

**Mercury Challenge Submission for Steven Erly**

1. **Introduction**
   1. Topcoder Handle: serly
   2. Marathon Match Placement/Award Received: CU Egypt Top Place, Milestone 1 Prize
   3. About You:

I work as an HIV epidemiologist for the Washington State Department of Health while pursuing a PhD at the University of Washington. I received my bachelor’s degree in neuroscience from the University of Arizona in 2014 and my master’s in public health from the University of Arizona in 2016 with a focus on epidemiology. Prior to my current role, I worked as a senior statistician for psychometrics and quality of life research in pharmaceutical development and as a statistical consultant in neurological research.

* 1. Why you participated in the Marathon Match?

I participated in this contest because I have a passion for using machine learning and statistical computing to improve the lives of those around me. I have had success in prediction contests in the past, and thought that the topics proposed by the Mercury Challenge were particularly interesting.

1. **Technical Solution**

My solutions drew on my experience in infectious-disease epidemiology and the recognition that the topics under consideration followed patterns similar to disease outbreaks: periods of low baseline activity broken by periodic spikes in counts. Given such a structure, I chose to model the events as a combination of two separate processes: a baseline rate and outbreak rate. The baseline rate was modelled using simple linear regression and the outbreaks were intended to be modelled via zero-inflated negative binomial regression. Given the short time period of the contest, it was determined that the overall probability of an outbreak for a given domain was low and the zero-inflated negative binomial regression was not needed.

The only data source used in my model was the gold standard reports released by the Mercury Challenge. The data was processed into a line list of events per day. Outbreaks were identified as any day where the number of events exceeded 2 standard deviations from the mean daily events over the time period represented in the dataset. The year, month, and day of each date were extracted into individual variables.

For all non-outbreak days, simple linear regression was used to model the number of events as a function of event date in order to capture long-term trends in the baseline rate. An attempt was made to capture periodic trends using day of the month and month of the year, but these models were found to have less predictive power than the simpler models, likely due to overfitting.

A linear model was fit at the beginning of period 2 (I did not participate in period 1.) A second model was fit halfway through the period to update the predictions based on the most recent results.

For the civil unrest and military action domains, the data was stratified by actor and region, and separate models were fit for each stratification. I felt that my models would perform best on actors where sufficient data was available, so I only submitted predictions for the most active few actors in each region. To accommodate the geographic component of the military action domain, I plotted the longitude and latitude of each event on a 2x2 grid and selected the approximate geographic centroid of the event locations for each actor. This value was imputed as the predicted location of future events.

A technical diagram can be found in the attached image.

1. **Feedback**

-The problem statement was well defined and was suitable to the competition. It attracted my interest and contributed to my decision to participate.

-The data was of high quality and presented pertinent information in an accessible and manageable format. The embedded links to the news articles were very useful.

-The scoring did not occur with enough regularity and was not accessible enough to be able to adequately assess and modify the performance of the models. The system was very complex and it was not always clear how our predictions contributed to our score. An online interface or API to examine daily predictions and resulting scores would have been invaluable and would have encouraged me to pursue more sophisticated models.