Workshop on Data Science at the Intersection of Agriculture with its Affiliated Disciplines

Friday April 22, 2022; 8:30am-5:20pm

Texas A&M, Rudder Tower, Rm 701 and online

Program (updated 04/21/2022)

8:30-8:40am **Welcome and Introduction**

Nick Duffield (TAMIDS/ECE) & Rudy Nayga (AGEC)

8:40-10:20am **Session 1: Marketing, Forecasting, Choice, Statistics**

8:40-9:10am Keynote: Venky Shankar (MAYS), *Empirical application of autoencoder models in marketing.*

With artificial intelligence permeating conversations and marketing interactions through digital technologies and media, machine learning models, and in particular, natural language processing (NLP) models, have surged in popularity for analyzing unstructured data in marketing. We introduce the latest neural autoencoder NLP models, demonstrate these models to analyze new product announcements and news articles, and provide an empirical comparison of the different autoencoder models along with the statistical NLP models. We discuss the insights from the comparison and offer guidelines for researchers. We outline future extensions of NLP models in marketing.

9:10-9:30am Ben Klopack (ECON), *One size fits all? The value of standardized retail chains.*

Multi-outlet firms, or chains, make up a large and growing part of the US retail sector and are the subject of important local policy; over 30 US cities ban or restrict the entry of chain firms. This paper quantifies the welfare and profit effects of standardized chains: chains face higher demand than independent firms, in part because of economies of scale in branding and advertising, but at the same time chains are less flexible in customizing product selection or prices across locations.

9:30-9:50am Jackson Bunting (ECON), *Continuous permanent unobserved heterogeneity in dynamic discrete choice models.*

In dynamic discrete choice (DDC) analysis, it is common to use finite mixture models to control for unobserved heterogeneity -- that is, by assuming there is a finite number of agent `types'. However, consistent estimation typically requires both a priori knowledge of the number of agent types and a high-level injectivity condition that is difficult to verify. This paper provides low-level conditions for identification of continuous permanent unobserved heterogeneity in dynamic discrete choice (DDC) models. The results apply to both finite- and infinite-horizon DDC models, do not require a full support assumption, nor a large panel, and place no parametric restriction on the distribution of unobserved heterogeneity. Furthermore, I present a seminonparametric estimator that is computationally attractive and can be implemented using familiar parametric methods. Finally, in an empirical application, I apply this estimator to the labor force participation model of Altug and Miller (1998). In this model, permanent unobserved heterogeneity may be interpreted as individual-specific labor productivity, and my results imply that the distribution of labor productivity can be estimated from the participation model.

9:50-10:10am Yang Ni (STAT) & Bani Mallick (STAT), *Ordinal causal discovery.*

Existing causal discovery methods for categorical data focus on inferring the equivalence classes, which leaves some causal relationships undetermined. This talk presents a new method that exploits the ordinal information contained in many real-world applications to uniquely identify the causal structure. The proposed method is applicable beyond ordinal data via data discretization. An R package OCD is freely available at https://web.stat.tamu.edu/~yni/files/OCD_0.1.0.tar.gz

10:10-10:20am Discussion (lead: Venky Shankar, MAYS)

10:20-10:30am **Break**

10:30-11:40am **Session 2: Agriculture, Nutrition, and Health**

10:30-11:00am Keynote: Patrick Stover (IHA), *Institute for Advancing Health through Agriculture*.

Historically, agriculture and food systems were designed to produce food, fiber, and fuel in abundance to ensure agricultural products were plentiful, affordable, and accessible. Presently, agriculture and food systems have additional expectations with respect to supporting human health and lowering health care costs, protecting the environment, and ensuring producer profitability. As highlighted by the 2015 National Academies Report[1]: A Framework for Assessing Effects of the Food System,

achieving these outcomes will require knowledge of the complex, dynamic systems associated with food, human health, the environment and the economy, and their interactions. Next generation data science applications will be essential to position agriculture as the solution to human health, environmental health and economic health. [1]A framework for assessing effects of the food system. Washington, DC: The National Academies Press. National Academies of Sciences, Engineering, and Medicine. 2015. https://doi.org/10.17226/18846.

11:00-11:30am Keynote: Kathy Baylis (UC Santa Barbara), *Machine Learning in Agricultural and Resource Economics*

With the substantial growth in novel data sources and computational power, machine learning holds great potential for economic analysis. However, like any new approach, the strengths and weaknesses of these tools need to be considered when deciding where and how they can be successfully applied. In this talk, I explore the potential of ML to fill gaps in our current methodological toolbox. I discuss use cases like the need for flexible functional forms, the use of unstructured data, and large numbers of explanatory variables in both prediction and causal analysis. I highlight these issues drawing from existing examples in agricultural and applied economics. To unpack the 'black box' of ML, I present numerous approaches used in computer science and statistics for model interpretability. I argue that economists can play a vital role in adapting ML methods for the use in economics by combining them with our domain knowledge of economic mechanisms, and our approach to causal identification.

11:30-11:40am Discussion (lead: Yong Liu, AGEC)

11:40-11:50am Break

11:50am-12:10pm Panel Discussion: Industry Perspectives on Data Science

Sihong Chen (Amazon), Guo Chris Cheng (STATA), Chen Gao (Meta)

Moderator: Sam Priestly (AGEC)

12:30-1:30pm Lunch (provided for in person registrants)

1:30-2:50pm Session 3: Policy & Sustainability

1:30-2:00pm David Viviano (UC San Diego), *Policy design in experiments with unknown interference.*

In this talk, I will discuss the problem of experimental design for estimation and inference on welfare-maximizing policies in the presence of spillover effects. As a first contribution, I introduce a single-wave experiment that estimates the marginal effect of a change in treatment probabilities, taking spillover effects into account. Using the marginal effect, I propose a practical test for policy optimality. The idea is that researchers should report the marginal effect and test for policy optimality: the marginal effect indicates the direction for a welfare improvement, and the test provides evidence on whether it is worth conducting additional experiments to estimate a welfare-improving treatment allocation. As a second contribution, I design a multiple-wave experiment to estimate treatment assignment rules and maximize welfare and derive guarantees on the proposed procedure. I illustrate the benefits of the method in simulations calibrated to existing experiments on information diffusion and cash-transfer programs.

2:00-2:30pm Luis Ribera (AGEC) & Juan Landivar (AgriLife CC), *Data-driven In-season management and yield forecasting of cotton for sustainable production.*

This project establishes a transdisciplinary team of scientists with expertise in remote sensing, agronomy, geomatics, civil engineering, big data analytics, precision agriculture, crop physiology, socioeconomics, and extension. The integration of multiple disciplines is expected to improve our understanding of the big data, the chance to understand stakeholders needs, and use the combined knowledge to develop tools for near time in-season management decisions. The availability of big data obtained from multiple remote sensors at high spatial and temporal scale makes it possible to develop data-driven models to forecast future plant growth, in-season management plans, yield, and marketing. Moreover, currently the team is working on measurements and forecast of cotton quality to improve crop marketing strategies.

2:30-2:40pm Discussion (lead: Grace Melo, AGEC)

2:40-3:00pm Break

3:00-4:10pm Session 4: Animal Data Science

3:00-3:20pm Louis Tedeschi, Karun Kaniyamattam, Egleu Mendes (ANSC), *Data-driven precision livestock farming for sustainable beef production in Texas.*

Over the last three decades, our laboratory has developed and implemented different decision support systems for beef cattle, focusing on precision livestock farming (PLF), such as the ruminant nutrition system, beef cattle nutritional requirements model, and cattle value discovery system (nutritionmodels.com). PLF is relevant to many fields of livestock production, including measurement of forage, monitoring animal location, behavior, early detection of diseases, milk composition, reproductive measurements, feed intake, and greenhouse gas emissions. There are many possibilities to improve animal production through PLF, but the combination of PLF with computer modeling is necessary to facilitate on-farm applicability. Different artificial intelligence (AI) based techniques are being used in beef systems to evaluate the data from IoT sensors. However, the domain parameter relationships provided through the AI infrastructure need to be integrated with existing decision support systems, using a hybrid approach. The current need is a data ecosystem that merge the data, expertise, and knowledge from ranches, feedlots, and retail logistics-based stakeholders. This data ecosystem requires a transdisciplinary collaborative approach between soil, plant, animal, veterinary, engineering, and social science expertise available at Texas A&M University, in order to implement a data-driven decision-making pipeline required for sustainable beef production in Texas.

3:20pm-3:40pm Sushil Paudyal (ANSC), *Churning data from the dairy farm.*

Modern dairy farms are increasingly adopting precision dairy cattle monitoring technologies to manage day-to-day activities of dairy cows. These sensor devices measure behavioral parameters (like fitbits for humans, e.g., rumination, activity, lying time, steps, lying bouts) and are increasingly used on farms to detect diseases and to breed cows. As dairy cows go through the milking process at least twice a day, these sensor devices coupled with milking robots produce millions of data points from dairy cows every day. In addition, the milk being produced by the cows in every milking is also being evaluated on-farm for fat, protein, and lactose content in addition to other milk quality parameters like somatic cells and electrical conductivity by the automatic milking robots. These wide varieties of variables being measured on a single dairy cow every single minute make the dairy farm a dynamic data manufacturing machine. However, because of the incompatibility of data sources, these data are stored and utilized on-farm as disjoint silos and have not

been analyzed and used to their full potential. There are opportunities for collaboration between data scientists and dairy scientists by utilizing big data tools to evaluate the dairy farm data to improve milk production and detect disease events in dairy farms.

3:40-4:00pm Yalong Pi (TAMIDS) & Egleu Mendes (ANSC), *Convolutional neural network implementations for beef cattle precision livestock farming.*

Precision livestock farming (PLF) utilizes technological advancements in sensors, data collection, and statistical tools to improve farming management. Traditionally, PLF data such as animal feeding quantity and quality, body sizes and weights, movements, and early illness symptoms are measured by a physical sensor or a trained operator, which are resource-intensive processes. Computer vision systems (CVS) have been applied as a low-cost and non-invasive diagnostic tool in research and commercial software for PLF. However, the traditional CVS techniques (e.g., engineered shape) are difficult to generalize due to the diverse backgrounds, animal features (color and pose), and camera placements. The emerging convolutional neural network (CNN) is a more generalizable technique, and it has outperformed humans in image recognition tasks. The collaboration between TAMIDS and the Department of Animal Science aims to investigate the fundamental aspects of implementing CNN in PLF, including datasets, model architecture, camera parameters, and limitations. The available information categories from CNN (e.g., animal activities) will be established by reviewing literature and experimenting with video samples from an experimental animal feed station at Texas A&M. New CNN models will be designed and trained considering the unique output needs of PLF, which are different from general CVS tasks. The domain experts and physical sensors will be involved in generating a diverse cattle dataset.

4:00-4:10pm Discussion (lead: Yvette Yu Zhang, AGEC)

4:10-4:15pm Break

4:15-5:00pm Session 5: Food Insecurity, Education

4:15-4:35pm Senarath Dharmasena & David Bessler (AGEC), *Endogeneity and Instrumental Variables in a Complex Economic System: Graph Theoretic Approach.*

Multitude of variables and their interactions form complex systems. When faced with such systems, one has to carefully identify variables and interactions among those variables to make prudent predictions about decision variables that are important for policy making. Traditional models are heavily dependent on assumptions in identifying variables affecting principal decision variables and their interactions with co-variates around them. These assumptions could come from various sources, such as, prior theory, researchers' subjective assessments and piecemeal-wise identification of variables found in the likelihood function which governs the stochastic characteristics of the observed variables. In this study, we investigate methods to identify endogenous and instrumental variables in a complex economic system using cutting-edge graph-theoretic approaches implemented via machine learning algorithms applied to variables governing the U.S. food environment system where many variables interact in complex ways. Specific objective is to identify instrumental and endogenous variables among a set of variables affecting the food environment using Instrumental Variable Discovery Algorithm (IVDA) of Phiromswad and Hoover (2013). We use the data on sixteen variables associated with the U.S. food environment. Preliminary results show that obesity and food insecurity variables are strictly endogenous while income is exogenous. Income acts as an instrument for poverty.

4:35-4:55pm Grace Melo (AGEC), Pourya Valizadeh (AGEC) & Rudy Nayga (AGEC).

The role of SNAP in the well-being of low-income households with children post COVID-19 onset.

To confront the adverse well-being effects of the COVID-19 pandemic (e.g., decline in economic activity, a spike in unemployment rates, and an increase in food prices), the US government authorized several temporary changes to the Supplemental Nutrition Assistance Program (SNAP) to ensure the provision of adequate resources to struggling families, particularly low-income families with children, beginning April 2020. These changes included: providing participants with emergency allotments of at least $95 per month, a 15% increase in the maximum SNAP benefits, replacing the missing free or reduced-price school meals due to school closures with benefits through a new SNAP program, called Pandemic-EBT (P-EBT), and expanding the SNAP Online Purchasing Pilot Program (OPPP). To our knowledge, the distinct role of participation in SNAP in determining childrens' food insecurity and health outcomes after the start of the COVID-19 pandemic has not been studied. The objective of this project is to examine the role of SNAP on the well-being measures and food insufficiency and physical/mental health status of low-income households with children in pre-and post-onset of the COVID-19 pandemic (when temporary changes in SNAP occurred).

4:55-5:15pm Geoffrey Pofahl (Arizona State University), *What is missing from Data Science education? A perspective based on 10 years of Data Science leadership.*

The continued development and growth of education programs focused on Business Analytics and Data Science is a welcome one. However, based on 10 years of leadership in the field, I continue to see a handful of gaping holes in Data Science training. What are those gaps? Why should we be doing more to fill them?

5:15-5:20pm Discussion (lead: Senarath Dharmasena, AGEC)

5:20pm Conclusion