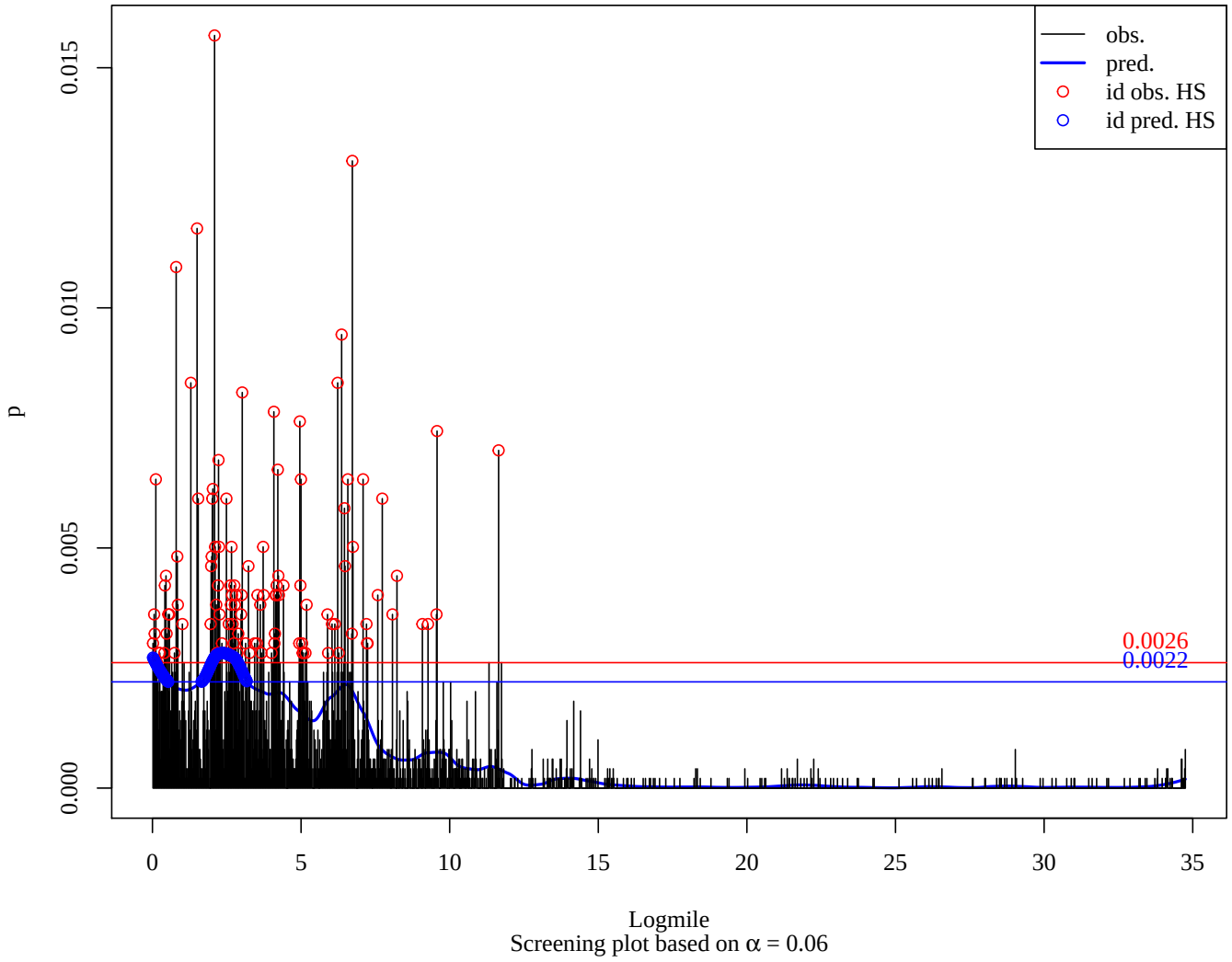
Level set setting

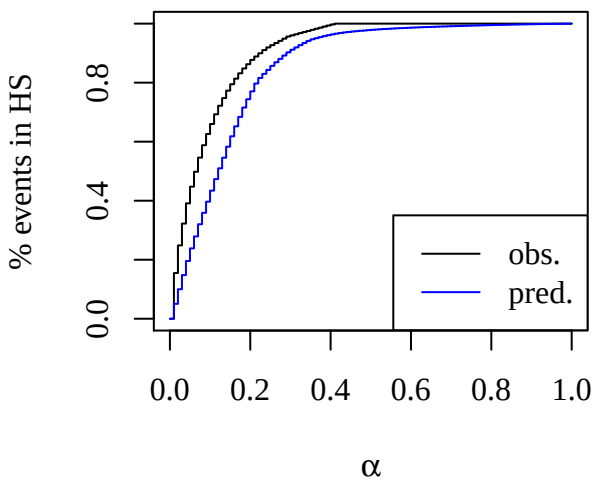
$\gamma_X(\alpha) = \sup_n\{X : \frac{\|C_X(n)\|}{N} \leq \alpha\}$, where n is the number of Logmile segments needed to satisfy the condition.

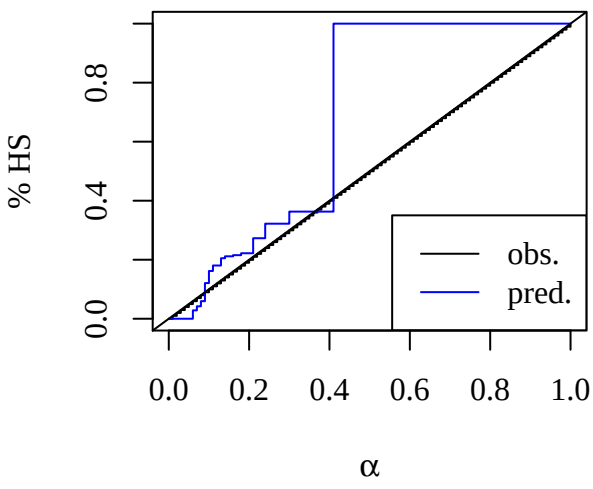
$C_X(n) = \{X : f_X(x) \geq f_{X_{(N-n+1)}}\}$, where $f_{X_{(N-n+1)}}$ is the $(N-n+1)^{th}$ ordered crash rate (from minimal to maximal crash rate).

$f_{X_{(N-n+1)}} = \phi$, where ϕ is the threshold that satisfies $\frac{\|C_X(n)\|}{N} \approx \alpha$.

Suppose $\alpha = 0.06$, we can find out the threshold for observations is $\phi_{obs} = 0.0026$, whereas the threshold for predictions is $\phi_{pred} = 0.0022$ and the proportions of blues and reds are approximately α respectively (as plot showed below).

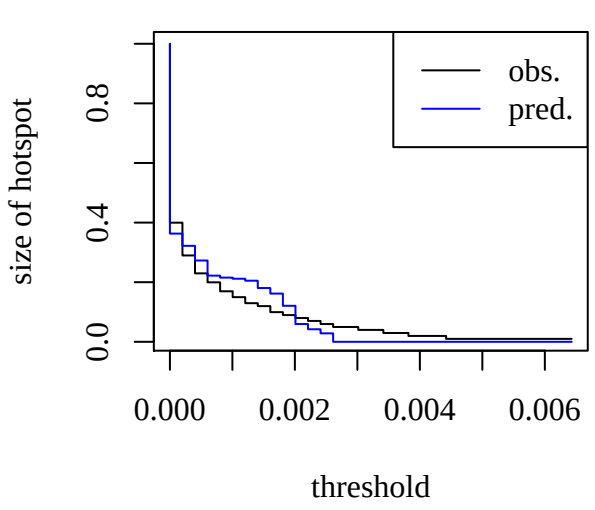


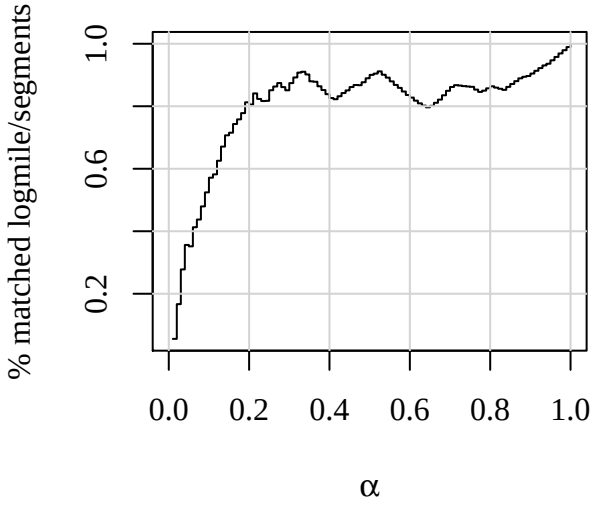




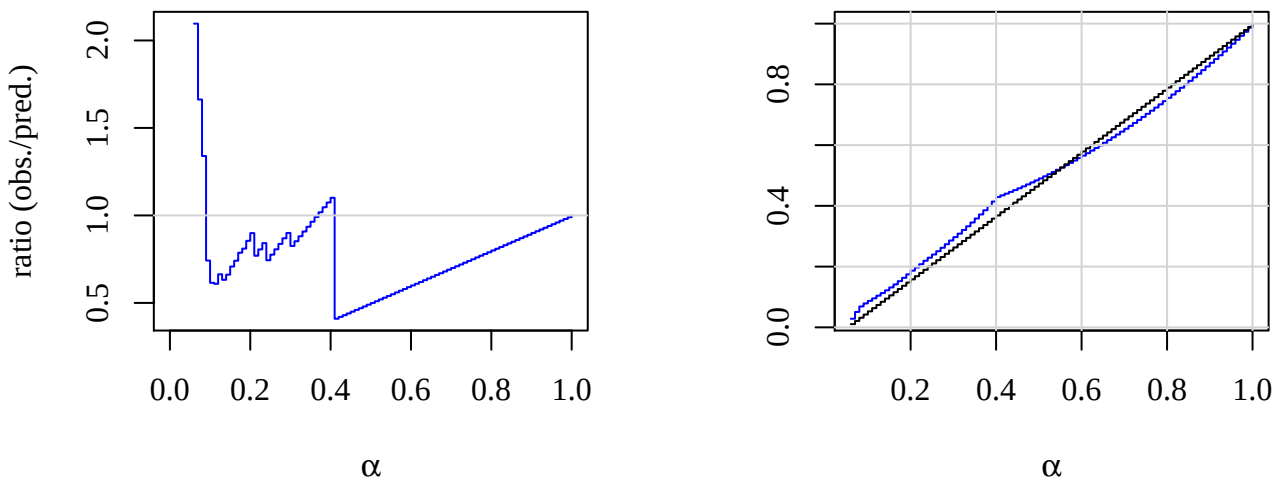
A sequence of α values is used to observe the pattern in different levels, $\alpha \in [0, 1]$ with width of interval 0.01. Y-axis is the cumulative capture crash rate in HS and X-axis is level of α. Suppose $\alpha = 0.1$, roughly 65% and 40% of crashes in observations and predictions respectively happened on the top 10% of Logmile segments (figure showed above left). Then, we observe the pattern under the same threshold for observations (figure showed above right). Since the proportion of observations above threshold is built based on α, it is similar to a straight line with slope of 1.

A quasi-KS test is performed in our surveillance plot. the statistic $\max_x |S_i(x) - S_j(x)|$, where $S_i(x)$ and $S_j(x)$ represent the CDFs for i and j repectively. And if the maximal gap between $S_i(x)$ and $S_j(x)$ is significant, then $S_i(x)$ is not the same distribution as $S_j(x)$. In here, we say our prediction does not fit observed data.





On the top right plot, it shows the percentage of matched prediction and observed segments under different levels of α. When top 20% of segments are identified as hotspots in both prediction and observed data, 80% of individual segments are matched.



relative distribution of predictions and observed data. Convert relative distribution to cumulative probability function (convert top left to top right).

$\lambda(x)$, intensity/rate for segment x represent observed data.

$\hat{\lambda}(x)$, intensity/rate for segment x represent predictions.

ratio $r = \frac{p(x:\lambda(x)\geq\phi_\alpha)}{p(x:\hat{\lambda}(x)\geq\phi_\alpha)}$

In prediction manner, if we expect our prediction is completely as same as observed data, the ratio r should always be 1 in the relative distribution plot, and be straight line with slope of 1 in CDF plot. To see the devation of our prediction and data, quasi-KS and chi-squared tests are performed and shown no significance exists.