The Effectiveness of Place for Spatial Big Data Analytics

A position paper by May Yuan

The proliferation of various sources of Spatial Big Data has challenged the traditional thinking of spatial statistics in finding meaningful patterns with strong conceptual and theoretical foundations. Frequency analysis is dominant in Big Data Analytics due to its simplicity. Likewise, the most common approach is kernel density mapping to identify hot spots or, furthermore, changes of hot spots over time. Nevertheless, most kernel density maps on Big Data, especially social media data, are overpowered by the so-called "bull's eyes" which indicate strong local distributions likely unfit for creating a density surface. While google flu forecasts were shown to be over-estimated, kernel density maps are likely subject to similar issues. In addition, the issues of false negatives would be more prominent in Spatial Big Data Analytics then general Big Data Analytics because social media behavior may be better (or easier) approximated by temporal interpolation than spatial interpolation.

The position paper posits that Spatial Big Data Analytics may consider to adopt the idea of "place" as the spatial unit of analysis for two reasons: (1) while there is no limit of how fast Big Data has been growing and how enormous Big Data has become, the number of places of interest is likely to be limited. Hence, places can be used as an effective organizer of and make Spatial Big Data manageable; (2) by identifying and learning about places based on known Spatial Big Data, the knowledge learned (such as correlates, associations, co-location, co-actions, etc.), the knowledge is likely better suited for spatial interpolation to places where have records of Spatial Big Data but exhibit the learned characteristics. Knowledge of what have happened to a place over time is essential to understand the potential attractors of the place to the happening (i.e. tweets on floods), and understanding place characteristics associated with the potential attractors, moreover, can help identify other potential places that things of interest could have happened but "Big Data" are lacking.

Using 2009-2011 crime incidents data from Tulsa, Oklahoma, we developed a method to identify "places" based on experiences of crime incidents over time. The data set include a total of 183,101 crime incidents classified into 12 crime types, including murder, aggravated assault, forcible rape, burglary, robbery, larceny from building, larceny from vehicle, motor vehicle theft, shoplifting, drugs, part II crime, and all other crime. Conventional spatial criminology analyses would either evaluate one crime type at a time or consider occurrences regardless of crime types. In contrary, we consider all crime types simultaneously to explore the following questions: How to identify places that are defined by crime occurrences? How to summarize crime experiences at these places? How to reveal site characteristics and situation characteristics of these places in relation to their unique crime experiences?

We first used random grids to identify places that experienced crimes over the study period. Then we applied Kohnen self organizing map algorithm to reveal categories of these places. Each category corresponded to a distinct pattern of crime experiences. We then applied life course analysis to elicit common crime-type sequences at places, which would summarize something like rubbery taking place commonly after multiple occurrences of larceny from building. We are also working to derive the likelihood that a certain crime type would go after another particular crime type at a place, and how the likelihood varies across different categories of crime places. While our experiment is on crime data, the methodology ought to be useful for other types of spatial events from social media since the method does not assume statistical distributions of the event of interest.