

INTERNATIONAL CONFERENCE ON DATA SCIENCE 2023
CONFERENCE PROGRAM
 November 8 to 10, 2023 www.icds2023.cl

WEDNESDAY 8 NOVEMBER 2023

07:30	08:30	Registration	
08:30	08:50	Welcome Ceremony	
08:50	09:50	Keynote Speaker 1: Peter Filzmoser	Auditorium BNP
		Chair: Javier Trejos	

		Authors	Title	Abstract
		Peter Filzmoser , Marcus Mayrhofer, Una Radojicic, Horst Lewitschnig	Explainable outlier identification for matrix-valued observations	Outlier detection techniques are well established for multivariate observations (vectors). We extend the ideas to matrix-valued objects, where the measurements are arranged in the rows and columns of a matrix. An example are image data, where the pixel information is presented in a rectangular matrix. The concept of matrix-valued data is not new at all, and a prominent distribution in this context is the matrix normal distribution. There are different proposals in the literature on how to estimate the parameters of this distribution. It is also possible to define a Mahalanobis distance, and the concept of robust covariance estimation can be modified to obtain robust estimators for the matrix-valued case. We present an adaptation of the well-known MCD (Minimum Covariance Determinant) estimator to this situation. Moreover, the concept of Shapley values, which has been successfully used in the context of Explainable AI, is extended in order to explain the reasoning behind the outlyingness. For example, one can identify outlying images and explain which pixels contribute to this outlyingness. A more detailed background, as well as illustrative examples, will be provided in the presentation.

09:50	11:00	IPS Session: Multivariate Analysis and Statistical Learning Applications	
		Chair and organizer: Javier Trejos	Auditorium BNP

		Authors	Title	Abstract
		Alejandro Tenorio-Sánchez , Javier Trejos-Zelaya, Juan José Leitón-Montero	Reducing Complexity in Route Optimization Analysis of Weather Monitoring Networks: A Comparative Evaluation of MDS and t-SNE Techniques	Weather monitor networks play a crucial role in collecting meteorological data from multiple locations to understand current weather conditions, climate patterns, and atmospheric changes. To ensure the reliability and availability of data, proper maintenance of these stations is essential. However, maintaining a network of national scale poses challenges due to limited personnel and resources. This research aims to approximate an optimal strategy for visiting weather monitor network stations, enhancing their overall efficiency by combining dimensionality reduction techniques and clustering algorithms. Specifically, T-distributed stochastic neighbor embedding (t-SNE) and multidimensional scaling (MDS) are employed to create a low-dimensional space based on road distance similarities between stations. The results are compared with the use of k-means clustering and Travelling Salesman Problem (TSP) heuristics to assess the effectiveness of the low-dimensional space in terms of traveled distance. T-SNE exhibits a unique characteristic by grouping similar data points together in the low-dimensional space it constructs. Leveraging this property, it becomes a valuable tool for identifying stations that should be visited together, enabling superior scheduling and resource allocation for maintenance operations than MDS.

		Victor Adolfo Rojas Cruz	Association of task effectiveness expectations with academic performance in engineering and science college students.	The association of the expectations of effectiveness and the value of the tasks with academic performance in university students of engineering and sciences has been investigated. Three data collection processes were carried out on students of a calculus 1 course at the University of Costa Rica. The first consists of carrying out 10 cognitive interviews in order to improve the understanding of the scale, a pilot to 128 students that was applied to calculus students, the results of both processes yielded results that helped to improve the understanding of the scale and the structure of the same, the analysis that was made in the pilot was an exploratory factorial analysis, so that they showed substantial improvements at the theoretical level. And finally, the final application for a sample of 428 students, an analysis of structural equations was carried out, in which the differences related to gender in relation to achievement motivation were shown, the results obtained show a fairly good theoretical consistency, in addition to yield models in which the percentage of variance explained is high.
		Luis Eduardo Amaya Briceño, Jilberrés Urbina Calero, Edison Quesada Lopez	Inclusión de la dimensión geoespacial y contexto socioeconómico en la relación cliente-institución financiera	En la mayoría de modelos econométricos de asignación crediticia, se consideran variables que se centran en la relación cliente-institución financiera, dejando por fuera otras variables que involucran la interacción de variables relacionadas al contexto del cliente. En nuestro trabajo, compartimos nuestra experiencia desarrollada y resultados previos en la obtención, limpieza y uso de datos geoespaciales obtenidos con un recubrimiento de bolas, por medio de una Api de Google.
09:50	11:00	IPS Session: Robust Data Analysis and Prediction		Auditorium FSSH
		Chair and organizer: Christophe Croux		
		Local Chair: Fabrizio Ruggeri		
		Jakob Raymaekers, Peter J. Rousseeuw	The cellwise Minimum Covariance Determinant estimator	The usual Minimum Covariance Determinant (MCD) estimator of a covariance matrix is robust against casewise outliers. These are cases (that is, rows of the data matrix) that behave differently from the majority of cases, raising suspicion that they might belong to a different population. On the other hand, cellwise outliers are individual cells in the data matrix. When a row contains one or more outlying cells, the other cells in the same row still contain useful information that we wish to preserve. We propose a cellwise robust version of the MCD method, called cellMCD. Its main building blocks are observed likelihood and a sparsity penalty on the number of flagged cellwise outliers. It possesses good breakdown properties. We construct a fast algorithm for cellMCD based on concentration steps (C-steps) that always lower the objective. The method performs well in simulations with cellwise outliers, and has high finite-sample efficiency on clean data. It is illustrated on real data with visualizations of the results.
		Anthony-Alexander Christidis, Stefan Van Aelst, Ruben Zamar	Subset Selection Ensembles	Two key approaches for high-dimensional regression are sparse methods such as best subset selection and ensemble methods such as random forests. Sparse methods have the advantage that they yield interpretable models. However, they are often outperformed in terms of prediction accuracy by “blackbox” multi-model ensemble methods. We propose an algorithm to optimize an ensemble of penalized regression models by extending recent developments in optimization for sparse methods to multi-model regression ensembles. The algorithm learns sparse and diverse models in the ensemble simultaneously from the data. Each of these models provides an explanation for the relationship between a subset of predictors and the response variable. To initialize our algorithm forward stepwise regression is generalized to multi-model regression ensembles. The resulting ensembles achieve excellent prediction accuracy by exploiting the accuracy-diversity tradeoff of ensembles. The ensembles can outperform state-of-the-art competitors on both simulated and real data.
		Christophe Croux	Robust Automatic Forecasting with exponential smoothing	We provide a framework for robust exponential smoothing. For a class of exponential smoothing variants, we present a robust alternative. The class includes models with a damped trend and/or seasonal components. We provide robust forecasting equations, robust starting values, robust smoothing parameter estimation and a robust information criterion. The method is implemented in an R package, allowing for automatic forecasting. We compare the standard non-robust version with the robust alternative in a simulation study. Finally, the methodology is tested on data.

11:00	11:30	Coffee Break		
11:30	13:00	IPS Session: Methods and Applications of Stochastic Simulation in Data Science		
Chair and organizer: David Muñoz				Auditorium BNP
Alfredo Garbuno-Iñigo, David Fernando Muñoz	Calibrate, emulate, sample	Many parameter estimation problems arising in applications are best cast in the framework of Bayesian inversion. This allows not only for an estimate of the parameters, but also for the quantification of uncertainties in the estimates. The overarching approach is to first use ensemble Kalman sampling (EKS) to calibrate the unknown parameters to fit the data; second, to use the output of the EKS to emulate the parameter-to-data map; third, to sample from an approximate Bayesian posterior distribution in which the parameter-to-data map is replaced by its emulator. This results in a principled approach to approximate Bayesian inference that requires only a small number of evaluations of the (possibly noisy approximation of the) parameter-to-data map.		
José Pablo Rodríguez, David Fernando Muñoz	Optimizing bus utilization and cost minimization in the Mexico City Metrobus: A simulation approach	The Mexico City Metrobus is a bus-rapid transit-system (BRTS) whose goal is to combine the capacity and speed of a subway system with the low costs and flexibility of a passenger-bus system. The system has 7 lines, each with various routes and, additionally, some routes operate across two different lines. An important problem to solve in a BRTS is the scheduling and dispatching of vehicles to serve passengers in a reasonable time, without incurring high costs. Scheduling in a stochastic environment is a difficult problem and we decided to solve this problem using a simulation model developed in Simio to analyze line number 1, which has four exclusive routes, and three routes that connect lines 1 and other routes. We used an optimization engine to find the optimal number of buses needed by route, type, and hour, while maintaining an average bus utilization of 75% to minimize operational costs. To model the input flow of passengers we used data from 2022 provided by Metrobus through the National Transparency Portal and historical data to model the probabilities of passengers entering each route using a Bayesian statistics to incorporate the uncertainty on the value of these probabilities.		
Luis A. Moncayo-Martínez, Elias H. Arias-Nava	Simulation-based Optimisation to Guarantee the Cycle Time in the SLAB-2 Problem in the Presence of Stochastic Environment.	In this work, a three-part algorithm is proposed to solve the Simple Assembly Line Balancing 2 (SALB-2) problem in which some manufacturing cells are known, and the objective is to minimise the cycle time. In the first part, a Mixed-integer Programming (MIP) model solves the problem of minimising the deterministic cycle time. In the second part, it is proven that the solution of the MIP model is not achievable when implemented in environments with stochastic parameters such as the number of workers, speed of the material handling system, and inter-arrival times. Therefore, modelling with SIMIO the problem, some scenarios are created to assess the effect of parameters' variability in the cycle time value. Finally, using OptQuest, a stochastic optimisation model is solved to minimise the cycle time given some value of the stochastic parameters. We prove our algorithm to solve an instance reported in the literature.		
David Fernando Muñoz	Empirical comparison of maximum likelihood and minimum Cramér-von Mises for input analysis in stochastic simulations	In a previous paper, Muñoz and Villafuerte (2015) introduced a software (Simple Analyzer) to fit sample data to the most frequently used probability distribution families as well as to generate sample data from each distribution to test how fitting procedures perform. This software was intended for input analysis of stochastic simulations and, in addition to the estimation of the parameters of the corresponding family of distributions, the joint estimation of a shift and a scale parameter is considered for each family of distributions. In most cases, the maximum likelihood (ML) method was applied to estimate the parameters of a family of distributions, except for the case a ML estimator may not exist. In this paper, we introduce a new version of the Simple Analyzer that incorporates graphs of the adjusted density functions and P-P plots, for a set of selected families (in a single figure) to facilitate visual comparison of fitting the sample data for different distribution families. We also incorporated parameter estimation based on the minimization of the Cramér-von Mises statistic to produce reasonable estimates with almost no requirements on the input data. Experimental results are discussed.		
11:30	13:00	IPS Session: Graph Navigation		
Chair and organizer: David Banks				Auditorium FSSH
Local Chair: Javier Trejos				
William Caballero, David Banks, Keru Wu	Defense and Security Planning under Resource Uncertainty and Multi-period Commitments	The public sector is characterized by hierarchical and interdependent organizations. For defense and security applications in particular, a higher authority is generally responsible for allocating resources among subordinate organizations. These subordinate organizations conduct long-term planning based on both uncertain resources and an uncertain operating environment. This article develops a modeling framework and multiple solution methodologies for subordinate organizations to use under such conditions. We extend the adversarial risk analysis approach to a stochastic game via a decomposition into a Markov decision process. This allows the subordinate organization to encode its beliefs in a Bayesian manner such that long-term policies can be built to maximize its expected utility. The modeling framework we develop is illustrated in a realistic counter-terrorism use case, and the efficacy of our solutions are evaluated via comparisons to alternatively constructed policies.		

		Polat Charyyev, Li Zhou, Elvan Ceyhan	Dependence of the Traversal Length of a Disambiguating Agent on the Obstacle Pattern: From Clustering to Regularity	In the optimal obstacle placement with disambiguation (OPD) problem, we investigate how traversal length depends on the spatial pattern of the obstacles in the entire traversal window. An obstacle placing agent (OPA) wishes to insert obstacles of two types as true or false obstacles in an environment to maximize the traversal length of a navigating agent (NAVA). NAVA is equipped with a sensor that can only assign probabilities to each obstacle as being a true obstacle but does not know the actual status of the obstacle. When NAVA comes by the obstacle, it disambiguates the status of the obstacle with a cost added to the traversal length. We investigate how NAVA's traversal length changes when OPA equips the navigation window with obstacles (only false, only true, or mixed type obstacles) whose pattern is changing from uniformness to regularity and to clustering. Monte Carlo simulations indicate that on the average the traversal length tends to increase as the obstacle pattern changes from uniformness to regularity and decrease as the obstacle pattern changes from uniformness to clustering regardless of the type of obstacles.
		David Banks	Bayesian Graph Traversal	Many decision-theoretic problems can be modeled as graph navigation problems, with costs for edge traversal and payoffs at vertices. The costs and payoffs are generally unknown, but there may be Bayesian beliefs that are updated by experience to guide the search. The goal is to maximize the difference between payoff and costs. This talk explores several versions of the problem, especially one in which an opponent can increase the cost of an edge or lower the payoff at a node.
11:30	13:00	IPS Session: High-Dimensional Data Analysis		
		Chair and organizer: Rosaria Lombardo		Room S501 FC
		Local Chair: Arthur Tenenhaus		
		Se-Kang Kim	Segment Profile Analysis (SEPA): "Peeling Off" Person Profiles of Observed Scores across Two-Dimensional Biplots for Better Assessment	Segment profile analysis (SEPA) is introduced as a novel statistical technique based on singular value decomposition and the biplot paradigm. SEPA estimates dimensions from a profile-type dataset in which rows represent individual profiles of column variable scores. SEPA, unlike other profile analyses, does not interpret dimensions as core profiles but rather uses them to construct two-dimensional biplots. Each biplot is constructed with two dimensions that are used only once, rendering them independent. In each biplot, SEPA then analyzes (1) the relationships between person points and domain lines, (2) the projections of person points onto domain lines in each plane, and (3) the contribution of individual domains. Each biplot defines the projections of the person points exclusively. The projected person profiles are "peeled off" the actual person profiles by the biplot. These "peeled off" or segmented person profiles provide information about profile patterns that is unavailable in the actual person profiles. To demonstrate its utility, SEPA was applied to the seven domain scores of the Woodcock-Johnson IV Cognitive Abilities, three independent biplots were identified, and the results were used for both individual- and domain-level assessments. SEPA is the method for simultaneously assessing individuals and domains across multiple biplots.
		Gianmarco Borrata, Mario Musella, Ida Camminatiello, Rosaria Lombardo	Evaluating regression models' effectiveness in high-correlation scenarios	Ordinary least squares method is one of the most widely used regression tools for modelling the dependence relationship between dependent and independent variables. However, the presence of a strong correlation among the independent variables, known as multicollinearity, can have negative effects on the model estimation. It leads to wider standard errors of the regression coefficients, resulting in less precise and stable estimates of the coefficients. Different models have been proposed in literature to address this problem. In this work, the authors intend to evaluate the effectiveness of four regression models which deal with multicollinearity, i.e. the ridge regression, the least absolute shrinkage regression, the linear and non-linear partial least squares regression and the elastic net. These models aim to reduce the variance of the regression estimators while accounting for the presence of multicollinearity. Both simulated and real data will be considered for evaluating the regression models' efficiency in various scenarios (such as different correlation structures among the predictors, different numbers of observations and explicative variables).
		Amalia Vanacore, Armando Ciardiello	The behavior of classifier performance measures when dealing with class imbalance and instance hardness	In this study the behavior of several classifier predictive measures is investigated under different conditions of class imbalance and classification hardness. The investigation has been conducted through an extensive comparative analysis where several classification algorithms (i.e. 8 algorithm-level methods and 4 hybrid methods) have been applied to artificial data sets generated for multi-classification problems in multi-dimensional space and covering a wide range of class imbalance and instance hardness levels. Specifically, the data generation process has been controlled through a set of properties providing the characteristics of the generated data (i.e., number of attributes, $p=3, 5, 7$; number of classes, $k=2, 3, 5$; class frequency distributions, representing 6 increasing levels of Imbalance Ratio; instance type frequency distributions, representing 4 increasing levels of instance hardness). Study results highlight that, although the investigated performance measures quite agree for easy classification tasks (i.e. with balanced datasets containing only easy-to-classify instances), their behavior significantly differs when dealing with difficult classification tasks (i.e. increasing class imbalance and instance hardness) which is a rule in many real-word classification problems.

13:00	13:15	Conference Official Photo	
13:00	14:00	Lunch	
14:00	15:00	Keynote Speaker 2: Katherine Ensor Auditorium BNP	
		Chair: David Muñoz	
		Katherine Ensor	Statistics: A foundation for innovation Statistical foundations are without question at the core of modern innovation. In today's economy, a common phrase is "data is the new gold". Certainly, we live in an age where data is large, ubiquitous, and comes in many forms. The contributions from the statistical sciences go beyond "data". We are emerging from a pandemic where statisticians around the globe saved lives by contributing critical understanding to vaccines, treatments, pandemic policies, and management. The contributions are universal - from self-driving cars to Mars rovers, to sustainable and improved infrastructure, to clean energy and environmental stewardship, to financial markets and investing, to advances in medicine and medical practices, and even toward a better understanding of the communities in which we live, work, learn and play. This talk will highlight these important contributions and the innovations they made possible and will look to innovations on the horizon.
15:00	16:00	IPS Session: Multiblock Methods and Supervised Learning Algorithms for Data Science Auditorium BNP	
		Chair and organizer: Alba Martínez-Ruiz	
		Alba Martínez Ruiz , Natale Carlo Lauro	Incremental singular value decomposition for some numerical aspects of multiblock redundancy analysis Simultaneously processing several large blocks of streaming data is a computationally expensive problem. Based on the incremental singular value decomposition algorithm, we propose a new procedure for calculating the factorization of the multiblock redundancy matrix M, which makes the multiblock method more fast and efficient when analyzing large streaming data and high-dimensional dense matrices. The procedure transforms a big data problem into a small one by processing small high-dimensional matrices where variables are in rows. Numerical experiments illustrate the accuracy and performance of the incremental solution for analyzing streaming multiblock redundancy data.
		Agostino Gnasso , Massimo Aria and Luca D'Aniello	Exploring the Applications of Explainable Ensemble Trees: Real-World Scenarios Ensemble methodologies, such as random forest, employ supervised learning algorithms to ensure precise solutions by training multiple models. Random forest constructs decision trees on diverse samples, using majority voting for classification and averaging for regression. Its main advantage is superior predictive accuracy, yet interpretability remains a drawback in several research domains such as finance and medicine. To tackle the issue of interpretability, we proposed a novel approach called Explainable Ensemble Trees (E2Trees), which aims to elucidate and visually represent the relationships and interactions among the variables employed in the random forest algorithm. Our proposal amalgamates the advantages of decision trees and random forest models. In essence, we offer a dendrogram-like structure capable of explaining all the information encapsulated within the output of a random forest algorithm. We provide concrete real-world examples to illustrate how practitioners can effectively employ the E2Tree framework for evaluating and understanding interpretations.
		Arthur Tenenhaus , Michel Tenenhaus	RGCCA for Structural Equation Modeling with Latent and Emergent Variables In this work, we show how to use Regularized Generalized Canonical Correlation Analysis (RGCCA) in structural equation modeling with latent and/or emergent variables. This new approach produces consistent and asymptotically normal estimators of the parameters. RGCCA relies on a well-grounded optimization problem and the global convergence of the algorithm used to solve this problem is guaranteed. We also propose a maximum likelihood (ML) estimation method for estimating the parameters of the model. RGCCA and ML are evaluated in a Monte Carlo simulation and lead to similar results. RGCCA and ML are also compared on the ECSI data for the mobile phone industry and produce very close results.
15:00	16:00	IPS Session: IPS Recent Developments on Computational Statistics for the Modelling of Multivariate Auditorium FSSH	
		Chair and organizer: Mauricio Castro	
		Pedro Luiz Ramos , Eduardo Ramos, Francisco Rodrigues, Francisco Louzada	A generalized closed-form maximum likelihood estimator The maximum likelihood estimator plays a fundamental role in statistics. However, for many models, the estimators do not have closed-form expressions. This limitation can be significant in situations where estimates and predictions need to be computed in real-time, such as in applications based on embedded technology, in which numerical methods can not be implemented. This paper provides a generalization in the maximum likelihood estimator that allows us to obtain the estimators in closed-form expressions under some conditions. Under mild conditions, the estimator is invariant under one-to-one transformations, strongly consistent, and has an asymptotic normal distribution. The proposed generalized version of the maximum likelihood estimator is illustrated on the Gamma, Nakagami, and Beta distributions and compared with the standard maximum likelihood estimator.

	Luis Gutierrez	Bayesian flexible models for N-way analysis of variance	Analysis of variance (ANOVA) literature is antique and based on restrictive assumptions. Thus, existing models are inappropriate for analyzing complex and diverse data correctly. To circumvent these limitations, this work focuses on developing flexible models for N-way ANOVA designs under a Bayesian nonparametric inference approach. Specifically, our strategy considers the definition of latent binary variables, which identify each cell with different random distributions. Such distributions are modeled via infinite mixture models, where the mixing distributions follow a dependent Dirichlet process with shared weights. In our specification, the binary latent variables map the hypotheses. Then, the prior distribution on the hypothesis space is defined with a distribution over the binary vector. In summary, we relax the classical ANOVA assumptions, propose models for data in different supports, and study the Bayes factor consistency for these models. Our methodology is implemented in an R-package called ANOVABNPTesR, which provides easy-to-use functions for continuous, counting, and binary responses.
	Mauricio Castro , L. Hernández-Velasco, C. Abanto-Valle and D. Dey	A Bayesian Approach for Mixed Effects State-Space Models Under Skewness and Heavy Tails	Human immunodeficiency virus (HIV) dynamics have been the focus of epidemiological and biostatistical research during the past decades to understand the progression of acquired immunodeficiency syndrome (AIDS) in the population. Although there are several approaches for modeling HIV dynamics, one of the most popular is based on Gaussian mixed-effects models because of its simplicity from the implementation and interpretation viewpoints. However, in some situations, Gaussian mixed-effects models cannot: (a) capture serial correlation existing in longitudinal data, (b) deal with missing observations properly, and (c) accommodate skewness and heavy tails frequently presented in patients' profiles. For those cases, mixed-effects state-space models (MESSM) become a powerful tool for modeling correlated observations, including HIV dynamics, because of their flexibility in modeling the unobserved states and the observations in a simple way. Consequently, our proposal considers a MESSM where the observations' error distribution is a skew- t . This new approach is more flexible and can accommodate data sets exhibiting skewness and heavy tails. Under the Bayesian paradigm, an efficient Markov Chain Monte Carlo algorithm is implemented. To evaluate the properties of the proposed models, we carried out some exciting simulation studies, including missing data in the generated data sets. Finally, we illustrate our approach with an application in the ACTG-315 clinical trial data set.
16:00	16:30	Coffee Break	
16:30	18:00	IPS Session: Applied Modeling Using Data Science Tools	
		Chair and organizer: Julio López	Auditorium BNP
	Sara Arancibia , Sebastian Gomez	Factors that influence the adoption of technologies in MSMEs	This research study aims to identify the factors influencing the intention of micro, small, and medium enterprises (MSMEs) to digitally transform. By applying the Theory of Planned Behavior (TPB) and the Technology Acceptance Model (TAM), a model was developed to analyze these factors using structural equation modeling (PLS-SEM). A questionnaire was administered to a sample of 1,199 respondents, primarily owners and/or leaders of MSMEs out of a total pool of 140,000 companies. Among the main findings of the research, it is obtained that the adoption of technologies and the digital transformation process must be addressed according to the stage of maturity of the company, its size and the economic sector in which it operates. Effective communication of the benefits of digital transformation, supported by specific and relevant examples that align with the realities and needs of the companies, is crucial. Furthermore, it is important to emphasize that the perceived benefits outweigh the costs associated with adopting digital transformation. This research contributes to the existing literature on digital transformation and provides valuable insights for practitioners and decision-makers in fostering digitalization among MSMEs.
	Louis de Grange , Matthieu Marechal , Rodrigo Troncoso	Discrete Choice Model Estimation using Instrumental Variables	In this paper, we present two new approaches that allow for the acquisition of estimators with consistency properties for the parameters of Multinomial Logit Models that include endogenous explanatory variables. Both approaches are based on the use of instrumental variables. The first approach corresponds to moment conditions that include instrumental variables. The second approach considers combining parameters obtained during two different stages of estimation, as well as the use of instrumental variables. We implement both new approaches using simulated data, and compare them with the classic Control Function method. As a result, we obtain similar estimators among the three methods for the parameters of explanatory variables; however, when estimating the model's constant terms (modal constants), our new approaches provided much more accurate estimates. The latter has consequences on the predictive capacity of the models, as well as on the estimation of marginal effects, elasticities, and social benefits (consumer surplus).
	Hugo Robotham	Causal relationships between nonlinear time series: An application using Convergent Cross Method.	The natural and ecological ecosystems in which many fishery systems are located are nonlinear and complex; therefore, appropriate data manipulation techniques that are better adapted to these systems must be considered. The identification of causality relationships is important to determine effective fishery policy and management recommendations. The convergent cross method (CCM) was used to infer causal relationships from time series of the clam fisheries. The causal structure of a set of indicators of the pillars (biological, environmental, social and economic) of the sustainable development of the clam fishery is described. The resulting causal structure reveals unidirectional and bidirectional causal relationships of the indicators. The causal relationships of the network can be used, to improve predictive models, optimise the control and follow-up of the resources monitoring, among others.

		Miguel Carrasco, Julio Lopez, Sebastián Maldonado	Robust Support Vector Machine Using the Cobb-Douglas function	We propose a novel machine learning approach based on robust optimization. Our proposal defines the task of maximizing the two class accuracies of a binary classification problem as a Cobb-Douglas function. This function is well known in production economics and is used to model the relationship between two or more inputs as well as the quantity produced by those inputs. A robust optimization problem is defined to construct the decision function. The goal of the model is to classify each training pattern correctly, up to a given class accuracy, even for the worst possible data distribution. We demonstrate the theoretical advantages of the Cobb-Douglas function in terms of the properties of the resulting second-order cone programming problem. Important extensions are proposed and discussed, including the use of kernel functions and regularization. Experiments performed on several classification datasets confirm these advantages, leading to the best average performance in comparison to various alternative classifiers.
		Julio Lopez, Miguel Carrasco, Sebastián Maldonado	Some robust formulations for binary classification problem	In this work, we present novel robust formulations for binary classification based on the Minimax Probability Machine (MPM) approach. The idea is to introduce a regularization term the MPM and Minimum Error Minimax Probability Machine approaches. This inclusion reduces the risk of obtaining ill-posed estimators, stabilizing the problem, and, therefore, improving the generalization performance. In addition, inspired by the twin support vector machine method, we study a twin version in this context. Finally, some experiments on well-known binary classification datasets demonstrate the virtues of these novel formulations in terms of predictive performance.
16:30	18:00	IPS Session: Public Health Data Science		
		Chair and organizer: Sandra Flores Alvarado		Auditorium FSSH
		René Lagos, Brian Fleiderman, Juan Cristóbal Morales	Predicting specialty consultations in public hospitals, a public health data science application	Public hospitals face a high patient load and should align medical programming with the needs of the population, but there is no established method for doing so. What methods can be used to predict the demand for interconsultations in the public health network? First, a descriptive time series analysis was conducted on interconsultations demand in a metropolitan Health Service. A subset of specialties with trend and stationary patterns was visually identified. Predictions were made using the methods of moving average, double exponential smoothing, and XGBoost. Mean absolute percentage error (MAPE) was calculated to compare the performance of the three methods. Five pediatric specialties with a trend of increase in 2022 were selected and compared to five adult specialties with a stationary pattern in the same period. Moving average was the best method for predicting demand for stationary specialties (21.73% MAPE) and double exponential smoothing performed better in specialties with emerging trends (19.64% MAPE). This application shows how data science for public health provides a methodological framework for monitoring the outpatient care needs of the population in an ethical manner, focused on the interpretability and reproducibility of the analyses.
		René Lagos Barrios, Sandra Flores Alvarado, Andrés González-Santa Cruz, Felipe Medina Marín	Conceptual and Ethical Challenges of Public Health Data Science	Digitalization is generating a profound transformation in society, and the pandemic has accelerated and deepened digitization in many areas, including public health. Numerous initiatives emerged and deployed their potential during this period, such as ICOVID Chile, Covid Analytics, or the Digital Hospital. Where is digitization leading the practice and research of public health? And what role is the public health community playing in this new field of development? The 80th anniversary of the School of Public Health and the 40th anniversary of the Master's in Biostatistics at the University of Chile provide an opportunity to discuss these changes, their potential, and their course in Chile. In this article, we highlight some concepts to outline the scope of the discussion such as public health data science and digital transformation in public health. We also discuss the main ethical challenges presented by the advancement of data science in the field of public health. The main ethical dilemmas being the potential threats to privacy and autonomy resulting from the use of digital models in public health, which may be influenced by economic and social control interests.
		Sandra Flores Alvarado, Cristóbal Cuadrado, Natalia Vergara, María Fernanda Olivares, Christian García	Assessment of Sentinel Surveillance Capacity for Early Detection of Respiratory Disease Outbreaks: Lessons from the COVID-19 Pandemic in Chile	Over the past century, we have witnessed recurrent outbreaks of emerging and reemerging respiratory diseases. Epidemiological surveillance for the early detection of outbreaks is crucial for an effective public health response. Sentinel surveillance relies on the selection of subgroups to identify outbreaks and changes in case trends. However, it faces several challenges such as case definitions, inadequate sample sizes, completeness, and timely delivery, leading to underestimation of disease incidence or untimely information. In Chile, during the COVID-19 pandemic, a universal surveillance system with mandatory reporting was implemented (census), collecting a large volume of data while maintaining sentinel surveillance for influenza and other respiratory virus. This study examines the capacity of respiratory disease sentinel centers in Chile to estimate the incidence of COVID-19 cases in an accurate and prompt manner. The COVID-19 census data is compared with sentinel surveillance using distance measures for time series, and the COVID-19 incidence rate is modeled using sentinel surveillance records and other covariates. The results provide insights into the effectiveness of sentinel surveillance in detecting epidemic outbreaks. These findings will contribute to strengthening surveillance systems, improving preparedness and capacity to respond to emerging respiratory diseases, providing key evidence for decision-making during public health crises.

		Daniela Varas, Andrés Iturriaga, Sandra Flores and Felipe Medina	Cluster Analysis to Characterize the Profile of the Aggressor Through of the ENVIF-VCM 2020	Gender-based violence, according to the United Nations, is defined as "harmful acts directed against a person or group of people because of their gender." It is a global phenomenon that is most of the time perpetrated by an intimate partner. The purpose of this study is to identify profiles that allow to characterize this partner, through the answers to questions from women about their partner or ex-partner that were provided in the 2020 National Survey on Violence against Women in the Context of Domestic Violence and in other spaces (ENVIF-VCM). For this, the agglomerative hierarchical clustering, k-medoids (PAM) and the mixture model for mixed type variables are used. Through these cluster methods, it is sought to detect patterns of behaviors that help to know how these perpetrators are and to validate the quality of the profiles obtained. The results of this work allow to break down and reaffirm certain stereotypes that are held about the aggressors and can also contribute to the prevention of gender violence.
--	--	---	---	--

18:00	19:00	Keynote Speaker 3: Miguel de Carvalho		Auditorium BNP
		Chair: Mauricio Castro		

		Miguel de Carvalho	Data Science of Extreme Events	<p>Extreme events, such as hurricanes of unprecedented strength, heatwaves surpassing historical temperatures, and floods inundating regions previously deemed safe, have become alarmingly frequent in recent years. In this talk, I will highlight how Statistics and Data Science contribute to assessing the risk of extreme events, gauging their likelihood, and mitigating their impact on modern society. I will introduce methods from Extreme Value Theory and illustrate their applications using real data analyses. The methods that we will explore are being shaped by a community dedicated to extrapolating beyond observed data—into the tails of a distribution—drawing insights about the risk of extreme events, and understanding the dynamics governing extreme values across time and space. This talk is based on:</p> <p>- Coles, S., de Carvalho & Davison, A. C. (in preparation) An Introduction to Statistical Modeling of Extreme Values. Second edition. Springer: New York.</p>
--	--	---------------------------	--------------------------------	--

19:30	20:30	Welcome Reception		
-------	-------	--------------------------	--	--

THURSDAY 9 NOVEMBER 2023

08:30	09:30	Keynote Speaker 4: Karol Suchan		Auditorium BNP
		Chair: Julio López		

		Karol Suchan	Workload distribution in the yard of a multipurpose port terminal in Chile	<p>Our study delves into analyzing container flows within a maritime terminal's yard, focusing on workload distribution and vehicle movements. Our key objective is to create a yard simulator using statistical modeling, enabling us to evaluate diverse vehicle dispatching rules governing gate, storage block, and terminal traffic. Uniform workload distribution is proven to enhance productivity by minimizing unproductive container moves, shortening service queues, and easing personnel and machinery burden. Our investigation extensively analyzed workload distribution at a Chilean maritime terminal, using a dataset spanning 2016 to 2022. We revealed instances of non-uniform distribution, indicating potential for optimization. The terminal's layout includes storage sections and a break-bulk cargo area, each with unique gates for truck and rail cargo. Our approach heavily relies on statistical modeling to derive insights from data. Future work entails leveraging the simulator to devise strategies for optimizing distribution and ultimately bolstering terminal efficiency.</p>
--	--	---------------------	--	--

09:30	11:00	IASC-ARS IPS Session: Recent Developments for Effective Computation and Learning		Auditorium BNP
		Chair and organizer: Yuichi Mori		
		Chair: Luis Firinguetti		

		Yipeng Zhuang, Philip Yu	Graph-based preference learning with multiple network views	<p>In today's social media era, people interact and communicate across various networks. Their preferences (such as ratings and rankings) for items like movies and products are often incomplete. We propose a novel graph neural network model for preference learning that incorporates multiple network views and side information to improve the accuracy of predicting missing preferences. Our proposed model employs an attention mechanism to measure the influences from each multi-view network. We also introduce a new objective function to evaluate preference prediction performance. To test our model, we apply it to a few real-world movie recommendation datasets. The empirical results show that our proposed model outperforms existing models regarding preference prediction.</p>
--	--	---------------------------------	---	---

		Hung Hung, Su-Yun Huang , Shinto Eguchi	Robust self-tuning semiparametric PCA for contaminated elliptical distribution	The usual PCA is known to be sensitive to the presence of outliers, and thus many robust PCA methods have been developed. Among them, the Tyler's M-estimator is shown to be the most robust scatter estimator under the elliptical distribution. However, when the underlying distribution is contaminated and deviates from ellipticity, Tyler's M-estimator might not work well. In this talk, we apply the semiparametric theory to propose a robust semiparametric PCA, which is shown to be a re-weighted Tyler's M-estimator. The merits of our proposal are twofold. First, it is robust to both heavy-tailed elliptical outliers and non-elliptical outliers. Second, it pairs well with a data-driven tuning procedure, which is based on active ratio and can adapt to different degrees of data outlyingness. Simulation studies and image data examples will be presented.
		Chanmin Kim	Bayesian Ensemble Trees for Principal Stratification	Principal stratification analysis is a technique employed for assessing causal effects by examining the impact of treatment on an outcome through the lens of the relationship between treatment and an intermediate variable after treatment. When dealing with a continuous intermediate variable, it is generally necessary to model it. However, existing parametric modeling approaches struggle to fully capture the intricate relationship between variables. Moreover, when outcome and intermediate models are estimated separately, it becomes challenging to consider the uncertainty introduced by each model estimation when estimating the final causal effect. To address these issues, we propose a fully Bayesian method utilizing the Bayesian additive regression trees model. Unlike other methods found in the literature, our approach flexibly estimates all intermediate, outcome, and propensity score models using Bayesian nonparametric models. Additionally, our method can be applied to both targeted selection (specific confounding situations) and broad confounding scenarios. To demonstrate the performance of our proposed method, we conduct a series of simulation studies. Applying our method, we investigate the effects of the installation of sulfate abatement devices (scrubbers) in US coal-fired power plants on the surrounding concentrations of PM2.5 from various perspectives, considering the relationship between the scrubber and SO2 emissions.
		Masahiro Kuroda , Yuichi Mori	Speeding up the computation of a non-negative matrix factorization algorithm	Non-negative matrix factorization (NMF) is widely used in analyzing non-negative data matrices with high-dimension, such as image processing, text mining and spectral data matrices. NMF approximates a given non-negative matrix by the product of two non-negative matrices. Thus, the NMF problem finds these non-negative matrices as solutions. Several iterative algorithms are proposed to solve the problem. The multiplicative update (MU) algorithm is used as an efficient computational algorithm for solving the NMF problem. The algorithm is described by a simple computational procedure and has stable convergence, while its speed of convergence tends to be slow. To accelerate the convergence of the MU algorithm, we have developed an acceleration algorithm that utilizes the vector epsilon algorithm. The vector epsilon algorithm is applicable to other algorithms like the EM (expectation-maximization) algorithm and alternating least squares algorithm, and it can improve their speed of convergence. We provide the procedure for applying the vector epsilon acceleration to the MU algorithm and evaluate the performance of this acceleration algorithm through numerical experiments.
09:30	11:00	IPS Session: New Developments in Symbolic Data Analysis		
		Chair and organizer: Paula Brito		Auditorium FE
		Local Chair: Trevor Harris		
		Francisco de Assis Tenorio de Carvalho, Eufrazio de Andrade Lima Neto , Ulysses Rosendo	Interval Joint Robust Regression Method	Interval-valued data are needed to manage the uncertainty related to measurements or the variability inherent to the description of complex objects representing a group of individuals. A number of regression methods suitable to interval variables describing the variability of complex objects are already available. However, less attention has been given to methods that, simultaneously, take into account the full interval information and are robust to interval outlier observations, even with the frequent presence of atypical observations on interval-valued data sets. We propose a new robust linear regression method for interval variables, where the presence of outliers either in the centre or in the radius penalises both the centre and the radius regression models. Moreover, interval observations with outliers on both centre and radius are more penalised than those observations with outliers only in the centre (or in the radius). The parameter estimation algorithm estimates the parameters of the centre (or of the radius) model taking into account both information about the centre and the radius. The convergence and time complexity of the iterative algorithm is also presented. The new method's performance is compared with some previous robust regression approaches and evaluated on synthetic and real interval-valued data sets.

	Antonio Irpino	New visualization tools for numeric distributional data tables	Visualization tools are essential for showing patterns and suggesting analysis strategies to the user. Distributional data tables, where each observation is described by a vector of frequency or density distributions, are complicated to be visualized in a comfortable manner, and new visualization tools are then required. To answer the lack mentioned above, we present two new visualization tools for data tables described by numeric distributional data. The first tool, called Green Eye Iris (GEI) plot, is based on a polar coordinate-based representation of stacked bar charts. The tool allows users to compare two or more observations described by a moderate set (less than 12) of distributional variables. Colors play an important role, and a red-to-green diverging palette is used. A second tool is an extension of the classical heatmap plot, which is suitable for describing a dataset with many observations and more than 12-15 distributional variables. Both methods are based on visualizing the proportion of mass distributed on the domain variable using diverging color palettes for each distribution. Applications on some well-known distributional data will be presented to show the advantages of using the proposed tools.
	A. Pedro Duarte Silva, Peter Filzmoser, Paula Brito	Detecting Outliers in Distributional Data with Sparse Robust Estimators	We consider numerical distributional data, i.e., data where units are described by histogram or interval-valued variables, representing intrinsic variability of the corresponding observations. Parametric probabilistic models are introduced, which are based on the representation of each distribution by a location measure and interquantile ranges. Multivariate Normal distributions are assumed for the whole set of indicators, considering alternative structures of the variance-covariance matrix. A multivariate outlier detection method is proposed that is based on a sparse robust estimator of its inverse. The proposed methodology is evaluated on simulated data and then illustrated with real distributional-valued data.
	Oldemar Rodriguez Rojas	Symbolic t-SNE and UMAP methods for interval type variables	UMAP (Uniform Manifold Approximation and Projection) is a very new method for dimension reduction. UMAP method improve t-SNE (t-Distributed Stochastic Neighbor Embedding) method for data visualization and dimensionality reduction. The great advantage of UMAP is that it preserves better than t-SNE the global structure with superior run time performance. The foregoing makes UMAP an ideal method to be applied to the hyper-rectangles that are in the rows of the symbolic data table with interval-type variables, since UMAP compresses the structure inside each hyper-rectangle very well and at the same time better preserves the global structure of the clusters generated by each hyper-rectangle. This paper presents an adapted version of the t-SNE and UMAP methods for interval type variables. In addition, R and Python codes for both generalizations are presented.
11:00	11:30	Coffee Break	
11:00	11:30	Poster Session	
11:30	13:00	CA Session: Statistical and Machine Learning Methods	
		Chair: Natalia da Silva	Auditorium BNP
	Ignacio Alvarez-Castro	Predictive models using Learning Management System data In primary schools	Plan Ceibal is a public policy implemented in Uruguay, it is part of the global initiative One Laptop per Child (OLPC, 2005). The basic feature is providing every student and teacher in primary school with a laptop or tablet and internet access. Different data sets were combined, students and teachers activities registered in the Learning Management System (LMS) and student's performance in national standardized tests. Data were used to compute student's engagement indexes, combining motivation, creativity, velocity and performance. Statistical models were used to determine key drivers of LMS use, this is relevant to define educational policies based on evidence. Models for LMS use are fitted for several regional levels. Additionally, statistical learning methods were fitted to predict student's performance in national standardized test using as predictor variables different constructed usage indexes from the LMS platform. A major challenge was how to deal with sub-grouping data structure into machine learning algorithms, usually developed for independent observations. Initial results suggest school district is the main driver of the technology usage in the classroom.
	Adolfo Fuentes , Cristian Candia	Neural Networks Nominate for Predicting Political Coordinates	In this study, we used paired comparisons to study political participation in Chile. The public data was collected on a platform that asked participants to choose between two policy proposals. We used over 7 million pairwise preferences contributed by over 100,000 anonymous individuals. On one hand, we used a matrix of votes in support of a proposal to construct a similarity matrix among participants. We utilized this matrix to predict the preferences of participants by multiplying it with the matrix of users' preferences. We found that the model was able to predict participants' preferences with an accuracy of 72%. However, the accuracy of the model increased to 87% or higher when it was exclusively used to predict the preferences of participants for pairs of proposals that were highly divisive. On the other hand, a Neural Network model has been developed to obtain the political coordinates corresponding to each user and proposal in a two-dimensional space. Each coordinate represents the same concept used in the Nominate algorithm. We believe that these findings have significant implications for the study of political participation and the development of tools to increase citizen engagement.

		Martha Bohorquez Castañeda	Modeling multivariate spatial variability of soil deposits using functional random fields	A five step methodological framework to describe and model the spatial variability of a soil deposit, using spatial functional random fields, is presented and discussed. The method uses the functional description of the soil profile to account for the complex spatial anisotropy that arises in layered materials. By capturing the depth related variability within in the functions themselves, the method can be used to analyze large volumes of information and spatially correlated variables. The use of functional random fields overcomes many of the difficulties associated with scalar random fields and reduces the smoothing of features associated with geostatistical predictions. The final results obtained provide a measure of expected uncertainties and are compatible with further reliability-based calculations. A case study for is presented by using CPTu test measurements over a 1.62 km corridor in Bogotá city.
		Natalia da Silva	Projection pursuit supervised classification	Projection pursuit random forest (PPF) is an ensemble learning method for classification problems, built from trees utilizing combinations of predictors. PPF builds a forest from many projection pursuit trees (PPtree); trees are constructed by splitting on linear combinations of randomly chosen variables. Projection pursuit is used to find the linear combination of variables that best separates groups, and many different rules to make the actual split are provided. Utilizing linear combinations of variables to separate classes takes the correlation between variables into account, which allows PPF to outperform a traditional random forest when separations between groups occur in combinations of variables. PPF can be used in multi-class problems and is implemented into an R package PPforest. Some extensions of the individual trees in PPF are explored to make the classifier more flexible, to tackle more complex problems, while maintaining interpretability.
11:30	13:00	IPS Session: Modern Statistical Visualization		
		Chair and organizer: Juergen Symanzik		Auditorium FE
		Local Chair: Jorge Luis Bazán		
		Susan Vanderplas , Emily Robinson, Reka Howard	Multimodal User Testing: Producing comprehensive, task-focused guidelines for chart design	For at least the last 100 years, researchers have been testing statistical graphics and arguing about which chart designs are better. Many of these studies produce conflicting recommendations: should we use pie charts to display data about the relative proportions of a whole, or are stacked bar charts better? Much of the time, user testing statistical graphics takes a back seat to aesthetic preferences and gut feelings, but even when we test graphics, we often only use one methodology that is focused on a specific use case. For instance, visual inference is often used to determine whether someone can detect an effect, but it does not allow us to easily examine whether users can extrapolate from the data shown, or can draw logical conclusions from a chart. In this presentation, I'll discuss ongoing research examining chart design choices using multiple testing methods. Each of these methods has been designed to measure the usability of charts for a specific task: perception, estimation, and forecasting. We'll consider the benefits and drawbacks to this type of user testing and discuss the nuances of design decisions on chart usability.
		Xiaoyue Cheng , Bryan Vukorepa, Tamara Williams, Mahbul Majumder	Exploring Course-taking Pattern with an Interactive Visualization Dashboard	Course-taking data for K-12 students, particularly those in middle schools and high schools, has become an important support to education providers and policymakers, as it could offer valuable insights into student learning pathways and the obstacles they encountered on a large scale. This type of data can also be utilized to compare behaviors among various student groups, improve curriculum design, and predict course enrollment. However, visualizing course-taking data presents challenges due to three main components: students, courses, and time. This study aims to address these challenges by grouping students according to their yearly course selections, aligning course nodes chronologically and accommodating different term settings. Interactive visualization helps to reveal the three distinct learning pathways of math courses in high schools and illustrate variations among different groups of students in Nebraska, USA. A demonstration of the interactive R Shiny dashboard will be provided in the presentation to show the visualizations and findings.
		Ursula Laa	Tour visualizations for the interpretation of machine learning models	Modern statistical models are often black boxes that use high-dimensional input. Interpreting such models is a challenge, which has driven the development of new methods of explainable AI that can, for example, provide a local explanation for a single prediction. Complementary to such local approaches, we can get a global overview of a fitted model using tour visualizations. In this presentation, I will show a case study that illustrates how new developments in tour methods allow for better inspection of a fitted classification model. This includes new displays (the sage tour addressing the curse of dimensionality, and the slice tour that lets us look „inside“ a distribution) as well as new tour algorithms for better interaction with the visualization.

11:30	13:00	IPS Session: Exploring Multivariate Data: A Variety of Applications		Room S203 FP
		Chair and organizer: Sugnet Lubbe		
		Local Chair: Paulo Canas Rodrigues		
		Praise Obanya , Roelof Coetzer, Carel Olivier, Tanja Verster	Variable contribution analysis in multivariate process monitoring using permutation entropy In multivariate process monitoring, once a fault has been detected, fault diagnoses must be performed to identify the variables that are responsible for the fault or the deviation from normal operation. In this paper, we present a new criterion for variable contribution analysis to a fault based on permutation entropy. Specifically, permutation entropy (PE) is a statistical method for the measurement of complexity of a given time series. In this paper, we show how PE can be used to identify and interpret the variables responsible for or affected by faults in an industrial process, which is based on a change in the characteristics of the time series for the affected variables. The well-known Tennessee Eastman Process (TEP) is used to illustrate the application of PE for variable contribution analysis. Two sets of simulated TEP chemical process data, namely fault-free and faulty processes, are utilized to show that PE identifies the correct variables that contributed to specific faults. In addition, comparisons between the dynamics of the fault-free and faulty processes aid in the identification of the variables with the highest contribution to the specific faults. The results show that PE is an efficient analysis tool for specifying variable contributions to faults.	
		Ruan Buys , Carel van der Merwe	Visualising PCA Biplots With Density Axes: an R Package Principal Component Analysis (PCA) biplots are useful tools in visualising and analysing higher-dimensional data by projecting it to a lower-dimensional subspace. This is done by projecting the data, accompanied by calibrated axes, onto a surface spanned by a set of Principal Components. The calibrated axes often clutter the plotting space and may obscure interpretation of data projections to the axes. This paper proposes a package compiled in the R landscape to visualise the said biplot, embedded with an automated algorithm to translate the axes out of the plotting space: decluttering the view. Superimposed on the axes are density graphs per class group in the data. The package renders the biplot in HTML through the Plotly package, which converts arguments to JSON bindings utilised in the Plotly.js library. The paper further details how prediction lines can be inserted onto the biplot through the manipulation of the Document Object Model (DOM) with JavaScript code. The final leg of the paper considers some code optimisation techniques to efficiently draw the final view.	
		Sugnet Lubbe	Visualising multi-dimensional data with mixed scaled measurements Traditionally Principal Component Analysis (PCA) biplots are associated with continuous scale measurements and Multiple Correspondence Analysis (MCA) biplots with categorical scale measurements. Here we will venture into the grey area of mixed scaled measurements. Categorical PCA (catPCA) find optimal numerical scaled values for categorical variables, before applying PCA to the numeric version of the data set. On the other hand, fuzzy coded biplots are constructed by applying a fuzzy coding of ordinal variables based on partitioning of the continuous scale into consecutive intervals to numeric data. Biplots are then constructed with MCA on the fuzzy coded indicator matrix. While these methods aim to transform all variables to the same scale before applying a method designed for either continuous or categorical data, generalised biplots perform multidimensional scaling on dissimilarities computed from the mixed scaled measurements. We will compare the different methods and highlight the advantages and disadvantages of these.	
13:00	14:00	Lunch		
13:00	14:00	Poster Session		
14:00	15:00	Keynote Speaker 5: Diego Gallardo		Auditorium BNP
		Chair: Sandra Flores Alvarado		
		Diego Gallardo , Marcelo Bourguignon, Helton Saulo	A parametric quantile beta regression for modeling case fatality rates of COVID-19 Motivated by the case fatality rate (CFR) of COVID-19, in this paper, we develop a fully parametric quantile regression model based on the generalized three-parameter beta (GB3) distribution. Generally, beta regression models are primarily used to deal with data arising from rates and proportions. However, these models are usually specified in terms of a conditional mean. Therefore, they may be inadequate if the observed response variable follows an asymmetrical distribution, such as CFR data. In addition, beta regression models do not take into account the effect of the covariates across the spectrum of the dependent variable, which is possible through conditional quantile approach. In order to introduce the proposed GB3 regression model, we introduce a reparameterization of this distribution by inserting a quantile parameter, and direct inference in parametric mode regression based on the likelihood paradigm. Furthermore, we proposed a simple interpretation of the predictor-response relationships in terms of percentage increases/decreases of the quantile. A Monte Carlo study is carried out for evaluating the performance of the maximum likelihood estimates and the choice of the link functions. A real COVID-19 data set is finally analyzed to illustrate the proposed approach.	

15:00	16:00	CA Session: Statistical Modeling and Data Science		
		Chair and organizer Luis Firinguetti		Auditorium BNP
	Luis Gomez , Sebastian Parra	Genetic algorithms for optimizing hyperparameters in neural network design	In the 1950s, researcher Frank Rosenblatt created the first artificial processing unit, the perceptron, inspired by Warren McCulloch and Walter Pitts. Since then, neural networks have been recognized as an innovative and efficient solution to everyday problems. In addition, genetic algorithms based on Darwin's evolution are combinatorial optimization techniques that aim to reduce the computation time of classic numerical resources by selecting the fittest individuals among generations to optimize a specific function. By applying these evolutionary algorithms to optimize neural network design parameters, such as the number of nodes, number of layers, transfer functions, and more, highly capable architectures can be built for specific problems. In this work, we propose using genetic algorithms to design neural networks capable of solving the electrical resistivity inversion problem. The electrical resistivity inversion problem involves using electromagnetic methods to map the subsurface structure by identifying mineral or rock bodies based on the resistivity readings, which vary according to the material's electromagnetic properties. This problem is typically solved using techniques from geophysical modeling, and various software tools, such as pyGIMLi and BERT, have been developed for this purpose.	
	Tarik Faouzi , Luis Firinguetti, José Avilez, Rubén Carvajal	The α -Groups under Condorcet Clustering	We introduce a new approach to clustering categorical data: Condorcet clustering with a fixed number of groups, denoted α -Condorcet. As k-modes, this approach is essentially based on similarity and dissimilarity measures. The paper is divided into three parts: first, we propose a new Condorcet criterion, with a fixed number of groups (to select cases into clusters). In the second part, we propose a heuristic algorithm to carry out the task. In the third part, we compare α -Condorcet clustering with k-modes clustering. The comparison is made with a quality's index, accuracy of a measurement, and a within-cluster sum-of-squares index. Our findings are illustrated using real datasets: the feline dataset and the US Census 1990 dataset.	
	Luis Firinguetti , Manuel Pereira	Best Linear Unbiased Estimation in a System with Singular Seemingly Unrelated Regression Equations	We consider a system of Seemingly Unrelated Regression Equations where there are some equations containing matrices of explanatory which are not of full rank. We show that the simultaneous estimation of the complete system will produce unbiased and fully efficient estimates of the subsystem of equations with full rank explanatory matrices. We also provide some simulations and present an example	
15:00	16:00	CA Session: Bayesian Analysis and Time Series		
		Chair: Ricardo Ehlers		Auditorium FE
	Ricardo Ehlers , Ritha Condori	Stochastic Volatility Models using Hamiltonian Monte Carlo Methods and Stan	This work aims at evaluate and compare the performance of the No-U-Turn Sampler (NUTS) algorithm, implemented in the Stan software, in estimating the parameters of stochastic volatility models with leverage based on scale mixtures of the skew-normal distribution. These models can simultaneously capture important features of financial return series, such as leverage effect, heavy tails, and asymmetry. The results of simulation studies show that, according to bias and root mean squared error measures, the NUTS algorithm performs well. In particular, we observe that the R package stochvol has faster execution times, but NUTS can outperform it in terms of effective sample size. Additionally, we propose the use of WAIC and leave-one-out cross validation techniques for comparing and selecting stochastic volatility models. Finally, we apply the developed methodology to real time series of financial returns.	
	Cristian Cruz , Marvin Villafranca	Vector Autoregressive model with multivariate stochastic volatility	Vector autoregressive (VAR) models have proven to be efficient in capturing the dynamic relationships of multivariate time series. Multivariate stochastic volatility (MSV) models have shown to be useful for modeling the variance as it changes over time. Therefore, this article proposes the integration of a VAR model with an MSV model (VAR-MSV). The choice of the most suitable VAR-MSV is carried out by the Deviance Information Criterion (DIC). An application was made to two key macroeconomic variables for the United States. We aggregate the SP500 stock market performance index and interpreted the results. To estimate the parameters, Monte Carlo Markov Chains (MCMC) are used. The results indicate that the VAR-MSV model captures dynamic relationships as well as variance changing over time effectively.	
	Richard Warr	Dependent Random Partitions by Shrinking Towards an Anchor	Random partition models are flexible Bayesian prior distributions which accommodate heterogeneity and the borrowing of strength by postulating that data are generated from latent clusters. The Chinese restaurant process and other stick-breaking priors are popular exchangeable random partition models used in Bayesian nonparametrics. The exchangeability assumption is not appropriate when one has a notation of which items are likely to be clustered together. We call this "best guess" partition the anchor partition and define the Shrinkage Partition (SP) distribution that takes any random partition distribution and pulls its probability mass towards the anchor partition. Since prior knowledge about the clustering of items may be different across the items, our formulation allows for differential shrinkage towards the anchor. Our distribution has a tractable normalizing constant and easily fits into standard Markov chain Monte Carlo sampling algorithms for model fitting. We explore the properties of our SP distribution and compare it to related random partition distributions. We show how our SP distribution provides a general framework to build dependent random partition models, and demonstrate our method in an application of hierarchically-dependent and time-dependent random partitions.	

16:00	16:30	Coffee Break	
16:00	16:30	Poster Session	
16:30	18:00	IPS Session: IPS Recent Developments on Computational Statistics for the Modelling of Multivariate	
Chair and organizer: Mauricio Castro			Auditorium BNP

Rosangela Loschi	Handling Categorical Features with Many Levels Using A Product Partition Model	A common difficulty in data analysis is how to handle categorical predictors with a large number of levels or categories. Few proposals have been developed to tackle this important and frequent problem. We introduce a generative model that simultaneously carries out the model fitting and the aggregation of the categorical levels into larger groups. We represent the categorical predictor by a graph where the nodes are the categories and establish a probability distribution over meaningful partitions of this graph. Conditionally on the observed data, we obtain a posterior distribution for the levels' aggregation, allowing the inference about the most probable clustering for the categories. Simultaneously, we extract inferences about all the other regression model parameters. We compare our the proposed and state-of-art methods showing that it has equally good predictive performance and more interpretable results. Our approach balances out accuracy versus interpretability, a current important concern in statistics and machine learning. This is joint work with Tulio L. Criscuolo, Renato M. Assunção, Wagner Meira Jr., and Danna Cruz-Reyes. Financial Support: CNPq, CAPES and FAPEMIG.
Anuradha Roy	Computation of parameters in linear mixed effects model for multivariate repeated measures data	The number of parameters multiplies in a linear mixed effects model in the case of multivariate repeated measures data. Computation of these parameters is a real problem with the increase in the number of response variables or with the increase in the number of time points. The problem becomes more intricate and involved with the addition of additional random effects. A multivariate analysis is not possible for these models in a small sample setting. We propose a method to estimate these parameters in bits and pieces from child models, by taking an appropriate subset of response variables at a time, and finally integrate these bits and pieces at the end to get the parameter estimates of the mother model and draw appropriate conclusions. By exploiting this method one can calculate the fixed effects, the best linear unbiased predictions (BLUPs) for the random effects, and also the BLUPs at each time for each response variable to monitor the effectiveness of the treatment for each subject. The proposed method is illustrated with an example of multiple response variables measured over multiple time points arising from a clinical trial in osteoporosis.
Armin Schwartzman	Bootstrapping Gaussian Random Fields and the Distribution of the Supremum	Gaussian random fields are common noise models for 1D, 2D and 3D imaging data in neuroscience and climatology. Performing statistical inference on such data (e.g., for detecting the presence of signal or estimating the spatial extent of the signal) often requires the height distribution of the supremum of the noise field. How can such a distribution be estimated, especially when the noise is spatially nonstationary with an unknown covariance function? This problem can be solved using the multiplier bootstrap when multiple instances of the fields are observed. This involves paying close attention to the distribution of the bootstrap multipliers and proper normalization of the observed fields.
Fernando Quintana, Garritt Page, Matthew Heiner	A projection approach to local regression with variable-dimension covariates	Incomplete covariate vectors are known to be problematic for estimation and inferences on model parameters, but their impact on prediction performance is less understood. We develop an imputation-free method that builds on a random partition model admitting variable-dimension covariates. Cluster-specific response models further incorporate covariates via linear predictors, facilitating estimation of smooth prediction surfaces with relatively few clusters. Component kernels exploit marginalization techniques to analytically project response distributions according to any pattern of missing covariates, yielding a local regression with internally consistent uncertainty propagation that utilizes only one set of coefficients per cluster. Aggressive shrinkage of these coefficients regulates uncertainty due to missing covariates. The method allows in- and out-of-sample prediction for any missingness pattern, even if the pattern in a new subject's incomplete covariate vector was not seen in the training data. We develop an MCMC algorithm for posterior sampling that improves a computationally expensive update for latent cluster allocation. Finally, we demonstrate the model's effectiveness for nonlinear point and density prediction under various circumstances by comparing with other recent methods for regression of variable dimensions on synthetic and real data.

16:30	18:00	IPS Session: Applied Finance and Data Science	
Chair and organizer: Gabriel Pino			Auditorium FE

José Barralez-Ruiz, Gabriel Pino	On the effect of short-run and long-run US Economic expectations on oil and gold volatilities	This paper investigates the effect of US Economic expectations on oil and gold volatilities for different time horizons. To this end, we compute expectations using an MS-VAR model from one (short run) to (long run) sixteen quarters ahead. Then, we estimate the impact of an expectation shock on oil and gold volatility measures through an impulse response function estimating a VAR model from 1987Q1 to 2022Q1. Our results show that oil volatility is significantly affected by a shock to short-run expectations, while gold volatility is affected by a shock to long-run expectations. In particular, gold and oil volatilities decrease from one to two quarters after a positive expectation shock.
---	---	---

	Jocelyn Tapia, Fernando Diaz , Alba Martinez	A New Model-Agnostic Approach for the Estimation of Marginal Effects based on a Synthetic Prediction Sample	This investigation presents the new model-agnostic approach Synthetic Prediction Sample for computing marginal effects and constructing their corresponding confidence intervals in machine learning models. The main objective is to contribute to the interpretable machine learning literature by addressing the limitations of partial dependent plots or local interpretable model-agnostic explanations models. We evaluate the performance of the new approach by comparing it with traditional hedonic models and the plot marginal effect of variables of Ishwaran and Kogalur. In addition, we perform several experiments using a comprehensive database of real estate prices in Santiago, Chile, and data sets commonly used in the literature. Nonparametric bootstrapping is implemented to estimate confidence intervals.
	Jaime Lavin, Mauricio Valle, Nicolas Magner	Stock Market Pattern Recognition using Symbol Entropy Analysis	Financial assets strongly relate to market uncertainty dynamics. The emergence of global and local shocks provokes price patterns in stock markets that jeopardize whole market stability. Combining symbol entropy analysis, regressions, and forecast variance decomposition models, we study the relationships between the frequency of states associated with rising and down price patterns of the S&P 500 index, its implied volatility, and the fluctuations in the uncertainty of the global stock market -gauged by the symbol entropy of the All Country World Index (ACWI). We find a positive (negative) relationship between the increase (decrease) in global uncertainty and the frequency of extreme bear (bull) episodes. Similarly, we observe the same association between increases (decreases) in the implied volatility and the frequency of extreme bear (bull) episodes. Furthermore, joint increases (decreases) between implied volatility and global uncertainty are positively (negatively) associated with the frequency of bear (bull) episodes in the S&P 500 index. Finally, interconnectedness analysis confirms the relevance of the reciprocal influence and the dynamic nature of the phenomenon, which rises the probability of observing shock waves of financial instability throughout the global and local stock markets.
	Nicolás Hardy , Pablo Pincheira	"Fueling Predictability:" Can Commodity-Equities Forecast Fuel Prices?	In this paper we show that several MSCI stock indices have a remarkable ability to predict the returns of oil prices (WTI and Brent) and of three additional oil-related products: gasoline, propane and heating oil. The theoretical underpinnings of our findings rely on the present-value theory for stock price determination and on the strong co-movement displayed by some industrial commodity prices. Interestingly, this predictive ability is stronger than that embedded in commodity-currencies. We find substantial evidence of predictability both in-sample and out-of-sample. One distinctive feature of our paper is a focus on MSPE differences at both the population and sample level. While several papers in the literature have recently found predictability for commodity returns at the population level, they typically tend to show poor gains in MSPE at the sample level. We address this failure with a simple approach based on the covariance between the target variable and our forecasts. With our approach we find substantial evidence of predictability at the sample level as well, in sharp contrast with the weak results reported by the traditional Giacomini and White (2006)/ Diebold and Mariano(1995)-West(1996) test when evaluating MSPE differences.
18:00	19:00	Keynote Speaker 6: Trevor Harris Chair: Fernando Quintana	Auditorium BNP
	Trevor Harris	Multi-model ensemble analysis with neural network Gaussian processes	Multi-model ensemble analysis integrates information from multiple climate models into a unified projection. However, existing integration approaches based on model averaging can dilute fine-scale spatial information and incur bias from rescaling low-resolution climate models. We propose a statistical approach, called NN-GPR, using Gaussian process regression (GPR) with an infinitely wide deep neural network based covariance function. NN-GPR requires no assumptions about the relationships between climate models, no interpolation to a common grid, and automatically downscales as part of its prediction algorithm. Model experiments show that NN-GPR can be highly skillful at surface temperature and precipitation forecasting by preserving geospatial signals at multiple scales and capturing inter-annual variability. Our projections particularly show improved accuracy and uncertainty quantification skill in regions of high variability, which allows us to cheaply assess tail behavior at a 50 km spatial resolution without a regional climate model (RCM). Evaluations on reanalysis data and SSP2-4.5 forced climate models show that NN-GPR produces similar, overall climatologies to the model ensemble while better capturing fine scale spatial patterns. Finally, we compare NN-GPR's regional predictions against two RCMs and show that NN-GPR can rival the performance of RCMs using only global model data as input.
19:30	22:00	Coferece Dinner	

FRIDAY 10 NOVEMBER 2023

08:30	09:30	Keynote Speaker 7: Katherine Ensor		Auditorium BNP
		Chair: Luis Valdivieso		
		Katherine Ensor	Multivariate non-linear time series nowcasting with spatial considerations and applications to wastewater epidemiology	Of important consideration are multivariate nonlinear dynamic time series with low to high levels of spatial association. We explore a state-space hierarchical modeling approach, considering both a frequentist and Bayesian perspective. Key questions answered are natural clusterings of the time series, short-term deviations between the series, and short-term predictions based on the fitted models. The methodology is applied to fifty weekly time series spanning three years, representing wastewater signals for SARS CoV-2. Wastwater signals are compared to the corresponding observed cases. From this paradigm, a predictive model for emergent diseases is posited.
09:30	11:00	CA Session: Data Science on Climate Change and Engineering		Auditorium BNP
		Chair: Daniela Catro-Camilo		
		Holger Cevallos-Valdiviezo, Gema Zambrano-Zambrano	Intruder Detection in Security Videos: A Data-Based Approach Using Robust PCA Estimators	While video surveillance by guards is commonly employed as a strategy to detect burglars in private properties, it is acknowledged to be an expensive and sometimes ineffective method for intruder detection in certain cases. In this work, we propose using a data-based approach that utilizes vectorized images from security videos to detect intruders or other anomalies. Specifically, we identify suspect images in security videos by applying robust principal component estimators to the image data from the videos. Robust PCA aims to estimate the low-dimensional subspace that best approximates regular image data from the videos. This allows for the detection of outlying images by examining how far each image deviates with respect to the robust subspace (orthogonal distances) and/or how far each image deviates in terms of projections onto the robust subspace (score distances). In this work, we compare several robust PCA estimators and assess their effectiveness in detecting anomalies in security videos. Our data-based approach yields promising results for anomaly detection in security videos and can also be seen as a cost-effective alternative to motion sensors and video surveillance by guards. Moreover, our approach is easy to implement and can be seamlessly integrated with existing security systems.
		Martha Bohorquez Castañeda	Network Analysis for Detection of Spatio Temporal Patterns	Our goal is to provide new tools to analyze spatio-temporal event networks. We combine statistical methods with building-network methods, to find marked patterns of theft occurrences. Time series of directed event networks are constructed for a sequence of spatial distances. The spatial distance that generates the strongest change of the event network connections is found. In addition, the degree of centrality and assortativity are modeled. Finally, we propose an empirical random network event generator to detect significant motifs throughout time. This generator preserves the spatial configuration but randomizes the order of the occurrence of events. To prevent the large number of links from masking the count of motifs, we propose using standardized counts of motifs at each time slot. Our methodology can detect interaction radius in space, build time series of networks, and describe changes in its topology over time, using identification of different types of motifs that allows for the understanding of the spatio-temporal dynamics of the phenomena. We illustrate our methodology by analyzing thefts occurred in Medellín and Palmira, Colombia.
		Euloge Kouame, Falikou Dosso	Well allocation and performance prediction: approach by machine learning	In oil and gas fields, production from several wells is commingled. Prediction and classifying oil and gas production for each well is difficult and it's based most of time on individual well testing which may not be feasible due to operational constraints. The aim of this study is to find an appropriate method to allocate accurately the contribution of each well to the total oil production rate of an oilfield. We use an efficient implementation of machine learning classification algorithms and compare their performance to the approach used on the field. The model solution takes a set of reservoir and process parameters as inputs and returns the oil rate as output. The model is trained using actual well test data from producing wells. The results show the high accuracy of this approach. This approach gives real time well performance and allow better intervention planning and production forecast for the field

		Daniela Castro-Camilo , Erin Bryce, Luigi Lombardo	Improving landslide hazard modelling in Scotland: enhanced predictions, uncertainty evaluation and residual analysis for model validation	We aim to improve the landslide susceptibility model for Scotland, working closely with the British Geological Survey (BGS) to incorporate updated covariates and recent storm-triggered landslides. The current BGS model is constructed by assigning a binary response for the presence/absence of landslides, reducing the information to a single binary classifier that only accounts for spatial dependence through spatially varying covariates. Our approach uses a log-Gaussian Cox Process (LGCP) to predict landslide intensity by modelling the expected landslide count, which provides a more detailed and complete characterisation of these hazardous events. Fast and accurate inference is carried out using the integrated nested Laplace approximation (INLA) with the stochastic partial differential equation (SPDE) approach. To make our findings accessible to everyone, we propose a novel and intuitive method for communicating aggregated predicted landslide intensities and their associated uncertainties, adhering to the guidelines set by BGS. Finally, we explore the challenges of model validation for the LGCP, which is far from being straightforward since the interest lies in the spatial structure of points in space rather than some measure at a point. To this end, we study the benefits of residual analysis for LGCPs as a model validation tool.
09:30	11:00	IPS Session: R Packages for Data Science		Auditorium FP
		Chair and organizer: Han-Ming Wu		
		Local Chair: Natalia da Silva		
		Li-Pang Chen, Jou-Chin Wu	Causal Inference with Error-prone Treatments and Applications of the R Package caret.	In the framework causal inference, average treatment effect (ATE) is one of crucial concerns. To estimate it, the propensity score based estimation method and its variants have been widely adopted. In existing literature, parametric logistic regression models with precisely measured data is a standard setup to estimate the propensity score. However, they are restrictive to handle error-prone data and nonlinear confounders. To address those issues and derive reliable estimators of ATE, we aim to propose a corrected treatment to eliminate measurement error effects. After that, we implement the random forest method in the R package 'caret' to estimate the propensity score. Based on the new estimation of the propensity scores, the corresponding ATE can be accurately estimated. Numerical studies are also conducted to assess the finite sample performance of the proposed estimator, and numerical results justify that the estimator based on the random forest method outperforms the estimator with the propensity score being estimated by the conventional parametric methods.
		Li-Pang Chen, Cheng-Kuan Lin , Su-Fen Yang	EATME: An R package for EWMA control charts with adjustments of measurement error	In this paper, we introduce an R package EATME, which is known as Exponentially weighted moving average control chart with Adjustments To Measurement Error. The main purpose of this package is to correct for measurement error effects in continuous or binary random variables and develop the corrected control charts based on the EWMA statistic. In addition, the corrected control charts can detect out-of control process parameters accurately. The package contains a function to generate artificial data and includes functions to determine the reasonable coefficient of control limit as well as estimate average run length (ARL). Moreover, for the visualization, we also provide the plots to show the monitoring of in-control and out-of control process. Finally, the functions in this package are clearly demonstrated, and numerical studies show the validity of our method.
		Po-Wei Chen, Han-Ming Wu	dataSDA: Datasets for Symbolic Data Analysis in R	Nowadays, data collection has become increasingly complex and extensive. The representation of data is no longer limited to single values but also includes intervals, histograms, and/or distributions. These are examples of what is known as symbolic data. Symbolic data analysis (SDA) refers to the methods and techniques developed for analyzing such types of datasets. This study focuses on the development of an R package called dataSDA, which aims to provide a comprehensive repository for research and application in symbolic data science. The package is designed to collect a diverse range of symbolic data from various sources, including prominent SDA textbooks, as well as other literature and donated datasets. These datasets are categorized based on their applicability to different analysis tasks, serving as benchmarks for evaluating symbolic data analysis methods. Furthermore, dataSDA offers essential features such as calculating symbolic descriptive statistics, aggregating traditional data into symbolic format, converting between symbolic data formats, and applying preprocessing techniques like standardization. We will discuss the implementation of dataSDA and demonstrate its utility through various analysis tasks, including dimension reduction, clustering, classification, and regression, using symbolic datasets available within the dataSDA package.

11:00	11:30	Coffee Break		
11:30	13:00	IPS Session: Women in Data Science: Recent Theoretical Research and Applications		
		Chair and organizer: Carolina Marchant		Auditorium BNP
		Tamara Fernandez Aguilar, Nicolas Rivera Aburto	A general framework for kernel-based tests	Kernel-based tests provide a simple yet effective framework that uses the theory of reproducing kernel Hilbert spaces to design non parametric testing procedures. These procedures can be applied to many different testing problems, and to virtually any type of data. While kernel methods were initially investigated by the machine learning community, their applications in classical statistical problems have piqued the statistical community's interest. In this talk, I will introduce the main ideas of kernel-based testing and I will review some of the most important recent developments in the area. In particular, we will show new theoretical results that lead to a simple yet effective analysis of kernel-based tests. These results are based on limiting theorems for random functionals on separable Hilbert Spaces, rather than the lengthy and complicated V-statistics expansions commonly used in the literature. We will show practical applications to some well-known problems in statistics, such as conditional independence testing.
		Xavier Lopez	Ciencia de datos y aprendizaje de máquina: una mirada desde la data hasta modelos de aplicación	La presentación se enmarca en dar a conocer cómo desde la data en su formato más bruto, podemos llegar a su aplicación en modelos y sistemas inteligentes, los cuales permiten implementación de modelos de aprendizaje de máquina para la predicción de diversos objetivos de investigación. En específico, se mostrarán aplicaciones en el campo de la biomedicina, medio ambiente y energías renovables. Los modelos se basan en aprendizaje supervisado y redes neuronales profundas, con un enfoque de búsqueda bayesiana para optimización de hiperparámetros, además de implementación de técnicas apropiadas para abordar desbalanceo de clases.
		Yolanda Gómez, Diego Gallardo, Jeremias Leão, Vinicius Calsavara	On a new piecewise regression model with cure rate: Diagnostics and application to medical data	In this work, we discuss an extension of the classical negative binomial cure rate model with piecewise exponential distribution of the time to event for concurrent causes, which enables the modeling of monotonic and non-monotonic hazard functions (ie, the shape of the hazard function is not assumed as in traditional parametric models). This approach produces a flexible cure rate model, depending on the choice of time partition. We discuss local influence on this negative binomial power piecewise exponential model. We report on Monte Carlo simulation studies and application of the model to real melanoma and leukemia datasets.
		Carolina Marchant, Luis Sánchez, Germán Ibacache-Pulgar	New varying-coefficients quantile regression models with application to Chilean pollution data	Many phenomena can be described by random variables that follow asymmetrical distributions. In the context of regression, when the response variable Y follows such a distribution, it is preferable to estimate the response variable for predictor values using the conditional median. Quantile regression models can be employed for this purpose. However, traditional models do not incorporate a distributional assumption for the response variable. To introduce a distributional assumption while maintaining the flexibility of the model, we propose new varying-coefficients quantile regression models based on the family of log-symmetric distributions. We achieve this by reparameterizing the distribution of the response variable using quantiles. Parameter estimation is performed using a maximum likelihood penalized method, and a back-fitting algorithm is developed. Additionally, we propose diagnostic techniques to identify potentially influential local observations and leverage points. Finally, we apply and illustrate the methodology using real pollution data from Padre Las Casas city, one of the three most polluted cities in Latin America and the Caribbean according to the World Air Quality Index Ranking (https://www.iqair.com/).
11:30	13:00	IPS Session: Local Influence and Robustness in Regression Models: New Perspectives for Statistical Learning		
		Chair and organizer: Manuel Galea		Auditorium FP
		Felipe Osorio, Manuel Galea, Patricia Giménez	A robust approach for generalized linear models based on maximum likelihood procedure	In this work we propose a procedure for robust estimation in the context of generalized linear models based on the maximum likelihood method. An estimation algorithm that represents a natural extension of the usual iteratively weighted least squares method in generalized linear models is presented. The asymptotic distribution of the proposed estimator and a set of statistics for testing linear hypothesis are discussed, which allows the definition of standardized residuals using the mean-shift outlier model. Robust versions of deviance function and the Akaike information criterion are defined with the aim of providing tools for model selection. The performance of the proposed methodology are illustrated through a simulation study and analysis of a real dataset.
		Fernanda De Bastiani, Jonathan Acosta, Manuel Galea, Miguel Uribe-Opazo	Local influence for Gaussian spatio-temporal model	Spatio-temporal models are well known in the literature, however to present statistical methods to evaluate the temporal dependence between observations in this type of model is still a challenging task. We propose case weight perturbation scheme under local influence approach to check the assumptions of temporal structure in Gaussian spatio-temporal linear models with repeated measures. We evaluate the presence of temporal structure in a soybean productivity data set collected at the same sites over seven years, and on the explanatory variables, the soil chemical contents. This methodology can be considered as a goodness of fit, verifying assumptions such as independence across time and heteroskedasticity. The results showed that there is no misleading in considering a model with independent repeated measures for this data set. Extension of this methodology to other distributions such as the elliptical family of distributions and to explore assumptions as well as separable covariance functions are straightforward.

		Manuel Galea	Robust estimation in a functional measurement error model using the Lq-likelihood function	A procedure is proposed for robust estimation of the structural parameter in a functional measurement error model. The proposed estimator is obtained, based on maximum Lq-likelihood approach, first replacing the incidental parameters by estimators depending on the structural parameter. The estimation procedure can be implemented easily by a simple and fast re-weighting algorithm. Consistency and asymptotic normality is established and the covariance matrix is given. Theoretical properties, ease of implementability and empirical results on simulated and real data show the satisfactory behavior of the MLqE and its advantages over the MLE in presence of observations discordant with the assumed model.
		Alejandra Tapia	Perturbation selection and local influence for binary regression models	Perturbation selection and local influence are techniques used in statistical modeling to study the sensitivity of maximum likelihood estimators and other relevant quantities for the model. Binary regression models are statistical models used to analyze binary outcome variables. The goal of binary regression models is to understand and model the relationship between the predictor variables and the probability of the binary outcome occurring. This relationship is based on the structural assumption of a link function, which maps the linear combination of the predictor variables to the range [0,1]. In this work, we propose specifying the binary regression model using a general class of asymmetric link functions and discuss parameter estimation and hypothesis testing. We also explore the perturbation selection and local influence techniques to evaluate the symmetry of the link function, in the direction of asymmetric link functions. Specifically, we propose an appropriate perturbation scheme called skew-link perturbation to assess the symmetry of the link function of the model. Thus, under this skew-link perturbation, we implement a formal approach to sensitivity assessment of the estimators and odds ratio of the model. We present the results of Monte Carlo simulation studies and applications of the methodology to real medical datasets.
13:00	14:00	Lunch		
14:00	15:00	Keynote Speaker 8: Paulo Canas Rodrigues		
		Chair: Rodrigo Salas		
		Paulo Canas Rodrigues	The usefulness of singular spectrum analysis in hybrid methodologies for time series forecasting	Time series forecasting plays a key role in areas such as energy, environment, economy, and finances. Hybrid methodologies, combining the results of statistical and machine learning methods, have become popular for time series analysis and forecasting, as they allow researchers to compensate for the limitations of one approach with the strengths of the other and combine them into new frameworks while improving forecasting accuracy. In this class of methods, algorithms for time series forecasting are applied sequentially, i.e., the second model can be applied to the residuals that were not captured by the first by considering that the observed data is a combination of linear and nonlinear components. In this talk, I will discuss several strategies for time series forecasting that use singular spectrum analysis in hybrid methodologies, with application to electricity load forecasting and to PM10 (inhalable particles, with diameters that are generally 10 micrometers and smaller) forecasting.
15:00	16:00	CA Session: Bayesian Analysis, Psychometrics, and Signal Processing		
		Chair: Luis Valdivieso		
		Alex Rodrigo dos Santos Sousa	A class of priors to perform asymmetric wavelet shrinkage	In Bayesian wavelet shrinkage, the already proposed priors to wavelet coefficients are assumed to be symmetric around zero. Although this assumption is reasonable in many applications, it is not general. The present work proposes the use of asymmetric shrinkage rules based on the discrete mixture of a point mass function at zero and an asymmetric distribution in a class composed by beta, triangular, Kumaraswamy and skew normal distributions as prior to the wavelet coefficients in a non-parametric regression model. Statistical properties such as bias, variance, classical and Bayesian risks of the associated asymmetric rule are provided, Bayesian robustness is discussed and performances of the proposed rule are obtained in simulation studies involving artificial asymmetric distributed coefficients and the Donoho-Johnstone test functions. Applications in seismic and chemical real datasets is also analyzed.
		Renato da Silva Fernandes, Jorge Luis Bazán Guzmán, Mariana Cúri	Analyzing different Constraints on item parameters in the Bayesian estimation of G-DINA model	The Cognitive Diagnostic models are a class of discrete latent variable models where the probability of correct response to an item is defined in function of the individuals' latent attributes possession. The G-DINA model is a CDM with great flexibility, capable of incorporating aspects of several CDMs in its formulation. Despite its advantages, the generality of this model may lead to unexpected results, such as the less skilled individuals presenting the highest probability of correct response in certain items. To avoid these results, it is usual to impose some constraints on the item parameters. In this work, we analyze different restrictions on the parametric space of the G-DINA model and its effects on parameter estimation under a Bayesian approach. We conduct simulation studies to evaluate the parameters recovery and the estimation accuracy of the different constraints. Furthermore, we analyze and compare the effect of the constraints in a real data application example.

Auditorium BNP

Auditorium BNP

		Zaida Quiroz, Luis Valdivieso, Cristian Bayes	A Beta Inflated Spatial Model for Assessment of Reading Level	An indicator of a country's reading comprehension progress is the proportion of students who achieve, within a region, the highest reading level (HRP). Sometimes these proportions turn out to be zero, especially in regions where students have poor reading skills, and are spatially correlated between regions; that is, they are influenced by their closest neighboring regions. This work proposes a new beta inflated mean regression model with spatial effects that potentially explain the expected HRP. Their estimation is performed using a Hamiltonian Markov Monte Carlo algorithm implemented in Stan. The model is applied to the results of the census evaluation of students in the city of Lima-Peru, at the district level, where some bayesian divergence measures are included to identify districts with atypical HRP.
15:00	16:00	IPS Session: Machine Learning and Deep Learning Applications and Challenges		
		Chair and organizer: Rodrigo Salas		Auditorium FP
		Cesar Roudergue, Romina Torres	Predicting the Next Step of a Multistage Attack in CTF events using the Hidden Markov Model	Multi-Stage Network Attack (MSNA) are a sequence of correlated sub-attacks with a common aim. A team participating in Capture the Flag event (CTF) must in parallel mitigate MSNAs perpetrated by multiple attacking teams. Since a victim has limited resources, it is not possible to mitigate all MSNAs. Hence the importance of knowing the probable evolution of a MSNA for its prioritization. In this work we use Hidden Markov Models (HMM) to model the visible observations (alerts generated by Intrusion Detection System given the sub-attacks) and the hidden ones (stages through which an MSNA evolves). In particular, we model stages of MSNA using a simplified version of the Cyber-kill-chain and we implement our proposal using the hmmlearn library with unsupervised learning using the Baum Welch algorithm. We generated three models for three attacking teams seen from the point of view of a team receiving those attacks (a.k.a. victim) during the three-day DefCon CTF 22 event. Results are encouraging. Applied to the CTF data, using a fixed window of 200 events and random initialization, the victim obtained an hmm model with 72% accuracy for predicting next step and stage of one of their attackers.
		Pascal Sigel, Daira Velandia, Jorge Saavedra, Jorge Arévalo, Rodrigo Salas	Deep learning methods applied to the detection of lake surface changes using satellite images	Due to climate change, a megadrought has affected Chile since 2010, leading to a significant decrease in precipitation. In the V Region of Chile, Lake Peñuelas, a water resource for the city of Valparaíso, has experienced a substantial reduction in its capacity, forcing its use as a supply source to be suspended since January 2021. This study employs the Principal Component method, the Normalized Difference Water Index (NDWI), and the Mann Kendall test to perform a multitemporal analysis of the lake's surface between the years 2010 and 2022, using images obtained from Landsat-8 and Landsat-9 satellites. The obtained results indicate an accelerated decrease in the surface of Lake Peñuelas beginning in 2019, with an estimated loss of 94.8% of its surface by 2022. Moreover, a Convolutional Neural Network (CNN) model was implemented to predict NDWI images and quantify changes in water bodies according to the seasons of the year. These analyses and models can provide valuable information for implementing effective control measures and protecting the country's water resources.
16:00	16:30	Coffee Break		
16:30	18:00	IPS Session: Machine Learning and Deep Learning Applications and Challenges		
		Chair and organizer: Rodrigo Salas		Auditorium BNP
		Roberto Leon, Kevin Voss, Carola Blazquez	Application of clustering techniques for a combinatorial problem in the conformation of new ligands	New effective treatments for cancer are necessary. Several computational methods can assist these treatments. Tumor development is favored by a set of proteins called Hsp (Heat Shock Protein). These proteins favor cell proliferation and are related to the development of cancer. It is proposed to inhibit the Hsp to reduce their incidence in cancer development. This inhibition can be possible with the design and development of new components called ligands, which interact with the Hsp. A computational approach proposed focused on a Fragment-Based Drug Design technique consisting of randomly selecting Hsp90 ligands and deconstructing them into fragments. These fragments are combined, and new ligands are generated. Their inhibition level to the Hsp90 is compared with previous ligands. A promissory ligand has a negative interaction energy with the Hsp90. Currently, the initial selection of ligands is manual. The goal is to apply clustering techniques in the selection of ligands before the deconstruction and reconstruction process. It is proposed to study and analyze different clustering techniques to generate new groups of known ligands and to study the impact of the quality of the generated ligands. Three clustering techniques will be applied: k-means, DBSCAN, and Hierarchical clustering.

		Alejandra Bravo-Diaz , Sebastian Moreno, Javier Lopatin	Analyzing transfer learning for Pinus radiata detection from images captured by drones using convolutional neural networks	Pinus radiata (P. radiata) is a highly invasive species in native forests from Chile, affecting the functioning and structure of ecosystems. Models based on convolutional neural networks (CNN) are a promising alternative to detect P. radiata in high-resolution remote sensing data. However, current studies are limited in their evaluations of transferability to new sectors, hampering the ability to use these models in a real environment. We evaluate the transferability prediction of P. radiata from images captured by drones with RGB sensors using different CNN-based architectures. Five sectors of the Maule coastal were considered, where the fit and transferability of models trained with information from a single sector and different sectors were evaluated to include different levels of spatial variability. We also searched for an empirical model that maximizes transferability using labeled data in the zones to be transferred. The results demonstrate that pre-trained networks with fine-tuning benefits the performance of the models and their ability to transfer learning. In addition, single-sector models are ideal for predicting within the same sector (low spatial heterogeneity); however, the results show a high variability when applied to dissimilar areas. In contrast, multi-sector models, on average, are better at transferring learning, indifferent over territorial situations.
		Leondry Mayeta , Julio Sotelo, Steren Chabert, Marvin Querales, Francisco Torres, Rodrigo Salas	Fuzzy Inference System for brain tumors segmentation based on Magnetic Resonance Imaging and Deep Learning	Brain tumors represent a tumor in the central nervous system, and their detection at early stages is a key issue for providing improved treatment. Magnetic resonance imaging (MRI) is a widely used imaging technique to assess these tumors, but the large amount of data produced by MRI prevents manual segmentation in a reasonable time. Deep Learning (DL) methods obtain high segmentation performances, but these models do not provide any evidence regarding their process to perform this task. To overcome this issue, we propose a novel Fuzzy Inference Systems that includes a deep learning network to segment tumors from MRI. We have applied the proposed model to the Multimodal Brain Tumor Segmentation Challenge where we have obtained more reliable segmentation results.
		Gabriel Guerra , Rodrigo Salas	Classification of Parkinson's Disease based on Biomedical Voice Measurements using Explainable Machine Learning models	The objective is to discriminate between healthy subjects from those with Parkinson's disease based on biomedical voice measurements using explainable machine learning models. Data separation based on Subjects allocations in the training and testing datasets is explored to evaluate performance differences. We highlight the importance of the hold-out-subject (HOS) scheme to evaluate the generalization performances using health data. Additionally, assessing the feature importance using SHAP (SHapley Additive exPlanations) is essential to obtain more reliable predictions. Our research contributes to developing a robust classification framework for predicting Parkinson's disease based on voice measurements. This work emphasizes the need to consider the HOS scheme to ensure reliable and effective Parkinson's disease classification using explainable machine learning models.
16:30	18:00	IPS Session: Advanced Topics in Machine Learning and Complex Data		
		Chair and organizer: Eufrásio de Andrade Lima Neto		Auditorium FP
		Local Chair: Jonathan Vergara		
		José Natanael Andrade de Sá , Marcelo Ferreira, Francisco de Assis Tenório de Carvalho	Co-clustering based on kernel functions with Variable Weighting	Kernel functions have been used successfully in clustering algorithms to deal with overlapping clusters efficiently. Bringing this idea to co-clustering, we propose two kernel-based fuzzy co-clustering algorithms based on the fuzzy double Kmeans (FDK). The first proposed algorithm, the Gaussian kernel fuzzy double Kmeans (GKFDK), is based on FDK and computes the cluster prototypes in the original feature space. The second algorithm, the Weighted gaussian kernel fuzzy double Kmeans (WGKFDK), is an extension of the GKFDK with automated variable weighting, that distinguishes the relevance of the variables in each cluster. Experiments performed with both synthetic and real data, in comparison with previous state-of-the-art co-clustering algorithms, showed the effectiveness of the proposed algorithms.
		Marcelo Ferreira	Spectral clustering of planar shapes	With the advancement of technology, the collection of geometrical information from images has become common. Statistical shape analysis utilizes statistical methods to analyze geometrical structures and can be widely applied. One particular area of interest is the adaptation of classical methods for shape data or the development of new methods. In statistical shape analysis, there is often a need to cluster shapes to obtain similar groups. The k-means algorithm is one of the most widely used methods. However, despite its simplicity and efficiency, the k-means algorithm has some limitations. Therefore, it is important to propose alternative methods that can be effective where the k-means algorithm fails. Spectral clustering originates from the spectral theory of graphs, and the clustering problem can be formulated as a graph cut, where an appropriate objective function needs to be optimized. In this study, we propose an adaptation of the Ng, Jordan & Weiss spectral clustering algorithm for planar shape data. We conducted experiments on 14 shape datasets and found that the proposed algorithm, considering both the full Procrustes distance and the Euclidean distance in the shapes' tangent space, outperforms the k-means algorithm for planar shapes, corroborating that the proposed adaptation is efficient for shape data.

	Telmo Silva Filho	Advances in Machine Learning Evaluation using Item Response Theory	Item response theory (IRT) offers the possibility of evaluating respondents with different task-related abilities, by taking into account differences in difficulty of the items that make up an evaluation setup. In machine learning, IRT has been successfully applied to evaluate instance hardness and the abilities of models to correctly predict the target values of instances in a test set. Here, correctness can be seen as a binary value (0, for wrong predictions, 1 for correct ones), or a continuous value, e.g. the probability assigned to the correct class or, in regression, (the inverse of) the magnitude of the error associated to a prediction. Recently, IRT has also been investigated for evaluating clustering models and some work has been done to understand the actual implications of IRT's parameters when applied to machine learning problems, including questions such as how to select a good pool of models to evaluate instance hardness, what would be the ability of the Bayes optimal model, and what is an instance's true hardness. Thus, this talk aims to cover these recent developments aiming to contribute to advances in machine learning evaluation.
	Eufrasio de Andrade Lima Neto, Georgina Cosma, Axel Finke, Jonathan Bailiss, Jo Miller	Ethical AI for Enhancing Decision-Making Processes in Young People Requiring Early Help Services	Local authorities in the UK provide early help services aimed at supporting young people and their families. The COVID-19 pandemic revealed a significant decrease in referrals, suggesting an over-reliance on the referral process for identifying support-needing individuals. To address this issue, this study investigates the application of data science techniques and machine learning algorithms to support the decision-making process of referrals and assessments. Experimental validation and testing of various machine learning models were conducted, with the most accurate models recommended for evaluation on unseen test data. Moreover, fairness techniques were employed to mitigate biases and prevent discrimination against minority groups within the machine learning models. The results revealed that decision-making models can support the process of identifying vulnerable young people who may need early help services.
18:00	19:00	Keynote Speaker 9: Fabrizio Ruggeri Local Chair: Luis Firinguetti Fabrizio Ruggeri	Auditorium BNP
		Advances in Adversarial Classification	In multiple domains such as malware detection, automated driving systems, or fraud detection, classification algorithms are susceptible to being attacked by malicious agents willing to perturb the value of instance covariates in search of certain goals. Such problems pertain to the field of adversarial machine learning and have been mainly dealt with, perhaps implicitly, through game-theoretic ideas with strong underlying common knowledge assumptions. These are not realistic in numerous application domains in relation to security. We present an alternative statistical framework that accounts for the lack of knowledge about the attacker's behavior using adversarial risk analysis concepts.
19:00	19:30	Closing Ceremony	
POSTERS PRESENTATIONS			
Thursday 9 November 11:00 to 11:30, 13:00 to 14:00, 16:00 to 16:30			
	John L. Santibáñez, Diego I. Gallardo, Yolanda M. Gomez	A modified cure rate model based on the piecewise regression distribution with applications to cancer dataset	A new cure rate survival model is proposed, using the power piecewise exponential distribution for failure times and the binomial, poisson, negative binomial, Haight, Borel, logarithmic, and restricted generalized poisson distributions for model the number of concurrent causes. The properties of the model are studied, the estimation of parameters is carried out by a classical approach through the EM algorithm. A simulation study is presented to demonstrate the consistency of the estimators in finite samples. Finally, an application is made with data from the medical area, to demonstrate the effectiveness of the model in relation to others in the literature.
	Cristian Bayes, Luis Valdivieso	A bayesian graph-based cluster model with effect fusion	We propose in this article, a graph-based cluster model for areal data that generates the clusters by removing edges from a given spanning tree. A key assumption of our model is that the effects inside a cluster region are nearly identical but these effects are different from other clusters. To achieve this goal, we have considered a spike-slab prior for the differences between effects. Estimation is carried out from a Bayesian perspective, using a Gibbs sampler algorithm to sample from the posterior distribution of the parameters. An application to a real dataset is presented to illustrate this new model.

Francisco Segovia, Luis Gutiérrez, Ramsés Mena	Bayesian model selection for some useful regression models	Regression analysis aims to explore the relationship between a response variable and predictors. A key aspect of regression analysis is model selection, which allows the investigator to decide which predictors are relevant to the response distribution considering a parsimony criterion. A standard frequentist model selection strategy is to explore the model space using, for instance, a Stepwise strategy based on some goodness of fit criteria. A popular Bayesian model selection strategy is the spike-and-slab methodology, which assigns a specific prior to the predictor coefficients by defining a latent binary vector that will indicate which predictors are relevant. Such a strategy includes a prior over the binary vector to penalize complex models. In this talk, we discuss a general Bayesian strategy for model selection in a broad range of regression models, using the spike-and-slab strategy and a data augmentation technique. We show that if the likelihood function follows certain conditions, the asymptotic good behavior of the Bayes factor is guaranteed alongside the availability of closed-form expressions for the posterior distribution. We present five regression models based on different choices for the response distribution, providing the necessary details for each model to be implemented alongside a Monte Carlo simulation study.
Hamel Elhadj	TCL of MSE of functional regression estimator	A popular stochastic measure of the distance between any unknown function is the integrated square error. In this work, we study the central limit theorem (CLT) for quadratic error (ISE - MISE) of Nadaraya-Watson regression estimator when the explanatory variable X is valued in some abstract semi-metric functional space and the response variable Y is real-valued. Our results may be applicable for various nonparametric regression models and bandwidth selection.
Matilda Tapia, Alba Martínez, Pablo Lemus	Functional Data Analysis of the Temperature Patterns in Chile	Climate change has triggered an accelerated increase in temperature worldwide, which has negative consequences for the environment and our society. This study aims to determine the monthly temperature patterns in six Chilean cities: Calama, Antofagasta, Santiago, Valparaíso, Chillán, and Concepción. The study includes the years from 1980 to 2022. The data analysis combines a statistical description of the data and functional principal component analysis (PCA). B-splines with roughness penalty is used to transform the data to smooth functions and generalized cross-validation to determine the value of the penalty parameter. Functional PCA makes it possible to explain the covariance structure of the data and to identify and model the main modes of variation of data as continuous functions. This provides much more information about the phenomenon under study. Results show that the minimum monthly temperatures in Calama and Santiago have increased steadily over the last 40 years, shortening the difference between maximum and minimum temperatures over time.
Fabián Gómez, Andrés Iturriaga	Analysis of fine particulate matter 2.5 during the winter periods from 2018 to 2022 in the city of Santiago, Chile, using functional data analysis tools	Fine particulate matter 2.5 (PM 2.5) is a type of harmful particle to health, and its monitoring aims to establish the air quality that a region of a country may have. In this work, functional data analysis tools are used to analyze the concentration of PM 2.5 during the winter periods from 2018 to 2022 at the Parque O'Higgins monitoring station. Firstly, the smoothing of the curves is performed, describing their hourly functional behavior for each day in all winters. Then, a functional analysis of variance is conducted to study if there are any differences in the average curves of each winter, seeking patterns of behavior between the years, in contrast to the current decontamination plans in Santiago, Chile. Finally, the incidence of NO ₂ on the concentration of PM 2.5 is studied through a functional regression model for all periods, with the purpose of exploring if there are any variations in this incidence over the years.
Brian Vergara Bravo, Alba Martínez Ruiz, Pablo Lemus Henríquez	A Comparison of Methods for Time Series Cross-Validation	This research aims to compare and evaluate four procedures for time series cross-validation: splitting, growing and sliding window approach, and the nested forward-chaining system. These methods mainly attempt to preserve the temporal structure of the data, which is affected by resampling. However, there is little evidence about which approach is more appropriate given the nature of the data. We implemented the procedures to validate a feedforward multilayer neural network and a support vector machine. These supervised learning models are trained to predict the IPSA (índice de precios selectivo de acciones) financial series and components. Predictive performance is reported in terms of mean square error and mean absolute percentage error.

Alex Centeno	An Entropy in Complex Networks with Latent Interaction	This paper introduce a latent interaction index and examine its impact on the formation, development, and stability of complex networks. To overcome the limitations of traditional compositional similarity indices, particularly when dealing with large networks comprising numerous nodes, this index takes into account both observed and unobserved heterogeneity per node. In this way, it effectively captures specific information about the participating nodes, while mitigating the reflection effect, a common problem in estimation methods based on network structures. We develop a Shannon-type entropy function to characterize the density of networks and establish optimal bounds for this estimation by leveraging the network topology. Additionally, we demonstrate some asymptotic properties of pointwise estimation using this function. Through this approach, we analyze the compositional structural dynamics, providing valuable insights into the complex interactions within the network. Our proposed method offers a promising tool for studying and understanding the intricate relationships within complex networks and their implications under parameter specification. We perform simulations and comparisons with the formation of Erdős-Rényi and Brabási-Albert type networks and Erdos-Renyi and Shannon-type entropy.
Alba Martinez Ruiz	Multidimensional Perspectives: A Patent Data Set for Analyzing Technological Development	Technological development is an unobserved variable in a multidimensional model with latent variables such as science development and geographical conditions. Estimating the holistic model is still an unresolved issue, and the quality and reliability of data are essential for obtaining rigorous results. This work presents a characterization of the utility patents granted in the United States of America and published between 1976 and October 10, 2012. The research has several objectives. First, we provide an overview of the complete data set in terms of patent indicators as a basis for further research in the field of technological development. Second, we describe the technical and computational complexities of retrieving, processing, and storing the patents full text in XML and ASCII format. Nowadays, advances in computer science allow faster processing of data. However, the treatment of patent data is a complex task, and it requires a deep understanding of the meaning of each piece of information. Third, we report the statistical description of patent indicators computed for granted patents applied for USA companies. This includes the missing value treatment, and the description per application year and International Patent Classification section.

BNP: Nicanor Parra Library
FSSH: Faculty of Social Sciences and History
FE: Faculty of Education
FC: Faculty of Communications
FP: Faculty of Psychology