by Lars Stentoft

Title: Internship report on auditing Trade Floor Risk Management (TFRM) with the scope of Market Risk Valuation and Governance at Scotiabank's Audit Department Summer 2019

Abstract: This report summarizes the Audit Department's scope and responsibility, the functionality of Trade Floor Risk Management, and my experience during the internship at Scotiabank. The Audit Department, as the third line of defense, is responsible for providing independent assurance reports of the effectiveness of operational procedures, risk management, control function, and governance processes and procedures to the management. Trade Floor Risk Management (TFRM) provides market risk oversight for the Bank's trading, balance sheet, and liquidity management activities. TFRM also makes sure that the above-mentioned activities are in line with the Bank's risk appetite. During my internship, I supported Global Banking and Markets (GBM) Audit group in auditing TFRM and developed a tool using python to assist GBM's operation group to check whether email conversations between GBM employees of the Bank contained any sensitive word, which is not in line with the Bank's professional communication policies.

Johanna de Haan Ward, Statistical Sciences, supervised by Simon Bonner and Doug Woolford

Title: Plastic in our Waterways: Characterizing Plastic Resin Pellet Pollution on Great Lakes Beaches

Abstract: Plastic resin pellets are an industrial product manufactured to be molded into final products for consumers. Increasingly, they are being found in the waterways and coastlines of the world, presumably released during manufacturing and transport. Recently, such pellets have been identified by the thousands on the shores of the Great Lakes. Plastic pellets in water and on beaches have a negative impact as wildlife confuse these for food, resulting in illness or death. Additionally, the pellets act as a vehicle for pollutants and can release toxic substances into the water. The aim of this project was to determine which geological and anthropogenic factors influence the number of pellets found at a given location using data collected concerning microplastic abundance on beaches around the Great Lakes. Two count models were investigated to accurately model the number of pellets occurring at each location to account for the relatively high number of locations where no pellets were observed: the zero-inflated model and the hurdle model. The covariates considered in these models include human-related factors such as population, proximity to industrial plastic facilities and major transport routes, as well as environmental factors such as the shape of the coastline, the grain size of the beach sand and the proximity to a river mouth.

Jun Fang, Statistical Sciences, supervised by Jiandong Ren

Title: Statiscal Methods in Claim Reserving

Abstract: In this project, we consider various statistical methods used for loss reserving in property/casualty insurance. First, we introduce the classical Chain-Ladder method and then investigate using curve fitting for estimating the loss reserves. Second, we apply the Generalized Linear Model in loss reserving; it is shown that GLM method produce results like those from the basic CL method. However, GLM may provide statistical characteristics of the model parameters such as variance. Therefore, a confidence interval for the loss reserves can be derived. In addition, the variability of the model parameters is evaluated though bootstrap method. By using the Chain Ladder package in R, these techniques are illustrated with numerical examples.

Madhusoodan Gunasingam, Statistical Sciences, supervised by Hristo Sendov

Title: On Distribution of Zeroes in the Complex Plane

Abstract: In 1943, M.Kac began research which investigates the properties of roots of random polynomials as the degree is taken to be arbitrarily large. Here we define a random polynomial as a complex valued function with finitely many roots and random variable coefficients. He considered polynomials of degree n whose coefficients are independent and identically distributed real normal variables and showed that the expected number of real roots of such a polynomial is $\frac{2\pi}{n+1}\log(n+1)$. Over time, many renowned mathematicians have contributed and even improved upon this result. Others have worked on generalizing with different conditions and for various distributions of random coefficients. A current objective found in the literature is to understand the asymptotic distribution of complex roots of complex random polynomials of the form $G_n(z)=\xi_0+\xi_1z+\dots+\xi_nz^n$ as the degree is taken to be arbitrarily large. We particularly turn our focus to the work published by I.Ibragimov and D. Zaporozhets in 2013, which shows under weak constraints imposed on the coefficients, the roots of $G_n(z)$ tend to concentrate around the unit circle. The principle objective of this article is to provide a historical account of related results and to see how they have evolved over the years. Furthermore, it is to provide a self-contained, and complete understanding and justification of the last three results of I.Ibragimov and D. Zaporozhets.

Utsa Karan, Statistical Sciences, supervised by Ricardas Zitikis

Title: Bridging natural and machine intelligence for calculating insurance premiums

Abstract: InsurTech refers to the marriage of insurance and technology with a goal to transform the insurance industry by using modern technologies to generate efficiencies. It is proposed to develop new technology based solutions for more realistic insurance premium calculations using data analytics based on individual customer behavior patterns. Applications such as activity trackers, in-car mobility devices are already being used by insurance companies to monitor and manage insurance coverage and payment. However, opportunities exist to bring a transformational change and explore options of developing cost effective links with other emerging technologies to minimize the possibilities of frauds, enhancing customer experience and, simplifying internal insurance processes including but not limited to premium determination. With limited existing research done hitherto, this research project aims to explore the use of Artificial Intelligence (AI) and Machine Learning (ML) to examine real-life and real-time observations of data to develop premium calculation principles that are better aligned with real life.

Keywords: InsurTech, FinTech, Artificial Intelligence, Machine Learning

Shanshan Liu, Statistical Sciences, supervised by Doug Woolford

Title: Evaluation on the effectiveness of a Speed/Weight Indicator for Wildland Fire Response

Abstract: In the Province of Ontario wildland fires used to be managed through a set of pre-defined Fire Management Zones, which had pre-defined response objectives on fires (e.g., fully suppress, monitor…). The response decision for a fire depended only on its location. In recent years, the management strategy of fires changed from location-based to situation-based, where ecological, spatial features, the infrastructure of the fire regions, fire behaviour and so on are considered in fire response decision-making. The Speed/ Weight Indicator (SWI) is an index that was developed to provide a one number summary of key variables that should drive response, namely the current fire behaviour, the initial attack distance, weather and values in the local area of a given fire at a given date. The higher the SWI, the more attention the fire should attract, which in turn should result in a shorter response time (defined as the time between report time and the onset of initial attack efforts). In this project we evaluate the effectiveness of the SWI. We present initial analyses using survival analysis methods for modeling response time that show that the SWI is an effective index indicating the degree of concern for responding to a fire. Ongoing work includes assessing the amount of response effort that is applied.

Hadi Vafadar-Moradi, Financial Modelling, supervised by Matt Davison

Title: Validation of wholesale credit risk models

Abstract: Risk is inherent in any financial models because model outputs are estimates that rely on statistical techniques and data and/or mathematical approximations to simulate reality or provide estimates of future outcomes. Model risk also arises from potential misuse of models by the users. Risk models are primarily anchored on the three components of Expected Credit Loss (ECL) estimate namely, 1) Probability of Default (PD) during lifetime, 2) Loss Given Default (LGD) accounting loss forecast derived from economic loss forecast, and 3) Exposure at Default (EAD). Model development (MD), model validation (MV) and corporate audit are the 1st, 2nd and 3rd lines of defense in the risk management strategies of an enterprise. Model validation is an integral component of the bank's well-established model risk management framework that manages model risk in the model life cycle. Part of the second line of defense, model validation provides 1) independent validation of model and effective challenge to its conceptual soundness, 2) approval / rejection of model and 3) ongoing monitoring of implemented model. Monitoring is value added, especially if there is material deterioration in model performance and / or changes in regulatory requirement, products, or market conditions necessitating re-assessment of the model. As a part of MV group at Bank of Montreal Financial Group (BMO), I have reviewed the performance monitoring report for the wholesale Advanced Internal Rating-Based (AIRB) models and also the annual review of International Financial Reporting Standards - Financial Instruments (IFRS 9) which documents the prediction of PD, LGD and EAD for different portfolios and assess the model performance ratings and compliance with appropriate regulatory standards. Performance monitoring report documents credibility of the implemented models based on the realized PD, LGD and EAD over one quarter and determines if any additional step proposed by MV has to be taken by the MD. Linear and logistic regressions have been widely used in these models. In particular, linear regression has been used to predict the LGD and EAD based on the macro-economic variables like gross domestic product (GDP), S&P 500, S&P/TSX composite index. Logistic regression has been widely used to predict the probability of default as a function of macro-economic variables. Thus, I found Regression and Advanced Data Analysis courses extremely useful in the industry.

Golara Zafari, Financial Modelling, supervised by Marcos Escobar-Anel

Title: On the Simulation of Robust Portfolio Optimization of an Ambiguity Averse Insurer

Abstract: Among different portfolio optimization techniques, robust portfolio optimization secures the investor against the worst-case scenario. In this project, we consider an ambiguity-averse insurer, who finance its surplus by means of re-investing it on three different assets, i.e., risk-free bank account, a mean reverting commodity and a risky bond. This problem relies on solving some PDEs and MATLAB is used to find their numerical solutions. Next, these solutions are used to simulate the optimum strategy, which maximizes the minimum expected utility. In practice, however, one may have an incorrect estimate of parameters used in the model or ignore them for the sake of simplicity. The strategy obtained by incorrect or ignored parameter is known as sub-optimal strategy and results in loss. To this end, we consider several sub-optimum cases and calculate their corresponding losses, which give an insight to the investor on how different parameters impact the surplus.

Bowei Zhang, Actuarial Sciences, supervised by Ian McLeod

Title: Supervised Machine Learning Methods in Classification of High Dimensional Data

Abstract: High dimensional data is essential for its capacity to include enough information yet difficult to classify due to its complexity. In this paper, several machine learning methods are applied in classifying the phoneme dataset extracted from TIMIT database. This dataset has 4509 observations and each observation is a vector of dimension 256 which is transformed from the log-periodogram of a phoneme. The unsmooth and erratic shape of the vector further increases the difficulty of classification. With five classes labeled for the dataset, the error rate in the test set and computation time are used to evaluate the performance of each model. The paper contains models using single algorithm to classify the inputs into five classes as well as models combining two algorithms to classify the inputs into five classes. Based on the experiment results,modelscombining twoalgorithms on average generate higher classification accuracy but are more time-consuming, among which the model combining the eXtreme Gradient Boosting (an advanced boosting algorithm that produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees) and Convolutional Neural Network (a class of deep neural networks that takes advantage of the hierarchical pattern in data and assemble more complex patterns using smaller and simpler patterns) method obtains the most satisfying result of the classification of the dataset.

Poster Session II

Shima Ahmandi, Financial Modelling, supervised by Lars Stentoft

Title: Evaluation of retirement portfolio using conditional Monte Carlo simulation

Abstract: The objective of this project is to evaluate risk characteristics of the retirement portfolio which consists of risk reduction pool (option strategy), low volatility fund and bond fund using Monte Carlo simulation. All these fund components are a function of the underlying stock return. However, the risk reduction pool is not a simple linear function of the underlying stock return and it should be evaluated through time on each simulated path. It should be considered that each path is also regime dependent and conditional on what happened prior to that point. To calculate the risk characteristics of the portfolio such as three months rolling returns, first covariance and transition matrix on each regime are estimated according to historical data. Then, the weekly simulated underlying stock return is generated which will require mean reverting adjustment. Afterward, it fed into the option strategy and finally the weekly profit and loss of the strategy with respect to simulated underlying stock return on each path is evaluated.

Mojtaba Dargahi, Financial Modelling, supervised by Mark Reesor

Title: 3-factor Loan Portfolio Credit Risk Model

Abstract: Traditional banking business takes deposits and lends on these deposits, thus bringing together borrowers and lenders and making profit on the interest rate differential. This business is not without risk. The focus of this project is on the credit risk associated with retail and commercial loan portfolios. Banking regulators require measures of this credit risk for capital requirements as do internal risk management processes.

In simple word, generally, Credit risk is defined as the risk of loss resulting from an obligor's inability to meet its obligations and it is the largest source of risk faced by banking institutions world-wide. There are three basic components of credit risk on an obligor level:

• exposure at default (EAD), the amount to which the bank was exposed to the obligor at the time of default,

• loss given default (LGD), the proportion of the exposure that will be lost if a default occurs,

• probability of default (PD) within a fixed time horizon, the probability that the obligor will default on his loan in a certain period, usually a year.

Historical evidence from retail banking data has indicated that PD, LGD, and EAD are not independent. This means that credit loss models which ignore this correlation will underestimate credit losses. Recent work by Avusuglo, Metzler, and Reesor provides a deep investigation into the properties of 2-factor models involving PD and LGD correlations.

In this project, we will extend the above mentioned 2-factor model framework to 3-factor models for PD-LGD-EAD dependency. This will be a first step in extending this work to 3-factor models. We will perform Monte

Carlo simulations and other numerical work to investigate the properties of this proposed model. The main attention will be given to understanding the effect of account-level correlations on the portfolio-level quantities of interest, namely default rate and portfolio loss given default.

Keywords: Loan Portfolio, Credit Risk Model, 3-factor models

Azar Eftekhari Targhi, Statistical Sciences, supervised by Serge Provost

Title: The Generalized Pearson Family of Distributions and Multivariate Extensions

Abstract: A moment-based density approximation technique that is based on a generalization of Pearson's system of frequency curves is introduced. More specifically, the derivative of the logarithm of a continuous density function is expressed as a ratio of polynomials whose coefficients are determined by solving a linear system, and a closed form representation of the resulting density function is provided. It is then explained that, when used in conjunction with sample moments, the methodology being herein advocated can be utilized for the purpose of modeling sets of observations, including those referred to as 'Big Data'. Bivariate extensions are considered as well. Several illustrative examples are presented.

Sean Farquharson, Statistical Sciences, supervised by Doug Woolford

Title: Overdispersion in Count Modelling with Applications to NHL Draft Data: Predicting Player Success

Abstract: Many hockey players are drafted by National Hockey League (NHL) teams each year. Unfortunately, some drafted players never actually get to play in an NHL game. Wishing to model what characterizes successful players, we present an analysis of NHL draft data. We first fit count models to predict the number of games played in the NHL from the time a player is drafted. However, the mean-variance relationship in the Poisson model is violated due to excess zeros causing overdispersion, where the data exhibits greater variability than expected. Statistical tests for overdispersion are implemented in R to show why common count models are not appropriate such a situation. We briefly discuss overdispersion in the case of zero-inflated data and outline modelling solutions to this problem. We investigate this issue using visual and numerical diagnostics and account for this in the model building process. A Hurdle model (two-part model) is fitted to the data to address the issue of our zero-inflated data set. We concluded that a Hurdle model framework is best fitting for our data, as it addresses the issue of overdispersion due to zero-heaviness by combining the power of a Bernoulli and a zero-truncated count distribution.

Xinyu Jiang, Statistical Sciences, supervised by Doug Woolford

Title: Analysis and Prediction of Ontario's Wildland Fire Cost

Abstract: Wildland fires pose great threats to public safety, property, forest resources and other economic assets, despite ecological benefits. Thus, they can be extremely costly. The total cost of fires includes cost of fire prevention and control, cost of firefighting, direct and indirect economic losses from fires and so on. This project focuses on characterizing and modelling the costs of responding to individual wildland fires in Ontario. Response actions are stratified by full suppression and monitored, where full suppression fires are further split by whether the initial attacks on fires are successful. Our objectives are to identify the main factors and relationships that historically driven costs using all information about the recorded fires and make a prediction about the cost of a fire at the report time using only the available information at that time, with the aim to support fire management decision making. Preliminary analyses found that monitored fires were mainly distributed in the high latitude areas of Ontario where the density of human activity and assets was low. Monitored fires were also found collectively in certain areas such as parks and islands. Most of them were caused by lightning and the cost of monitored fires are much higher in the Northwest Region than the Northeast Region on average. Durations and final sizes were found to be two main factors that have driven the cost of monitored fires. However, the vast majority of wildland fires were fully suppressed. Although most of them were responded quickly and incurred low response costs, fully suppressed fires were much costlier than monitored fires on average. The number of air tankers sent was found to be the most important factor that has historically driven costs of fully suppressed fires, regardless of districts. Future work will continue to focus on modelling, with the emphasis on creating confidence bands for the cost prediction of fully suppressed fires, comparing results across districts, and merging district models into a single province-wide model.

Jiliang Li, Statistical Sciences, supervised by Doug Woolford & David Stanford

Title: A study of the mortality of patients transported by Emergency Medical Services in Northwestern Ontario

Abstract: Patients in rural areas can find it difficult to access timely care. The time until care is especially crucial for critically ill patients, such as those triaged under the Canadian Triage and Acuity Scale (CTAS) as level 1 or 2. We examine the relationship between remoteness from hospital and mortality through an analysis of ambulance and hospital data from northwestern Ontario. The purpose of the project is to determine how the distance of patients transported and travel time by Emergency Medical Services to Emergency Department (ED) affects the 7-day mortality risk of patients. Logistic regression models were developed for groups of patients stratified by key predictors such as CTAS score, age and gender in an attempt to verify the hypothesis that mortality risk increases as distance to ED increases. However, counterintuitive results occurred; the farther from the ED, the lower the mortality risk. We suspect this is due to the censoring of data from patients who died while being transported to the ED. To investigate this further, we simulated the process of censoring by first generating completed records of patients with mortality increasing against distance to ED, and then removing patients flagged as died en route. A reversed relationship between the distance to ED and mortality risk were found again. We concluded that the censoring of data of patients who died en route likely concealed the true relationship between the distance to ED and mortality risk.

Shirzad Mohajerani, Financial Modelling, supervised by Mark Reesor

Title: Report for the summer internship in the AML/ATF and Sanction Internal Audit department at Scotiabank

Abstract: Criminals' attempt to use financial institutions to launder funds can risk their reputation and ultimately their safety and soundness. Therefore, over the past years, governments acted extensively to implement effective and permanent Anti-Money Laundering (AML) and Anti-Terrorist Financing (ATF) programs. AML/ATF and Sanction Internal Audit department at Scotiabank provides independent oversight and internal assessment of different business lines including Canadian Banking (CB), International Banking (IB) and Global Banking and Markets (GBM) to ensure the compliance with the OSFI's AML/ATF program, PCMTFA (Proceeds of Crime, Money Laundering and Terrorist Financing Act) and internal policies and procedures. In this report, more details about AML/ATF program, as well as some aspects of Audit methodology and my contribution to the team during my internship in this department is provided.

Sahba Salarian, Financial Modelling, supervised by Matt Davison

Title: Model Risk Management Focusing on Wholesale Credit Risk Model Validation

Abstract: Risk is inherent in all financial models since the outputs are estimates relying on data, statistical techniques and/or mathematical approximations to simulate reality and provide estimates of future outcomes. Model risk also arises from potential misuse of the models. As a fundamental necessity for financial corporations, model risk management includes the steps taken for identification, measurement, management, mitigation, monitoring and reporting of this model risk. Model risk management applies to the whole model life cycle and covers the key stages of initiation, data assessment, development, validation, implementation, model usage, ongoing monitoring, and decommissioning. Both the appropriateness and the adequacy of a model for its intended purpose across the model life cycle are assessed using Model Validation (MV) activities and processes. By means of credit risk quantification methodologies such as Probability of Default (PD), Loss given Default (LGD) and Exposure at Default (EAD), combined with economical and qualitative factors, MV provides effective assurance of the reliability of the model in compliance with the corporate model risk policies and the regulatory requirements. As a member of MV-Wholesale group at the Bank of Montreal, I have conducted the quarterly performance monitoring of Borrower Risk Rating (BRR) models and back-testing of PD parameters by assessing the required key performance indicators of Advanced Internal Rating-Based (AIRB) and International Financial Reporting Standards - Financial Instruments (IFRS 9) models. IFRS9 which is inaugurated after the financial crisis of 2008, is the current standard reference for clear documentation as well as classification, measurement and impairment of financial assets. Among the multiple layers of model risk assessment, quarterly performance monitoring focuses on post-implementation assessment of the performance of BRR models based on the realized PD, LGD and EAD over the past quarter. BRR models include both qualitative and quantitative factors necessary for proper default risk assessment of different borrowers at different sectors of industry. These models are developed based on a combination of statistical analysis and expert judgment. Macro-economic data such as gross domestic product (GDP), S&P 500 and S&P/TSX composite index as well as historical default counts and exposures are important inputs for linear and logistic regression analyses involved in these models.

Performance monitoring of BRR models analyzes the extent of credibility of these models and determines if any modification is necessary at any stages of the model risk prediction.

Yulin Wang, Statistical Sciences, supervised by Ricardas Zitikis

Title: Momentum trading strategy and time series prediction of stock market

Abstract: This project studies two types of financial anomalies: Momentum and Reversal effects, whose existence has been confirmed by scholars from various countries. However, the traditional financial theory does not provide a reasonable explanation of these effects. Using data from S&P 500 index to create the winner-loser combination, this project is devoted to empirical research of the aforementioned two effects in the US stock market. We aim at answering questions such as: Do these effects really exist in the stock market? Is the US stock market an effective market or not? What kind of trading strategy can be more profitable? On the other hand, we employ various machine learning methods to predict the stock-market prices. We also compare the performance of various methods, such as Long Short-Term Memory Neural Network (LSTM) and the ARIMA time-series models.

Keywords: Momentum effect; Reversal effect; Time series models