MEMORANDUM

DATE: October 22, 2020

TO: Faculty and Students

FROM: Professor(s) Robert McCulloch
Chair/Co-Chairs of Xuetao Lu
Defense for the PhD in Statistics
Committee Members Paul Hahn
Shiwei Lan
Shuang Zhou
Steven Saul

DEFENSE ANNOUNCEMENT

Candidate: Xuetao Lu
Defense Date: 11/06/2020
Defense Time: 3:00 PM
Virtual Meeting Link: https://asu.zoom.us/j/2210701708
Title: Spatial Regression and Gaussian Process BART

Please share this information with colleagues and other students, especially those studying in similar fields. Faculty and students are encouraged to attend. The defending candidate will give a 40 minute talk, after which the committee members will ask questions. There may be time for questions from those in attendance. But, guests are primarily invited to attend as observers and will be excused when the committee begins its deliberations or if the committee wishes to question the candidate privately.

ABSTRACT

-See next page-
Spatial regression is one of the central topics in spatial statistics. Based on the goals, interpretation or prediction, spatial regression models can be classified into two categories, linear mixed regression models and nonlinear regression models. This dissertation explored their real world applications. New methods and models were proposed to overcome the challenges in practice. There are three major parts.

In the first part, nonlinear regression models were embedded into a multistage workflow to predict the spatial abundance distribution of reef fish species in the Gulf of Mexico. There were two challenges in this application, zero-inflated data and out of sample prediction. The methods/models in the workflow could effectively handle the zero-inflated sampling data without strong assumptions. Three strategies were proposed to solve the out of sample prediction problem. The discussion of results showed that the nonlinear prediction has the advantages of high accuracy, low bias and well-performed in multi-resolution.

The question in the second part was asking to develop a spatial regression model for analyzing soil carbon stock (SOC) data. Different from the first application, the desired model should perform well in both prediction and interpretation. Unfortunately, as well known, there is a trade-off between the two goals. A two-stage model was proposed to break this trade-off. In the first stage, there is a spatial linear mixed model that captures the linear and stationary effects. In the second stage, a generalized additive model was used to explain the nonlinear and nonstationary effects. The results illustrated that the two-stage model has good interpretability in understanding the effect of covariates, meanwhile, it keeps high level prediction accuracy which is competitive to the popular machine learning models, like random forest, xgboost and support vector machine.

A new nonlinear regression model, Gaussian process BART (Bayesian additive
regression tree), was proposed in the third part of this dissertation. The model combined the advantages of the BART and Gaussian process to model the nonlinear effects of both observed and latent covariates. To implement this model, first, we generalized the traditional BART to accommodate the correlated data. Then the failure of likelihood based MCMC in parameter estimation was discussed. Based on the idea of analysis of variation, the methods, back comparing and tuning range, were proposed to tackle this issue. Finally, the effectiveness of the new model was examined with both one dimension and two dimensions (spatial) data.