Predictability of Crash Modelling From an Improved Quality of Data Using Level Sets Surveillance System

Abstract

1. Introduction

2. Evaluation Methodology

2.1 Level sets

Level set methods are widely used in image segmentations and processing. One of its applciations is for pattern recognitions and shape reconstructions. Instead of continuous and multidimentional space, same implementation of level set method is used as a measure of pattern recognitions for discrete space crashes analysis in this study.

Let X be highway segments, S_X be the support of X, $f(x)$ be the density of crashes at segment $X = x$, such that $\sum_{S_X} f(x) = 1$ (discrete spatial points). α is the top proportion of segments expected to be monitored. Then, the level set is defined as:

$\gamma = \{x \in \Omega | f(x) \geq c\}$

where γ is the set of segments x's that $f(x) \geq c$, c is a threshold satisfying $\inf_c \frac{\|\gamma\|}{\|S_X\|} \leq \alpha$ and c can be determined as $\min_{x \in \gamma} f(x)$.

Some drawbacks of using level set methods in discrete space data are such as approximation of desired levels and solution to the ties with the same rates. Considering these two main issues in this study, level $\hat{\alpha}$ of sample is contraint on the proportion of highway segments that is closer to desired α but not greater than α. Since same ranks represent that crashes are equally likely to happen to these tied segments, ties should be all inclusive or exclusive in our desired set as the level changes. These two issues are under control with two general approaches mentioned previously. Figure 2.1 showed below demonstrates a comparison of crashes given different time periods under a level set method with $\alpha = 0.05$ using highway data from Route 71, 2015. X-axis is the location (x = Logmile) on this route whereas y-axis represents proportion of crashes $f(x)$. Threshold, c, (grey line) is set to be the minimal crash rate in the desired set γ. The true α levels for weekdays and weekend data are 0.0481 and 0.0377 with thresholds 0.0032 and 0.0039 repectively. 88 highway segments out of 1729 (or with a proportion of 0.0481) on route 71 with a crash rate equal or greater than 0.0032 approximately meets the desired level $\alpha = 0.05$ for weekdays; similarily, 69 out of total 1729 segments on route 71 with a threshold equal or over 0.0039 satisfies this scenario. This setting is in a aspect of concertrating resources, say 0.05 in this example, and focusing on few highway segments with higher chance of having crashes with efficiency.

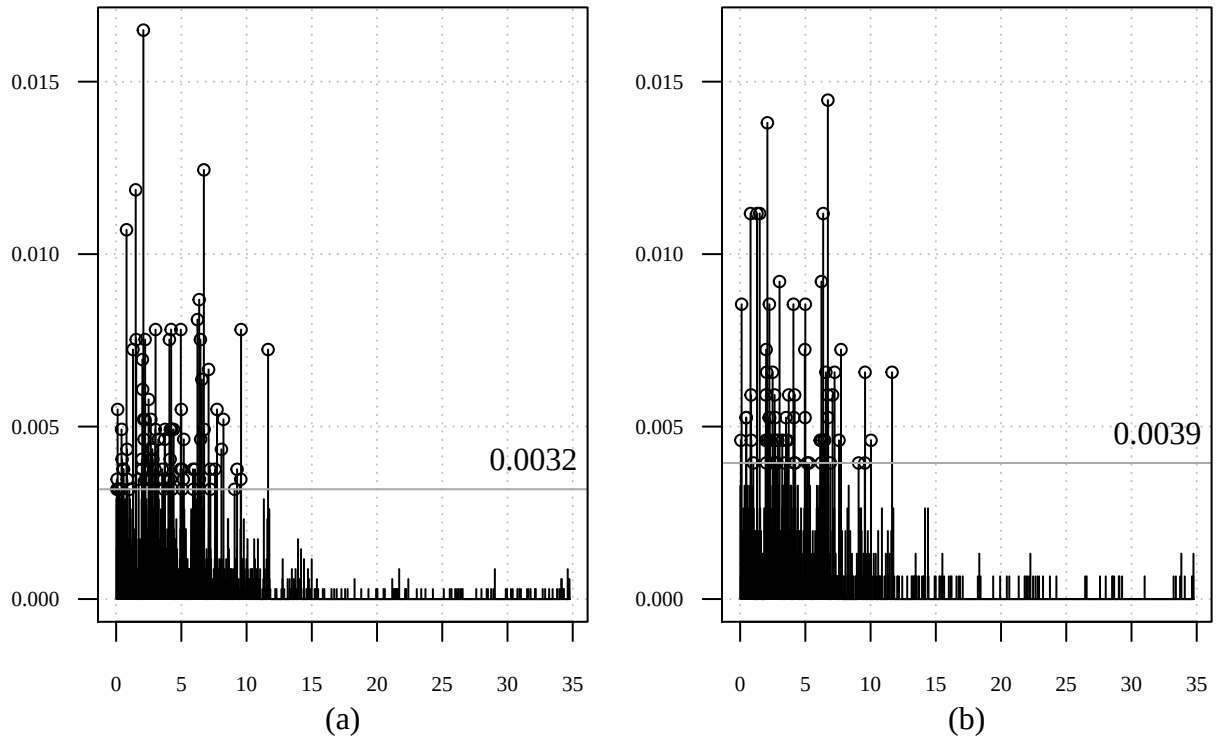

Figure 2.1: weekdays (a) and weekend (b) with $\alpha = 0.05$

https://link-springer-com.libdata.lib.ua.edu/book/10.1007%2F978-3-642-15352-5
https://link-springer-com.libdata.lib.ua.edu/book/10.1007%2F978-3-319-01712-9

2.2 Surveillance plots

Surveillance systems are built in respect of monitoring and capturing certain aberrations in the current or future process of tendency. These systems can be applied for detecting potential outbreak of epidemiological disease, and in essense used for evaluation of certain factors and profiles to acheive infection control in healthcare.

In conjuction with level set method, surveillance plots provide visualization of screening spatiotemporal highway data from distinct perspectives of monitoring crashes along highway segments. Under limited resources, say manpower, 10% or less of highway segments with the highest crash rates may be considered high priority and regularly kept under surveillance (fixed allocation based surveillance). Also, highway segments with a target percentage of crashes can also be determined and adjusted to capture different proportion of segments under various conditions (threshold based surveillance). Considering time influence, distributions of crashes can be expected either similiar or dissimilar. To observe and build up scenarios as described above, surveillance plots are essential to extend further exploratory and predictive analysis.

http://www.sciencedirect.com/science/article/pii/S187603411730206X
http://www.sciencedirect.com/science/article/pii/S2093791116300610

2.2.1 Fixed allocation based surveillance

The following is an example of fixed allocation based surveillance plots in two distinct way of presenting. Same data used for Figure 2.1 is also implemented in this example. Instead of single α level is observed in

Figure 2.1, a sequence of α levels (from 0 to 1 with a increment of 0.01) are now concerned and showed in Figure 2.2. For (a) and (b), y-axis represents a cumulative proportion of crashes, and x-axis individually represents cumulative ordered segments size and top proportion of segments expected to be monitored, which is α. The following paragraphs will describe this figures in detail.

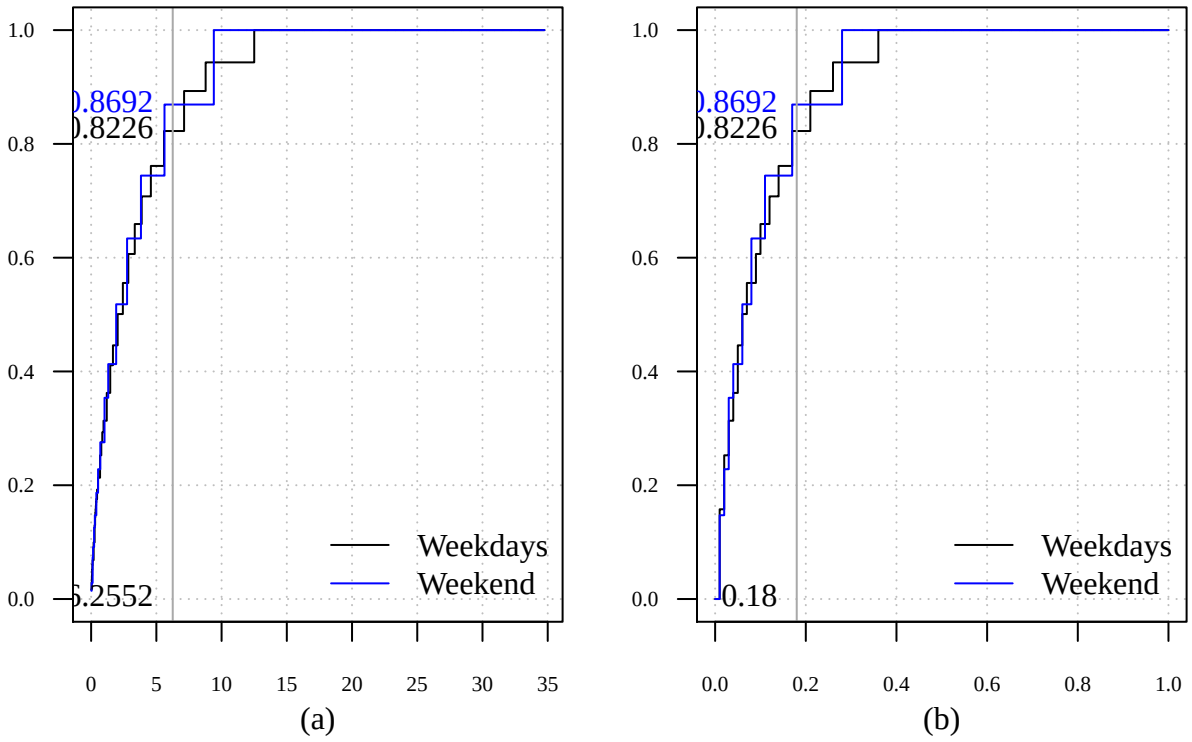

Figure 2.2: surveillance plots for weekdays and weekend under distinct perspectives

Table 2.1 is obtained with arrangement by order and shows the composition for Figure 2.2 (a). *diff* is the difference between each segments; *rank* is ranked by proportion of crashes in individual segments; p is proportion of crashes in individual segments; *Log.size* and *log.rate* are the sum of *diff* and p with the same ranked segments respectively (e.g. in $rank = 7$, 4 segments are tied, therefore with a $log.size = diff \times 4 = 0.0078$ and $log.rate = p \times 4 = 0.0145$, since the *diff* is all equal on route 71). It is plotted cumulative segmental differences versus cumulative proportion of crashes in ordered segments, and readible to see the surveillance under different desired lengths of highway.

Figure 2.2 (b), in analog of Figure 2.2 (a), will be showed simultaneously in the parentheses. The grey vertical line is the desired length of highway, top 6.26 *logmile* (or $\alpha = 0.18$) being under surveillance. Under this scenario, 82.26% and 86.92% of total crashes will be primarily on top 6.26 *logmile* (or top 18% of total segments) for weekdays and weekend respectively.

Table 1: Table 2.1: Partial output for Figure 2.2 (a)

diff	rank	p	log.size	log.rate
0.019	1	0.0164931	0.019	0.0164931
0.019	2	0.0124421	0.019	0.0124421
0.019	3	0.0118634	0.019	0.0118634
0.019	4	0.0107060	0.019	0.0107060
0.019	5	0.0086806	0.019	0.0086806
0.019	6	0.0081019	0.019	0.0081019

diff	rank	p	log.size	log.rate
0.019	7	0.0078125	0.076	0.0312500
0.019	8	0.0075231	0.076	0.0300926
0.019	9	0.0072338	0.038	0.0144676
0.019	10	0.0069444	0.019	0.0069444

2.2.2 Threshold based surveillance

Figure 2.3 shows an example of threshold based surveillance plot for weekdays and weekend. Y-axis is the proportion of segments above a given threshold whereas x-axis is desired threshold, which is the crash rate on a segment. Grey line is set to be a desired threshold 0.0019. 9.62% and 10.99% (or 176 and 201 segments) of total segments of their crash rates will be above a desired threshold of 0.0019 repectively. In comparison with fixed allocation based surveillance, it is also flexible to set up a desired threshold c instead of a desired proportion of segments α to be monitored.

Unlike the implementation of fixed allocation based surveillance, threshold based surveillance is directly built from raw data instead of ranking data first. Under a set of threshold based surveillance segments, this is a problem of giving priority to individual segment since no distinction shows beyond the entire set. Both systems have their own disadvantages and advantages; simultaneously using two systems provides an improved aspect of measuring the quality of data to set up a comparatively optimal cut-off point for surveillance.

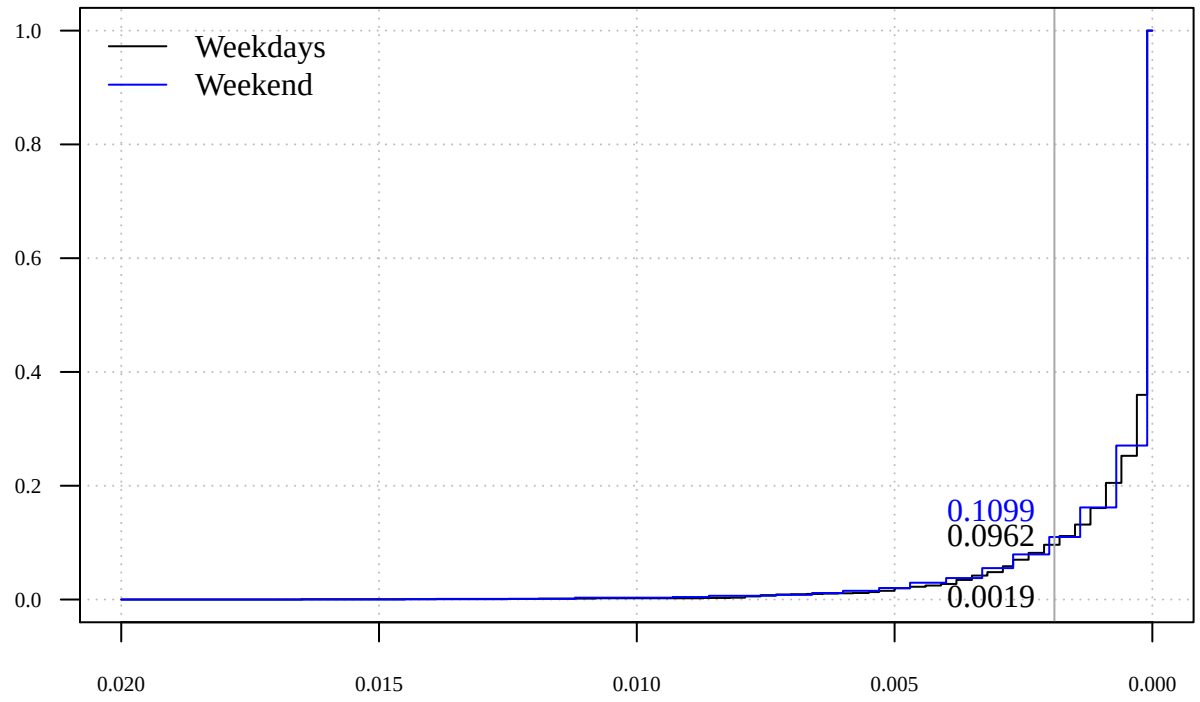

Figure 2.3: threshold based surveillance for weekdays and weekend

2.3 Kolmogorov-Smirnov test

In analog of parametric method, *Q-Q* plots or *P-P* plots for assessing normality use quantile and percentile statistics to compare the magnitude of deviations between estimated and hypothetical distributions. Kolmogorov-Smirnov test (KS test) is a common nonparametric method to evaluate the goodness of fit between estimated and hypothetical distribution. For two-sample KS test, two random samples are compared to observe magnitude of differences based on their distributions. The primary approach of finding estimated distribution of a random variable is empirical cumulative distribution function (*ecdf*) or empirical distribution function (*edf*). An *ecdf* $S_n(x)$ is defined as:

$S_n(x) = \sum_{i=1}^{n} \mathbf{I}(x_i \leq x)/n$

where $\mathbf{I}(x_i \leq x)$ is a indicator function that equals to 1 if $x_i \leq x$, and 0 otherwise.

For one-sample KS test $H_0 : S_n(x) = F_X(x)$, the D_n statistic is defined as:

$D_n = \sup_x |S_n(x) - F_X(x)|$

For two-sample KS test $H_0 : S_m(x) = S_n(x)$, the $D_{m,n}$ statistic is defined as:

$D_{m,n} = \max_x |S_m(x) - S_n(x)|$

If $D_{m,n} \geq d_0$, where $d_0 = c_\alpha \sqrt{\frac{1}{m} + \frac{1}{n}}$, c_α values can be found in any nonparametric materials, then H_0 is rejected, which implies $S_m(x)$ and $S_n(x)$ are from different distributions. Note that as $n, m \to \infty$, c_α is approximately $\sqrt{-\frac{1}{2}\ln(\frac{\alpha}{2})}$. Approximation of limiting c_α will be used for measure of distributions.

In this study, two-sample KS test is primarily conducted as a measure of association between distributions in analog of Least Significant Difference (LSD) for multiple comparisons in parametric methods. This test is contructed based on the snenarios used from a surveillance perspective whereas the surveillance is constructed based on the level set method; therefore, three ways of evaluations of highway data under different scenarios are seemingly distinct but closely tied together to organize a broaden perspecive of highway data structures.

Jean D. Gibbons, Subhabrata Chakraborti - Nonparametric Statistical Inference 4e (2003)

2.3.1 Demonstration of conducting KS test

Recall from section 2.2, surveillance using level set is comparing the proportion of crashes under various α values. Continuing using the same data as examples showed earlier, $S_m(x)$ and $S_n(x)$ are *cdf*'s for weekdays and weekend with same number of α values ($m = n = 101$) using level set method repectively. Aussume significance level is 0.05, then, $D_{m,n} = 0.085$ ($d_0 \approx 0.1911$) with *p-value* 0.96, and therefore, $H_0 : S_m(x) = S_n(x)$ is rejected. This is suggested that two distributions are indistinguishable. It is possible that model based on weekdays might have similar predictability with model based on weekend. Instead of two, one model might be sufficient for prediction and surveillance.

The purpose of conducting a KS test on a fixed allocation based surveillance is that patterns of the top desired percentage of total segments for compared distributions are intrigued instead of the entire original highway patterns. In ordered segments, unnecessary segments can be easily discarded (beyond a α level), and the desired set (within α level) can be used to construct merely one model to predict weekdays and weekend patterns with better performance and precision. Using this procedure, model complexity and performance can both be ideally achieved.

2.4 Jaccard Index plot

The Jaccard Index, or Jaccard similarity coefficient was introduced by Paul Jaccard and use as a measure of similarity of data. The Jaccard Index can be also extended to a Jaccard distance for measure of similarity in

multidimensional space. General intersection-uion relationship will be applied in this paper. Suppose two samples are drawn, it is defined as a fraction of intersection and union sets of two samples.

$J(A, B) = \frac{\|A \cap B\|}{\|A \cup B\|}$, where A and B represent two samples, $\|\cdot\|$ is the size of a sample.

Jaccard, Paul (1912), The distribution of the flora in the alpine zone

In accordance with level set method, Jaccard Index can be calculated under a sequence of α levels and the Jaccard Index plot is showed below in Figure 2.4. Grey lines are at α level 0.1 and 0.2, with percentages of matched segments 47.93% and 59.03% respectively for the Jaccard Index of weekdays and weekend sets. Note that percentage of matched rate stays constant starts from roughly $\alpha = 0.37$, this is caused by no crash records within the rest of highway segments.

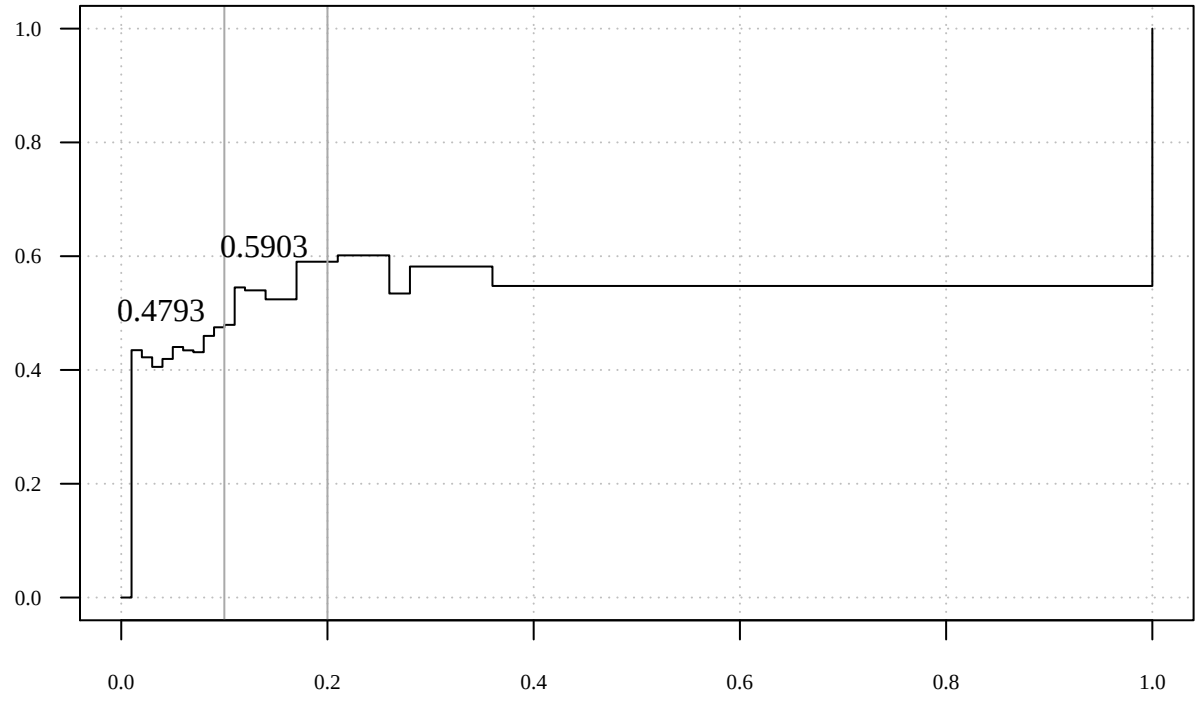

Figure 2.4: Jaccard Index for weekdays and weekend sets

3. Predictability of Crash Data

In this section, tools mentioned in section 2 will be implemented as a surveillance process of quality of data.

3.1 Data sources

The dataset used is composed of two subsets: *Highway* data and *Crash* data, and focus on the state of Arkansas region. *Highway* data recorded the highway surface characteristics across the entire highway system in Arkansas. Categories and variables such as location (logmile, longitude, latitude), traffic characteristics (average daily traffic), surface characteristics (type of road, sign of route, type of operation, median type and number of lanes) and others are included in this set. *Crash* data recorded individual crash cases at the spot. It includes time features (crash date, crash time), location (latitude, longitude, city, county, street),

highway features (class of property, classification of trafficway, road system, intersection, junction relation), environmental features (light condition, road way surface condition) and crash conditions (interaction type, crash manner, number of fatals, number of injuries, number of vehicles) and others are included in this set.

The data used in this paper is obtained from The Center for Advanced Public Safety (CAPS) at The University of Alabama. CAPS also cooperates with centers in other universities to work on various projects in different areas such as traffic safety and engineering and analytics, and has access to these numerous valuable data sources such as highway and crash data in different time and areas. This benefits this paper to conduct thorough analyses.

Grigorios Fountas a, Panagiotis Ch. Anastasopoulos, "A random thresholds random parameters hierarchical ordered probit analysis of highway accident injury-severities"

3.2 Concentration of crashes: empirical performance measures

In previous demonstrations, crashes behaviors along the highway in separated time, weekdays and weekend, are showed a process of screening data. Results are also presented that weekdays and weekend may consider both predictive, i.e., a model can be built using one of set, and this model may also capture a similar pattern for the other set. Other characteristrics of causes of crashes, e.g. time variation and segment variation, will be compared through the same process as demonstrations earlier to uncover patterns of latent connections.

3.2.1 Time variation

Given different time periods, distributions of crashes can be varied. Several comparisons of distributions from time to time will be illustrated and only insignificant groups are summarized under significance level of 0.05 in following tables.

To make identifications between months is arduous by starting with a plot of ordered segments by crash percentages for months of year (Figure 3.1). January has higher intensity on crashes for first couple segments whereas October has relatively flat coverage across entire set of segments. It is expected that January may have higher predictability than October because of its concentration on fewer segments.

Table 2: Table 3.1: KS tests using level set for months of year

	Subject	Insignificance
1	Jan	Jun (0.1906)
2	Feb	Apr (0.1630), Jul (0.0622), Nov (0.0602)
3	Mar	Apr (0.0395), May (0.0798), Sep (0.0280), Nov (0.1600)
4	Apr	May (0.1193), Sep (0.0675), Nov (0.1292)
5	May	Sep (0.0518), Nov (0.1881)
6	Jul	Nov (0.0748)
7	Aug	Oct (0.0530)
8	Sep	Nov (0.1794)

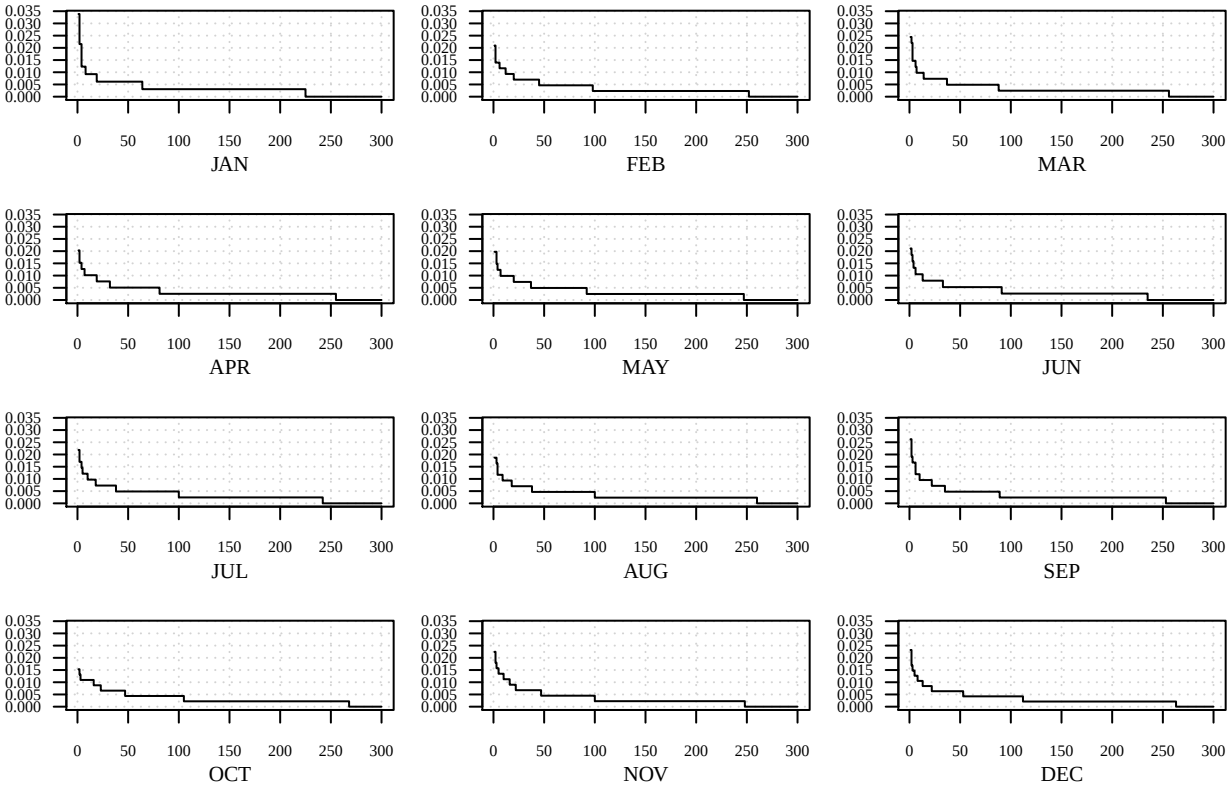

Figure 3.1: ordered segments by percentages of crashes for months of year

In conjuction with KS test results as showed in Table 3.1, subtle diversities in months become detected from the KS statistics. December is comparatively distinct from any other months, and then January, June, July and November. On the other hand, February through May and August through October may be individually similar to each other respectively. Note that these classifications or similarities may follow a pattern of an academic year. A scenario can be established for distinctions between semester and off-semester periods. Also, off-semester months are varied in each other whereas semester months are similar; this results that semester months may be more appropriate for prediction than off-semester months due to the diversities of distributions in off-semester period.

Table 3: Table 3.2: KS tests using level set for days of week

	Subject	Insignificance
1	Sun	Fri (0.1906)
2	Mon	Wed (0.1630), Sat (0.0622)
3	Tue	Wed (0.0395), Thu (0.0798)
4	Wed	Thu (0.1193)

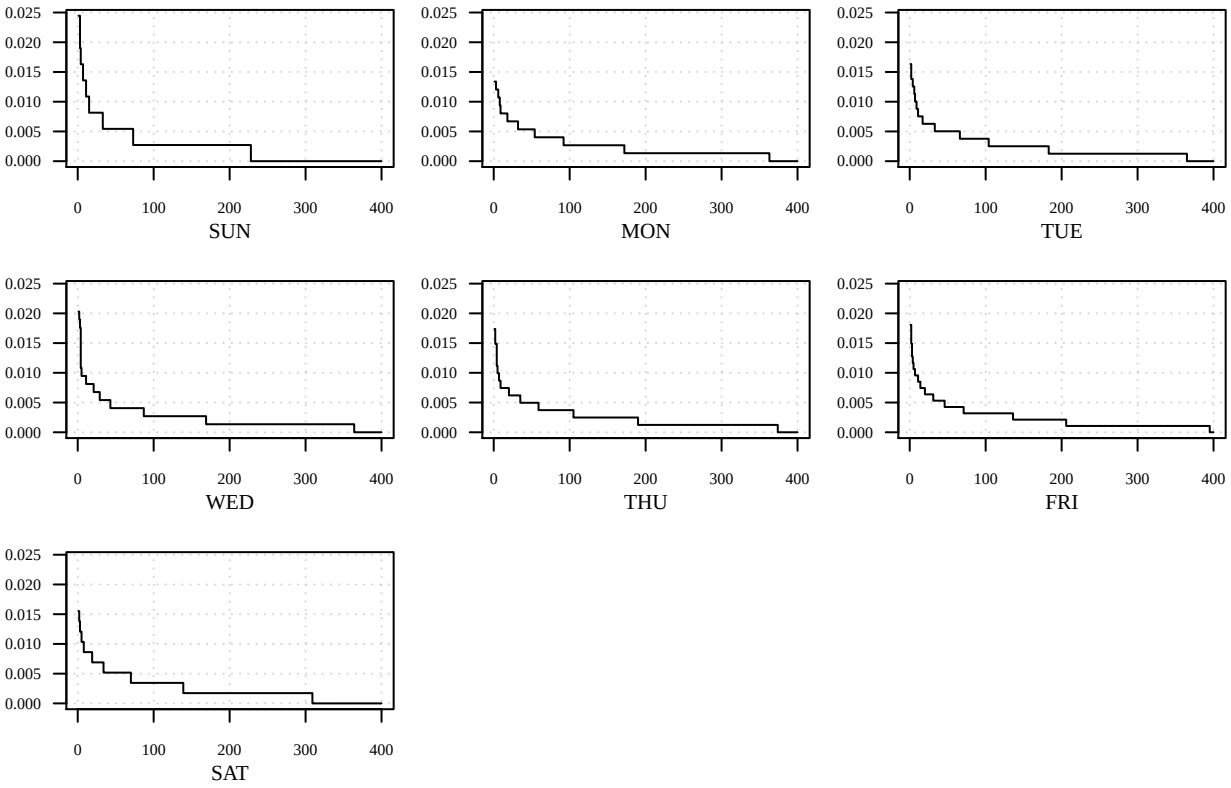

Figure 3.2: ordered segments by percentages of crashes for days of week

3.2.2 Segment variation

```
        - in extreme: for continuous time and space, then perfectly predictable.
        - another extreme: worst (optimal) performance should be from using full year and large grid/seg
- use type 1 to predict, use type 2 to evaluate
    - type 1: day-time, type 2: night-time (etc, many scenarios)
    - show performance according the the evaluation methods we defined in 1.
```

Crash Model Evaluation - We may not need this for the paper, I will evaluate after first two parts are written - Apply a standard model (no new methodology here) - evaluate according to the methods we outlined - show how close to optimal the method performs