The 22nd M.Sc. Day

DEPARTMENT OF
STATISTICAL & ACTUARIAL SCIENCES

July 30, 2026

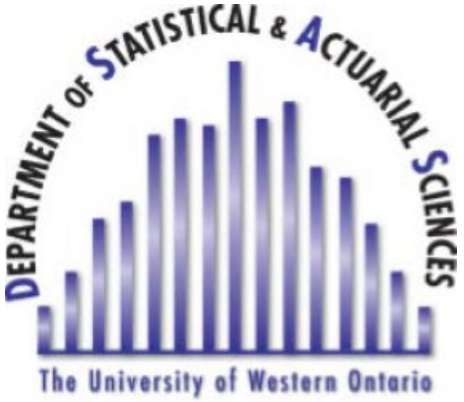

Table of Contents

Schedule of Events

Location: North Campus Building - 284

Time	Poster number	Last Name	First Name	Degree	Area	Supervisor
9:30-9:50		**MORNING COFFEE**				
9:50-10:00		**OPENING REMARKS**				
Session 1 10:00-10:50	1	Tang	Liangdan	MSc	AS	Dr. Ankush Agarwal
	2	Chen	Jiaye	MSc	FM	Dr. Lars Stentoft
	3	Deng	Jiyao	MSc	FM	Dr. Marcos Escobar-Anel
	4	Hong	Edward	MSc	FM	Dr. Lars Stentoft
	5	Li	Ying	MSc	FM	Dr. Guowen Huang
	6	Fontyn	Kevin	MSc	ST	Dr. Grace Yi
	7	Liu	Jie	MSc	ST	Dr. Jay Gweon
	8	Peng	Chen	MSc	ST	Dr. Camila de Souza
10:50-11:10		**COFFEE BREAK**				
Session 2 11:10-12:00	9	Xiang	Yu	MSc	AS	Dr. Jiandong Ren & Dr. Shu Li
	10	Liang	Yichen	MSc	FM	Dr. Marcos Escobar-Anel
	11	Qi	Hanning	MSc	FM	Dr. Jiandong Ren & Dr. Shu Li
	12	Wen	Xinyu	MSc	FM	Dr. Marcos Escobar-Anel
	13	Wettasinghe	Charith	MSc	FM	Dr. Cristian Bravo
	14	Raeisi	Mohammadreza	MSc	ST	Dr. Kristina G. Stankova & Dr. Amin Hassanzadeh
	15	Ren	Ziheng	MSc	ST	Dr. Grace Yi
	16	He	Qiduo	MSc-T	ST	Dr. Camila de Souza
12:00-1:00		**PHOTO TAKING + LUNCH IN WSC 263 & THE GRAD LOUNGE WSC 277**				
Session 3 1:00-1:50	17	Yin	Qingrong (Selena)	MSc	AS	Dr. Shu Li
	18	Xu	Chenhao	MSc	FM	Dr. Hao Yu & Dr. Reg Kulperger
	19	Zhang	Jiaxin (Yvonne)	MSc	FM	Dr. Marcos Escobar-Anel
	20	Zhang	Xi	MSc	FM	Dr. Cristian Bravo
	21	Snoddon	David	MSc-T	ST	Dr. Doug Woolford
	22	Tancredi	Bruno	MSc	ST	Dr. Simon Bonner
	23	Wang	Haiyue	MSc	ST	Dr. Wenqing He & Dr. Grace Yi
1:50-2:10		**CLOSING REMARKS**				

Titles and Abstracts

Liangdan Tang, AS, supervised by Dr. Ankush Agarwal

Actuarial Internship at WSIB: Supporting Valuation Process Through Experience Studies, Data Validation, and Workflow Automation

Currently, I am completing my internship as an Actuarial Summer Student at the Workplace Safety and Insurance Board (WSIB), an organization that provides workplace insurance coverage, wage-loss benefits, medical support, and return-to-work services for over five million workers across more than 300,000 workplaces in Ontario. This presentation provides an overview of how actuarial valuation supports WSIB's financial sustainability by estimating future obligations arising from workplace injuries, assessing funding sufficiency, and supporting financial reporting and planning. It also covers my contributions to this process across three key areas: conducting an Incurred But Not Reported (IBNR) experience study for assumption setting; performing data integrity checks for quarter-end reporting; and automating data preparation workflows using Power Query, Power Pivot, and DAX.

Jiaye Chen, FM, supervised by Dr. Lars Stentoft

GARCH-Based Pricing of European SPX Options under Non-Normal Innovations

This project investigates whether allowing for non-normal return innovations can improve the pricing of European S&P 500 index options within a GARCH framework. Traditional option-pricing models often rely on normally distributed returns and constant volatility, assumptions that are inconsistent with several well documented characteristics of equity-index returns, including volatility clustering, persistence, negative skewness, and heavy tails. To address these limitations, this study places particular emphasis on the distributional assumption of the innovation term and compares a standard Gaussian GARCH model with a GARCH model based on Student's innovations.

Daily S&P 500 index prices from January 2000 to August 2025 are used to construct log-return and excess-return series and to estimate the conditional volatility process. The effective federal funds rate provided by the Federal Reserve is used as a proxy for the short-term risk-free rate, while the OptionMetrics index dividend yield is incorporated to reflect the dividend-paying nature of the S&P 500 index. The option pricing sample consists of daily European SPX call and put quotations from January 2020 to August 2025. The option data include strike prices, expiration dates, bid and ask quotations, trading volume, open interest, and implied volatility.

The estimated GARCH specifications are incorporated into a risk-neutral option pricing framework to generate model-based option prices. These prices are compared with observed market mid-quotes and with prices obtained from the Black–Scholes model, which serves as a conventional benchmark. Pricing performance is evaluated using measures such as mean absolute error, root mean squared error, and relative pricing error. The analysis also examines whether pricing accuracy differs across option moneyness levels, times to maturity, and periods of different market volatility.

Particular attention is given to high-volatility periods, during which the Gaussian assumption may be especially restrictive. The main objective is to determine whether the use of heavy-tailed innovations provides a more realistic representation of SPX return behavior and leads to more accurate and stable pricing of European index options. The study also evaluates whether the potential advantage of the Student's specification is concentrated in particular option categories or market conditions.

Jiyao Deng, FM, supervised by Dr. Marcos Escobar-Anel

Beyond the VIX: Incorporating the CBOE SKEW Index into the Joint Calibration of the Heston–Nandi GARCH Model on the S&P 500 and VIX

This project assess the importance of the CBOE SKEW index on pinning down the distribution of S&P 500 returns in the presence of VIX. A Heston–Nandi GARCH model and a Black–Scholes benchmark are calibrated to daily data from 2010 to 2019 under three nested information sets: returns only, returns with VIX, and returns with VIX and SKEW. We augment the Gaussian return quasi-likelihood with inverse-variance-weighted least-squares targets for the two indices. Naturally, fitting the VIX improves the variance fit but worsens the skew and returns fit, while adding SKEW reduces skew loss at a cost to variance and return fits. This is because the second and third moments compete for the same parameters. Nonetheless, the overall fitting is greatly improved when including VIX first and then SKEW to the calibration. This motivates the usage of extra data, simpler than options prices, as well as richer state-space extensions with enough degrees of freedom.

Edward Hong, FM, supervised by Dr. Lars Stentoft

A Multi-Stage Machine Learning Framework for Financial Fraud Detection: From Customer Onboarding to Transaction Monitoring

Financial fraud remains a major challenge for financial institutions as digital banking, online payments, and increasingly adaptive fraud strategies continue to expand. While many fraud detection studies focus on transaction monitoring after an account or card has become active, fraud risk can also emerge earlier during the customer onboarding stage. This project develops a multi-stage machine learning framework that examines fraud detection across two points in the financial customer lifecycle: pre-approval onboarding fraud and post-approval transaction fraud. The first model uses the Bank Account Fraud (BAF) Base dataset as a practical proxy for application-stage fraud risk, while the second model uses a simulated credit card transaction dataset to represent post-approval transaction monitoring. For both stages, Logistic Regression, Random Forest, and XGBoost models are trained with class weighting to address severe class imbalance. Model performance is evaluated using ROC-AUC, PR-AUC, precision, recall, F1-score, and validation-selected false positive rate (FPR) thresholds. The results show that XGBoost provides the strongest performance in the onboarding fraud model, while a conservative XGBoost specification is selected for the transaction fraud model after removing identity-like and highly specific variables. The findings suggest that fraud detection requires different data, features, and

operating thresholds depending on where fraud occurs in the customer lifecycle. By combining onboarding screening with transaction monitoring, this project demonstrates how a multi-stage machine learning framework can support a more comprehensive approach to financial fraud prevention.

Ying Li, FM, supervised by Dr. Guowen Huang

Neural Networks as Flexible Score Models for Case-Crossover Analyses

Case-crossover models are widely used to study short-term environmental health effects because they compare exposure conditions at event times with nearby referent times from the same stratum. Standard analyses usually use linear score functions, which may be too restrictive when exposure-response relationships are nonlinear or when interactions among pollutants and meteorological factors are present. This project reformulates the case-crossover likelihood as a matched-set classification problem and replaces the usual linear score with a feedforward neural-network score. The proposed model preserves the case-crossover design while allowing nonlinear exposure-response patterns to be learned from the data. In simulation studies, when the true exposure-response relationship was linear, the three models produced similar results. In contrast, when the true relationship was nonlinear, the proposed neural-network case-crossover model outperformed the linear and spline-based alternatives. Future work will apply this framework to real-world environmental health datasets to evaluate whether neural-network score models can improve exposure-response estimation in practical air-pollution and mortality studies.

Kevin Fontyn, ST, supervised by Dr. Grace Yi

Some Results on Constrained Curve-Switching Over a Finite Horizon

Many multi-objective systems require a decision-maker to choose among competing options, each tracking a continuous outcome (e.g., value, utility, efficiency or reward) to be maximized over a finite horizon. However, within continuous systems where multiple options contend, it is often the case that no single outcome remains optimal over the horizon, and thereby a decision-maker may wish to switch among the options dynamically, following their initial commitment. While instantaneous switching risks either breaking the objective (i.e., remaining maximal) or introducing discontinuity, we constrain switching exclusively to points of parity, i.e., to the intersection points of the system where one's options are said to momentarily "agree". This constraint renders switching inherently path dependent. This project addresses two fundamental questions about these constraints. First, from a fixed initial decision, is there a path through the outcomes to the best one reachable at the end of the horizon? Second, is there a path that is indeed never bettered by others? We present a complete algorithmic framework for solving both. The key observation is that by enforcing parity-switching constraints, a continuous optimization collapses onto a finite combinatorial structure (a directed acyclic graph), over which these questions about paths are resolved. The central result uncovers that finishing "best" is governed by reachability rather than which option is currently leading. Consequently, a globally optimal endpoint may require following a temporarily suboptimal option, and an option with a suboptimal endpoint may nevertheless

provide the only path to the global optimum. The results are presented in full generality, illustrated by two worked examples whose differences are discussed.

Jie Liu, ST, supervised by Dr. Jay Gweon

When Do Predicted Labels Harm MetaBR? Selective MetaBR for Multi-Label Classification

Multi-label classification assigns several binary labels to each observation. Binary Relevance (BR) predicts each label independently and ignores dependencies among labels, while MetaBR adds the information from other labels as extra features. This creates a mismatch between training and testing. MetaBR is trained on true labels but uses BR-predicted labels at test time, and those predicted inputs can be unreliable. We study Selective MetaBR, which admits only reliable label predictions as meta-level inputs, where reliability is defined as a BR validation F1 score above a threshold. We conduct experiments on several multi-label datasets. Our results show that when MetaBR relies on many unreliable predictors, excluding them tends to improve accuracy. This suggests that the selective use of label predictions helps most when MetaBR relies on noisy meta-level inputs.

Chen Peng, ST, supervised by Dr. Camila de Souza

Fast Variational Bayesian Inference for the Weibull Accelerated Failure Time Model

The Weibull regression model is a widely used accelerated failure time (AFT) model for analyzing right-censored survival data. Although likelihood-based inference is readily available, Bayesian inference for Weibull AFT models commonly relies on Markov chain Monte Carlo (MCMC), which can be computationally expensive. In this work, we develop an alternative mean-field variational Bayes (VB) algorithm for estimating the parameters of the Weibull AFT model. Extending an existing variational framework for the log-logistic AFT model, we apply a piecewise approximation to the non-conjugate term associated with the extreme-value error distribution, resulting in tractable coordinate-ascent updates. The proposed algorithm is evaluated through simulation studies under different sample sizes and censoring rates and is compared with maximum likelihood estimation and MCMC. The results show that VB yields nearly unbiased estimates with estimation accuracy comparable to that of the competing methods, while producing posterior summaries that closely agree with MCMC. Across the simulated scenarios, VB achieves an average speedup of 495 times relative to MCMC. Finally, we illustrate the proposed approach using the publicly available METABRIC breast cancer dataset to assess which clinical, tumor-related, and treatment factors are associated with patients' overall survival times, measured from intervention to death.

Yu Xiang, AS, supervised by Dr. Jiandong Ren & Dr. Shu Li

Variable Payout Annuities and Retirement Income Strategies

This project studies variable payout annuities (VPAs) in retirement income planning. VPAs provide lifetime income through longevity pooling, but payments vary with investment performance and

mortality experience. Compared with a fixed annuity, a VPA gives retirees the possibility of higher future income, but also exposes them to the risk of declining payments. In this project, we follow Boyle et al. (2015) to study retirement income strategies involving VPAs, fixed annuities, equities, and risk-free assets. In particular, under certain assumptions of mortality rate, retirees' utility function, and bequest motive, we simulate and compare the income, wealth, and consumption paths resulting from different strategies. Our results show that VPAs can help retirees maintain desired consumption patterns and improve their expected utility. In addition, our findings suggest that a moderate hurdle rate provides a better balance between retirement income and long-term sustainability.

Yichen Liang, FM, supervised by Dr. Marcos Escobar-Anel

Statistical versus Machine Learning on Corporate Default Prediction for USA Listed Firms: A Trade-Off between Discrimination and Interpretability

This study compares statistical and machine learning models for corporate probability of default (PD) estimation on USA listed firms, and examines whether the improvement of machine learning over traditional benchmarks is driven by the size of the feature set at the expense of interpretability. The sample covers 870,234 firm-quarter observations for 21,259 non-financial, non-utility firms from 1990 to 2023, drawn from Compustat and CRSP through WRDS. Four models are built on the same CRSP-matched panel: (A) the 1968 Altman Z-score with its original coefficients, (B) a logistic re-estimation of the Altman ratios, (C) a discrete-time hazard logit on 22 variables covering accounting and market indicators, and (D) an XGBoost classifier on the same 22 variables with SHAP interpretation. A five-variable XGBoost using only the Altman ratios is also fitted to separate the method effect from the feature-set effect. On the CRSP-matched test set (76,136 obs / 252 defaults), test AUC rises from 0.6711 for (B) to 0.8515 for (C) and 0.9443 for (D), while the five-variable XGBoost reaches a very high 0.9243. The machine learning method contributes about twelve times as much to discrimination as the additional variables, at the cost of about one hundred times more parameters. The results suggest that the choice between logistic and tree-based methods is a trade-off between discrimination and interpretability.

Hanning Qi, FM, supervised by Dr. Jiandong Ren & Dr. Shu Li

Risk-Neutral Valuation of Variable Annuity Guarantees with Monte Carlo Simulation

In this project, we study a risk-neutral valuation framework for representative variable annuity guarantees, including guaranteed minimum maturity benefits (GMMBs) and guaranteed minimum death benefits (GMDBs). Contracts with fixed and ratchet guarantee structures are considered. Fair guarantee fees are determined analytically for contracts with closed-form solutions and through Monte Carlo simulation for more general path-dependent contracts. Two classical variance reduction techniques, antithetic variates and control variates, are implemented and compared to improve simulation efficiency. An insurer-liability formulation is also presented as a complementary perspective.

Numerical experiments demonstrate good agreement between analytical and simulation results for contracts with closed-form solutions, supporting the accuracy of the proposed valuation framework. The results also show that ratchet guarantees generally require higher fair guarantee fees than comparable fixed guarantees because of their enhanced benefit design. The proposed framework provides a practical foundation for the valuation of representative variable annuity guarantees and offers a flexible basis for future extensions involving surrender decisions, stochastic interest rates, and other insurer risk management applications.

Xinyu Wen, FM, supervised by Dr. Marcos Escobar-Anel

Market Microstructure-Based Anomaly Detection in Prediction Markets Using Polymarket CLOB Data

Prediction markets are event-driven trading environments in which contract prices represent market implied probabilities of future outcomes. Compared with traditional financial markets, they often operate with thinner order books, lower liquidity, and stronger sensitivity to discrete information shocks, making them suitable for studying abnormal trading dynamics. This project develops a market microstructure-based anomaly detection framework using publicly available data collected from the Polymarket Gamma API and Central Limit Order Book (CLOB) API. The empirical analysis combines midpoint prices, bid-ask spreads, liquidity measures, trading volume, and event resolution dates to construct time series and market microstructure variables.

An unsupervised machine learning framework based on Isolation Forest is implemented to identify statistically unusual trading periods. The analysis first examines a high trading-intensity cryptocurrency-related prediction market associated with the MegaETH airdrop event and then extends the framework to a cross-market comparison across twenty active prediction markets grouped into sports, politics, cryptocurrency, and other event-driven categories. The empirical findings suggest that lower-liquidity and speculative prediction markets are more likely to experience large discontinuous price adjustments and elevated volatility spikes. In particular, cryptocurrency and rumor-driven markets exhibit stronger abnormal price jump severity relative to political and sports contracts.

Charith Wettasinghe, FM, supervised by Dr. Cristian Bravo

Structure of Federal Reserve Yields: 2007-2008

This paper presents an exploration of the structure of Federal Reserve nominal zero coupon yield curves from 2007-01-01 to 2008-12-31 using techniques from Topological Data Analysis (TDA). First, the yield curve is formulated and the set of curves is classified as a point cloud, which justifies the use of persistence (and zigzag) homology to study the structure of the Vietoris-Rips filtration constructed from it. Then, the structure of this filtration made from the point cloud is analyzed through its separate spatial and temporal evolution, in order to separate the corresponding spatial and temporal properties. This is done by analyzing the persistence and zigzag barcodes, respectively, that encode the structure as a unique geometric fingerprint. Both of these objects, along with their associated techniques, are carefully explained. The results of the spatial analysis

indicate that the point cloud of yields during the 2007-01-01 to 2008-12-31 period is characterized by fragmented clusters and loops. The fragmentation is linked to underlying market multi-modality and the loops are hypothesized to positively correlate with uncertainty. This hypothesis is derived from the observation that loops are quasi-symmetric structures, and that symmetry in data is often generated by high-entropy processes. Such processes could reflect uncertainty through the lack of directionality. Further temporal analysis attributes the formation of loops to the subprime mortgage crisis of 2007.

Mohammadreza Raeisi, ST, supervised by Dr. Kristina G. Stankova & Dr. Amin Hassanzadeh

Flood Susceptibility Mapping for the Greater Toronto Area Using HAND and Satellite Precipitation Data

This study maps relative flood susceptibility across the Greater Toronto Area using terrain analysis of freely available satellite data. The result is a screening tool that ranks land by its exposure to flooding, not a prediction of actual floodwater depths. The method is based on the Height Above Nearest Drainage (HAND) index, computed from the Digital Elevation Model (DEM). Each land cell is assigned its vertical height above the nearest stream; cells that are low relative to the drainage network are flagged as flood-susceptible. Under a reference flood stage set from the terrain statistics themselves, about 37.6% of GTA land cells fall into the susceptible category. Two statistical analyses were built on this map. A Generalized Extreme Value distribution summarizes the distribution of relative depths across susceptible cells, and a multivariate regression confirms that HAND dominates, with terrain controlling relative flood depth. Extending the observed precipitation trend to 2100 under low, median, and high cases suggests the fraction of GTA land cells classified as flood-susceptible could range from roughly 11% to 79% by century's end, with the median case near 44%. These projections extrapolate a short, marginally significant 21-year trend and span a wide range of uncertainty. The framework's value lies in serving as a transparent first-pass screen for identifying areas that warrant detailed hydraulic modeling.

Ziheng Ren, ST, supervised by Dr. Grace Yi

Domain Adaptation with Limited Attribute Observation

Unsupervised domain adaptation uses labeled source observations and unlabeled target features to learn under a shift in the data-generating distribution. This project considers the additional restriction that prediction must use a fixed subset of p0 attributes. For a given subset A, we formulate the adaptation problem on the observed feature space XA and specialize the standard domain adaptation bound to this space. Motivated by the source-risk and discrepancy terms in the bound, we rank attribute j by the score Ij −γDj , where Ij measures source predictive utility and Dj measures source-target discrepancy. We evaluate this criterion on twelve ordered transfer tasks constructed from four clinical centers in the UCI Heart Disease dataset. Among the evaluated configurations of the proposed method, the largest mean target balanced accuracy is 0.692, obtained with p0 = 10 and γ = 0.25. The gains are not uniform across observation budgets and

penalty values, indicating that the usefulness of the discrepancy penalty depends on the transfer task and the available attribute budget.

Qiduo He, ST, supervised by Dr. Camila de Souza

Geographic Distribution of Autism Service Providers in Ontario:Examining Urban–Rural Differences in Service Availability

The prevalence of Autism Spectrum Disorder (ASD) has continued to increase in recent years, and many autistic children and their families require long-term support. Despite publicly funded autism programs in Ontario, many families still face challenges in accessing timely Applied Behaviour Analysis (ABA) services. An uneven geographic distribution of Registered Behaviour Analysts (RBAs) may contribute to disparities in service availability across Ontario. This study aimed to examine the geographic distribution of RBAs across Ontario, evaluate urban–rural differences in service availability, and identify socioeconomic and sociodemographic factors associated with provider availability. Ontario RBAs'location data were linked with 2021 Canadian Census data at the Forward Sortation Area (FSA) level. Geographic Information Systems (GIS) were used to visualize the geographic distribution of RBAs. Negative binomial regression was used to examine sociodemographic and socioeconomic characteristics associated with RBA provider rates. RBAs were concentrated in major urban centres, including the Greater Toronto Area, Ottawa, and Southwestern Ontario; however, 39.23% of FSAs in Ontario had no RBAs, and 67.12% had fewer than one RBA per 10,000 residents. Urban FSAs had significantly higher provider rates than non urban FSAs, indicating disparities in provider availability between urban and rural areas. Results from the negative binomial regression indicated that urbanicity and the proportion of residents with a bachelor's degree were positively associated with higher RBA provider rates. These findings may inform future workforce planning and support more equitable access to ABA services across Ontario.

Qingrong (Selena) Yin, AS, supervised by Dr. Shu Li

A Clustering-Based Predictive Analytics Approach for Telematics Data

Telematics data provide a more comprehensive representation of driving behavior and are increasingly relevant for usage-based automobile insurance. However, their predictive performance may be affected when telematics data are limited and biased. This project investigates a clustering-based framework for claim-frequency prediction using a synthetic automobile insurance telematics dataset. K-means and Gaussian mixture models are applied to identify latent policyholder risk groups. After clusters are ranked by claim rates and assigned corresponding risk levels, the GMM and K-means risk labels are paired to construct joint risk cells. Different groupings are then incorporated into Poisson frequency models and evaluated under random, age-based, and favorable-selection settings. The results show that joint-label clustering improves out-of-sample claim-frequency prediction in the main random setting, particularly under the three-cluster specification. Under biased selection mechanisms, however, the predictive advantage becomes

less stable. These findings highlight the potential of unsupervised risk segmentation while emphasizing the importance of selection mechanisms in claim frequency prediction.

Chenhao Xu, FM, supervised by Dr. Hao Yu & Dr. Reg Kulperger

A VIX3M-Informed Markov Regime Framework for Tail-Risk-Aware Portfolio Optimization

Traditional portfolio models rely primarily on historical returns and time-invariant distributions, limiting their ability to respond to changing market conditions and tail risk. This study develops a VIX3M-informed Markov regime framework for dynamic portfolio optimization. VIX3M is used as a forward-looking measure of expected three-month market volatility to identify Low-, Medium-, and High-volatility states. Regime transitions and state-dependent multivariate return distributions generate 20,000 scenarios over a 63-trading-day horizon. At each quarterly rebalancing date, a long-only portfolio is selected by maximizing expected terminal log return relative to Value at-Risk. The study considers an opportunity set of four assets: green equity, broad equity, green bonds and aggregate bonds. It also uses VIX3M data from 2009 to 2026. The framework is evaluated through out-of-sample and stress tests. Compared with an IID normal benchmark, it produces more defensive allocations in high-volatility regimes and can reduce selected tail risk measures, although sometimes at the cost of lower returns.

Jiaxin (Yvonne) Zhang, FM, supervised by Dr. Marcos Escobar-Anel

Pricing GMDB: A Joint Framework Using HN-GARCH and Gompertz Mortality Models

Guaranteed Minimum Death Benefits (GMDBs) embedded in variable annuities expose insurers to joint financial and biometric risk. This study builds a joint pricing framework: the GARCH model captures equity volatility clustering and leverage effects, while Gompertz law and Lee-Carter model policyholder mortality. The financial model is calibrated to 6,538 daily S&P 500 log-returns (2000–2025), spanning the dot-com, 2008, and COVID-19 crises; the NGARCH best fits historical dynamics (AIC 17611.3), while HN-GARCH is retained for tractability. Mortality is calibrated to 35 years of US HMD data (1990–2024, ages 40–95), with about 2% difference in death probability at 65 between Gompertz and Lee–Carter. The fitted aging-rate κ implies mortality roughly doubles every 8.3 years. Our joint model shows that mid-contract years contribute the most to the total GMDB cost and the maximum cost is not at the peak mortality e.g., for a 65-year-old buyer, the fair premium is 7.35 per 100 of account value, with peak year cost at 19 and peak mortality at 22 years.

Xi Zhang, FM, supervised by Dr. Cristian Bravo

High-Risk Credit Strategy Performance Evaluation

Background: Credit strategy performance monitoring helps banks assess whether risk treatments are working as intended. This project reviews a credit portfolio at a Canadian bank where a new treatment approach was introduced for selected higher risk accounts. The objective is to assess

whether the treatment improves portfolio outcomes and reduces potential credit loss compared with the existing Champion strategy.

Method: A coding-based ETL workflow was used to extract, clean, join, and aggregate large-scale account-level data into monthly monitoring cohorts. The processed cohorts were then used for a Champion-Challenger comparison across portfolio products. Performance was tracked through key credit KPIs, including Cure Rate, Roll Rate, account counts, balances, and estimated dollar impact.

Results: The analysis found stronger performance in one major portfolio segment, where the new treatment improved Cure Rate, reduced Roll Rate, and generated a meaningful estimated dollar benefit. Other portfolio products showed mixed results and did not demonstrate a clear financial benefit during the monitoring period.

Conclusion: The findings support a proposal to fully roll out the new treatment for the stronger-performing portfolio segment, while continuing to monitor the remaining products due to mixed results. The next step is to work with relevant business and strategy stakeholders to validate the recommendation, align on implementation readiness, and continue performance tracking after rollout.

David Snoddon, ST, supervised by Dr. Doug Woolford

Spatial and Temporal Patterns of Wildfires in Ontario, Canada, 1960–2025

Wildfires are a major natural hazard in Ontario, wildfires impact ecosystems, communities, and natural resource management across the province. Despite their importance for wildfire management, comprehensive analysis of how long-term wildfire ignition patterns across space, time, season, and fire size remains limited. This study examined provincial wildfire records from 1960 to 2025 to characterize long-term spatial, seasonal, and temporal trends in wildfire occurrence, ignition sources, and fire size across Ontario. Human-caused fires declined substantially over the study period and were surpassed by lightning-caused fires after approximately 2000. Lightning-caused fires accounted for approximately 42% of all wildfires but represented over 60% of fires that were at least 10 hectares in size and nearly 90% of fires exceeding 10,000 hectares. Human-caused fires were highest on average in May and occurred predominantly in southern Ontario. In contrast, lightning-caused fires were highest on average in July and occurred predominantly in northwestern Ontario. Both recreational and residential fires declined over time. Recreational fires became increasingly dominated by campfire ignitions as smoking-related ignitions declined, whereas residential fires remained distributed among several ignition types. On average, Residential fires occurred primarily in April and May, whereas recreational fires peaked during July and August. Because the provincial database excludes fires managed solely by municipal fire departments, the apparent decline in the number of fires, particularly those within municipal boundaries, may partially reflect changes in database coverage. These findings improve our understanding of how wildfire ignition patterns vary across region, time, and month in Ontario and can help improve future wildfire management strategies.

Bruno Tancredi, ST, supervised by Dr. Simon Bonner

Extending Subsampling Approaches for Occupancy Modelling with Large and Skewed Datasets

Occupancy models are widely used in ecology, but their application to rare species in high-resolution datasets continues to present computational challenges. de Haan-Ward et al. (2025) proposed a method for fitting occupancy models where the original dataset is subsampled, keeping all sites with at least one detection along with a random sample of sites with no detections. We extend this method by allowing new sampling methods of the no detections sites. Due to the risk that random sampling may discard observations associated with infrequent covariate values, our techniques were specifically designed to preserve those observations. Accordingly, we compute the sampling probability based on the covariate of interest, assigning higher sampling probabilities to less frequent values either through stratification or a kernel density estimation approach. To test our new techniques, we designed a simulation study comparing the results of the different methods under different sample sizes, covering scenarios where covariates with skewed distributions are possible. Our results show that by using only 5% or 10% of the data, we can achieve results similar to those obtained using the entire dataset and how, under certain scenarios, the new sampling techniques have outperformed random sampling.

Haiyue Wang, ST, supervised by Dr. Wenqing He & Dr. Grace Yi

Statistical Learning and AI Methods for Analyzing Chronic Pain Data

This study analyzes chronic pain data from the SPARC workbook where both mixed clinical and self reported variables are available. The analysis builds on a questionnaire-distance framework developed for self-reported chronic pain data (Deng, 2025) and extends it with two AI models. The main clustering analysis used 17 clinical and self-reported variables after range checking and complete-case filtering. The final clustering cohort contained 812 patients. Complete-linkage hierarchical clustering with the questionnaire distance selected two clusters, with 419 patients in Cluster 1 and 393 in Cluster 2. A random forest model identified GAD-7, CSI, and PHQ-9 as the most informative variables for distinguishing the clusters. The two extended AI models gave different forms of information. UMAP followed by HDBSCAN found one small noise group and two patient subgroups, although its silhouette width was lower than that of the hierarchical clustering. XGBoost with SHAP interpretation predicted the two hierarchical clusters with a held-out AUC of 0.973 and again pointed to GAD-7, PHQ-9, and CSI as important variables. Additional analyses showed that active myofascial trigger point dominance was associated with higher central sensitization scores, and that CSI and MPD helped predict nociplastic pain. These results suggest that self-reported chronic pain data can be organized into clinically interpretable patient phenotypes when the distance metric can well account for the questionnaire structure.