Variable Selection in Point Pattern Modeling

Maurits de Graaf, Marie-Colette van Lieshout, Changqing Lu

June 2022

Variable selection is an important but also difficult part in point pattern modeling. From the data science perspective, a efficient detection of explanatory variables can help build well-performing spatio-temporal models; from the statistical perspective, suffcient theory has to be established to explain variable selection results.

In the literature, there already exist many studies on this topic. Classic approaches for variable selection can be divided into two categories: statistical and machine learning methods. Statistical methods are often parametric, and include testing statistics (e.g., p-values, information criteria or specialized statistics for point pattern modeling) and regularized penalty functions. They usually have solid theory (e.g., asymptotics), however, behave unstable and even fail when the number of variables is large. Machine learning methods are mostly non-parametric, such as permutation importance of random forests and sparse learning of neural networks. They have been shown to successfully select significant variables from a large number of candidates and can thus obtain distinguished model performance based on the selections.

Back to the real world case, in a recent point pattern modeling study on chimney fires, we used the random forest methods to select the important variables and proposed a well-performing prediction model. However, in an extension of this modeling procedure to other fire types, we obtained confusing suggestions on variable selection. Moreover, the selection results is not convincing after a consultation with fire experts.

The goal of this master assignment is to study the variable selection methods for point pattern modeling. The possible research questions are: i) how do statistical and machine learning methods perform in different cases? ii) what can be a good variable selection method in data-driven point pattern modeling? iii) To what extent the machine learning methods can be explainable? The project has joint relations with data science and statistics. Further fire data can be collected and used to validate the utility of the proposed method or theory.