Proposal: KDE in the Congo

Jessica Moore

April 29, 2016

The Democratic Republic of the Congo (DRC) has a history of violent conflict dating back at least to its colonization by Belgium in the late 1800's (Rosen 2013). The colonial rule imposed by the Belgians was notoriously exploitative and left the country largely in ruins (Hochschild 1998). When the Congolese finally gained independence in the 1960's, the country had no government to speak of and, in spite of the DRC's plentiful natural resources, economic underdevelopment reigned (Rosen 2013). While progress has been made in the past half-century, the rampant poverty and lack of government oversight resulted in high levels of violence. Conflict in the country has killed more than 3.5 million individuals since 1996 (Rosen 2013). In 1998, a rebellion against the already weak Congolese government began in the Kiva region of the DRC, which lies on the country's east, near its borders with Rwanda and Burundi ("MONUSCO Background"). The United Nations (UN) brokered a ceasefire between the government and the rebels in 1999 and has maintained a substantive presence in the country ever since ("MONUSCO Background"). Currently, the UN oversees a force of nearly 20,000 military personnel in the DRC; the force attempts to quell violent uprisings and mitigate potential sources of conflict ("MONUSCO Facts and Figures"). However, in 2014 the UN Security Council identified the need for an exit strategy, the goal of which would be to remove UN forces from the country and transition military and police duties to the Congolese government ("MONUSCO Mandate"). In order to implement such a plan, the UN will need to gradually remove its troops, concentrating its remaining personnel in the most conflict-prone regions. This, however, is no small task, since the UN must first identify (with relative specificity) the areas of the country in which additional conflict is particularly likely.

We propose a paper evaluating kernel density estimation as a method of predicting the locations of future conflict in the DRC. We will, in essence, answer the question of whether KDE results in useful predictive models in this particular instance and, if so, which method of selecting tuning parameters appears to be optimal. If the models produced by this method are effective predictors of conflict, using them during the planning process would allow a more strategic draw-down by UN forces (i.e., first removing troops from the locations in which they are least necessary) and, as a result, a smoother transition from oversight by the UN to control by the Congolese government. This proposal proceeds in three sections. In the first section, we provide a brief overview of kernel density estimation and discuss the utility of the procedure in this particular context. That section serves to motivate our use of KDE above any other method, and specifically above parametric procedures. Next, we provide information about the data we intend to use for the paper. We review the data's origins and suggest certain subsets of the data set that may prove most interesting. In the same section, we describe the analyses that we intend to perform on this data. We indicate both how the density estimates will be built and how they will be evaluated. Finally, in the last section, we identify the tasks that will be necessary to progress towards the full paper. While we highly doubt that any upper-level UN policymaker is likely to read the proposed paper, we do hope that it may contribute to a larger discussion regarding strategies for smooth transitions of control from NGOs to local government forces in developing countries.

1 WHY USE KDE?

Kernel density estimation is a common nonparametric method for estimating the probability density function of a random variable based on observed realizations of that variable (Cai 2013). The technique is particularly useful when the data-generating distribution takes a complex form, because KDE imposes significantly fewer assumptions about the distribution that underlies the observed data than do most parametric procedures (Silverman 1986). Rather than beginning with the belief that the data take a specific distribution and estimating the parameters of the distribution based on the data, KDE allows the data to "speak for themselves" (Silverman 1986). Kernel density estimation attempts to represent the probability distribution from which a data set was generated by placing a small symmetric "bump" in probability density at every data point (Cai 2013). The bumps are scaled down and added together. The resultant height at any location is proportional to the probability assigned to that spot, that is, the predicted likelihood of future events occurring there (Cai 2013).

In implementing KDE, the statistician's primary job is to appropriately tune the form of the "bumps." We intend to use Gaussian kernels (far and away, the most common choice in practice), which will result in "bumps" that look like bivariate normal distributions centered at each location of previous conflict. We specifically select Gaussian kernels for their computational and mathematical tractability, as well as the fact that Gaussian kernels are the default in nearly all softwares for implementing KDE. Of significantly greater importance than our kernel will be the "bandwidth" of the kernel density estimate. The size of the bandwidth is directly related to the spread of the "bumps." It is commonly accepted that bandwidth selection is one of the most important and difficult components of creating a useful kernel density estimate (Cai 2013). The proposed paper would investigate a variety of bandwidth selection methods (plug-in methods, least-squares cross-validation, biased cross-validation, etc.), and evaluate how well each performs on the data set.

Kernel density estimation is a particularly useful technique in the proposed context for a few reasons. First, given the current instability in the DRC, the data required for a traditional parametric model are uniquely difficult to obtain. This is especially true given the detailed level at which one would need the data in order to make precise predictions. A conventional parametric approach would likely involve dividing the country into regions and obtaining data regarding both characteristics of those regions and the number of conflicts that occurred there during a specified time period. A parametric strategy would then fit a model to estimate how each of those characteristics is related to the amount of conflict a region experiences. Such a model might require data about public services, income, or demographics in a particular area of the country. However, it is, quite simply, easier to obtain data on previous instances of violence than it is to obtain data on local political, demographic, and economic indicators; the former are relatively well-publicized, whereas the latter are unlikely to even exist. Moreover, even if one were able to obtain the data needed to make accurate predictions, the high degree of interaction between variables would make a parametric model difficult to specify. Spatial data, such as the data set we intend to use, are unlikely to have easily-defined relationships between predictor variables and the outcome variable (Brunsdon and Comber 2015). KDE, because it simply reflects the data that it takes as input, has the ability to capture highly nonlinear and interactive relationships between variables, which might otherwise go unaccounted for in a parametric model.

Additionally, kernel density estimation brings with it a convenient visual depiction in the form of a heat-map. A heat map (sometimes referred to as a "hot-spot map") is an image that represents values in different areas of the map using different colors. E.g., areas that are colored red might have higher values and areas colored blue might have lower values. Heat maps based on kernel density estimates are often used in crime prediction, because they allow law enforcement agents to easily assess where crimes have historically occurred (Gerber 2014). More generally, heat maps provide a concise way of conveying information to an individual who is not familiar with the underlying statistics. Because, ideally, our analysis would be utilized by individuals who are subject-matter experts (in African politics, conflict resolution, etc.) rather than statisticians, it is uniquely helpful to have on hand a comprehensible visualization that corresponds to our predictive model.

2 Data & Analysis

The proposed paper would use data from the Armed Conflict Location and Event Data Project (ACLED), which collects real-time data on political violence in both Africa and South/Southeast Asia ("About ACLED"). The proposed paper would employ ACLED's Africa data set, specifically, the subset of that data set related to conflict in the DRC. ACLED records key information (e.g., date, location, actors, number of fatalities, etc.) about instances of conflict or events that may be precursors to it (e.g., the establishment of rebel bases). ACLED aggregates data from many sources, including local media and reports by nongovernmental organizations ("About ACLED"). Events recorded in ACLED's data range from nonviolent territory transfers to declared wars. The proposed paper would focus on the three most prevalent event categories out of those that appear to indicate true instances of conflict (in contrast to events such as rebel base establishment, which are not immediately violent). These categories are "Battle," "Violence against civilians," and "Riots/Protests."[1] While ACLED maintains data on events from 1997 onward, we intend to use only those observations from 2014 and 2015. We anticipate that using more recent data will result in a more accurate assessment of whether our models might be useful in predicting future events. Additionally, we intend to use aggregated at an annual level, because we do not believe individual months contain a enough data points to produce a viable density estimate.

We intend to build a variety of density estimates using 2014 data and evaluate them based on how well they predict the locations of conflicts that occurred in 2015. We plan to create models to predict conflict writ large as well as more specific models to predict only those conflicts falling into the categories described previously ("Battle," "Violence against civilians," "Riots/Protests"). In creating these predictors we also plan to test a range of different bandwidth selection methods in order to evaluate which of them tends to be most effective. We will use surveillance plots to compare the different models that these bandwidths produce. A surveillance plot shows the percentage of true events (here, 2015 conflicts) that occurred in the y% most-threatened area of the country, based on the model's prediction (Gerber 2014). Better models result in surveillance plot lines that lie closer to the upper-left corner of the plot area (which is a 1-by-1 square), because this indicates that the prediction captures true events relatively quickly (Gerber 2014). In addition to the surveillance plots, we will also produce single-number summaries for each of the models, which correspond to the area under the curve (AUC) in the surveillance plot. Per usual, higher values of AUC indicate better models. We believe that having a single value by which to judge each model will make model comparison relatively straightforward. We hope, at that point, to identify those models which might be helpful as the UN makes choices regarding where to station its military personnel. Our analysis will rely on historical data, for the sake of being able to test our predictions. However, if our method results in effective models, we would suggest that policymakers employ a similar strategy going forward, e.g., using the prior year's data to make predictions for the following year and placing troops in those areas with the highest predicted probability of conflict.

3 Next Steps

In order to progress towards the final paper, a significant amount of research and data analysis are still necessary. First, we will need to further explore the literature on kernel density estimation. While it is reasonably clear how univariate KDE operates, the multivariate extensions of the process will require additional investigation on our part. Another component of our research will involve gaining a better understanding of bandwidth selection methods, including their derivations and the motivations behind using any particular one of them. Moreover, we will need to begin the process of cleaning and understanding the data set that ACLED has compiled. Most significantly, this will involve the standardization of important fields, such as event category indicators. This undertaking will also involve checking for and handling missing data and inaccurate records (e.g., events listed as occurring in the DRC which, in fact, took place in another country). Finally, we will begin to build our kernel density estimates and create predictions regarding which regions of the country are most conflict-prone. Then we can evaluate the predictive power of each model and make recommendations as to which methods (if any) may be useful going-forward.

[1] The "battle" category is composed the event types "Battle-Government regains territory," "Battle-No change of territory," and "Battle-Non-state actor overtakes territory."

4 References

"About ACLED." ACLED. Web. 23 Apr. 2016.

Brunsdon, Chris, and Lex Comber. An Introduction to R for Spatial Analysis & Mapping. Los Angeles: Sage, 2015. Print.

Cai, Eric. "Exploratory Data Analysis: Kernel Density Estimation ? Conceptual Foundations." The Chemical Statistican, 9 June 2013. Web. 5 Apr. 2016. <https://chemicalstatistician.wordpress.com/2013/06/09/exploratory-data-analysis-kernel-density-estimation-in-r-on-ozone-pollution-data-in-new-york-and-ozonopolis/>.

Gerber, Matthew S. "Predicting Crime Using Twitter and Kernel Density Estimation." Decision Support Systems 61 (2014): 115-25. Web.

Hochschild, Adam. King Leopold's Ghost: A Story of Greed, Terror, and Heroism in Colonial Africa. Boston: Houghton Mifflin, 1998. Print.

"MONUSCO Mandate - United Nations Organization Stabilization Mission in the Democratic Republic of the Congo." UN News Center. UN. Web. 23 Apr. 2016.

"MONUSCO Facts and Figures - United Nations Organization Stabilization Mission in the Democratic Republic of the Congo." UN News Center. UN. Web. 23 Apr. 2016.

"MONUSCO Background - United Nations Organization Stabilization Mission in the Democratic Republic of the Congo." UN News Center. UN. Web. 23 Apr. 2016.

Rosen, Armin. "The Origins of War in the DRC." The Atlantic, 26 June 2013. Web. 20 Apr. 2016.

Silverman, B. W. Density Estimation for Statistics and Data Analysis. London: Chapman and Hall, 1986. Print.