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

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.

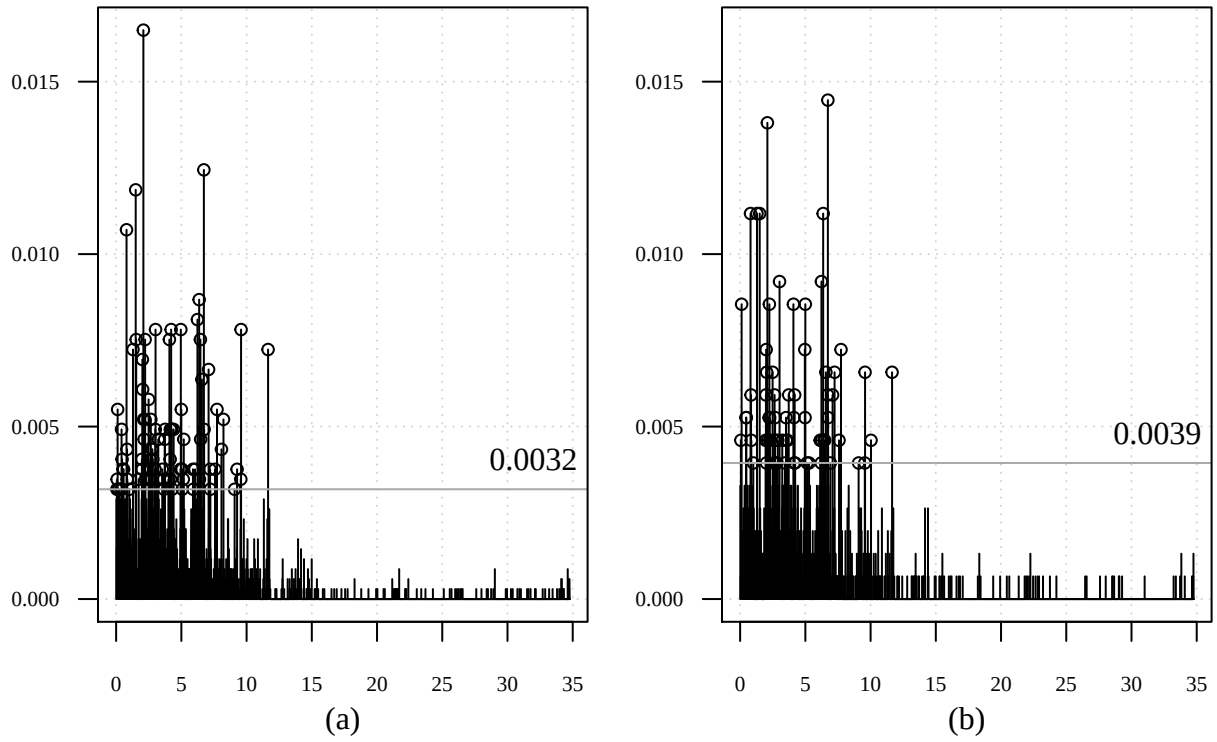

Figure 2.1: weekdays (a) and weekend (b) with $\alpha = 0.05$

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 by crashes and top proportion of segments expected to be monitored, which is α. The following paragraphs will describe this figures in detail.

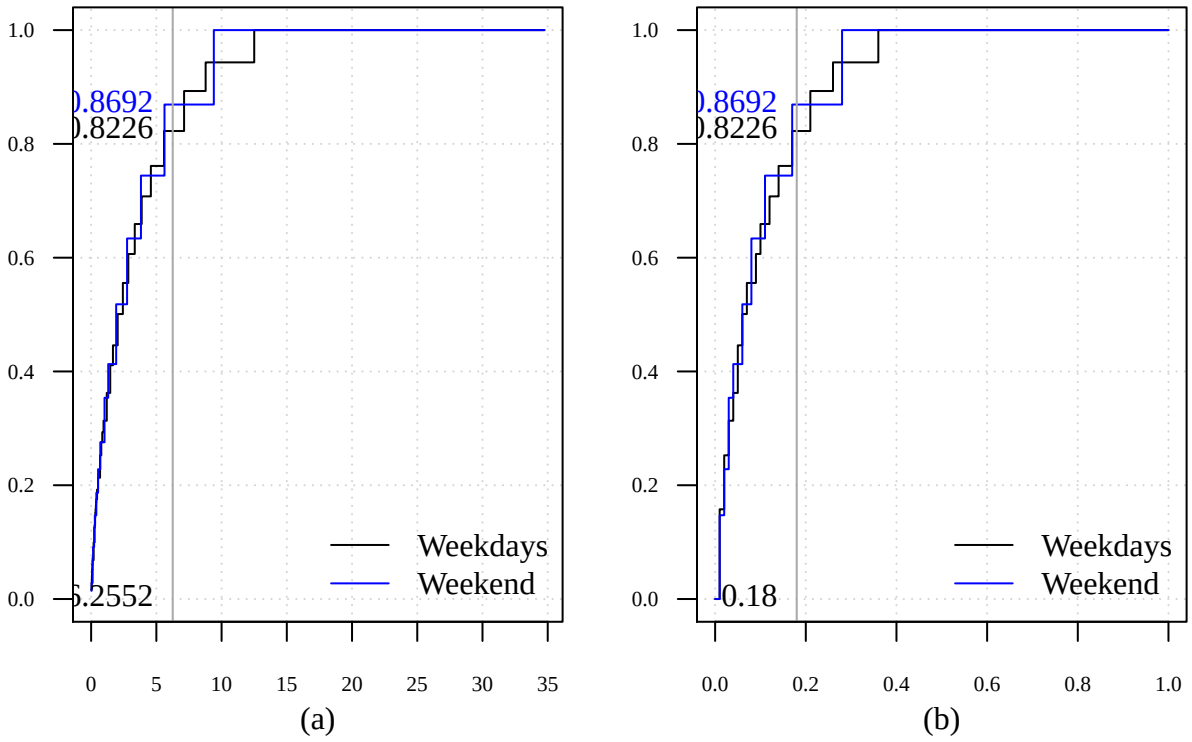

Figure 2.2: surveillance plots for weekdays and weekend under distinct perspectives α levels

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* × 4 = 0.0078 and *log.rate* = *p* × 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.

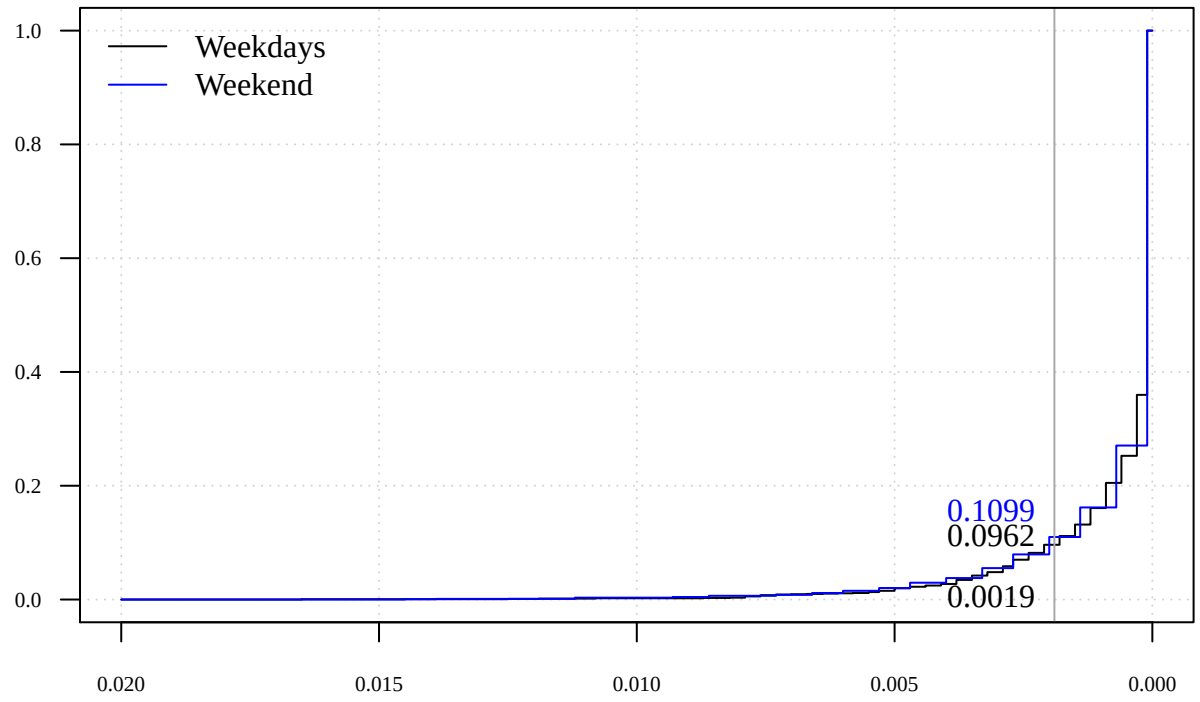

Figure 2.3: threshold based surveillance for weekdays and weekend

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.

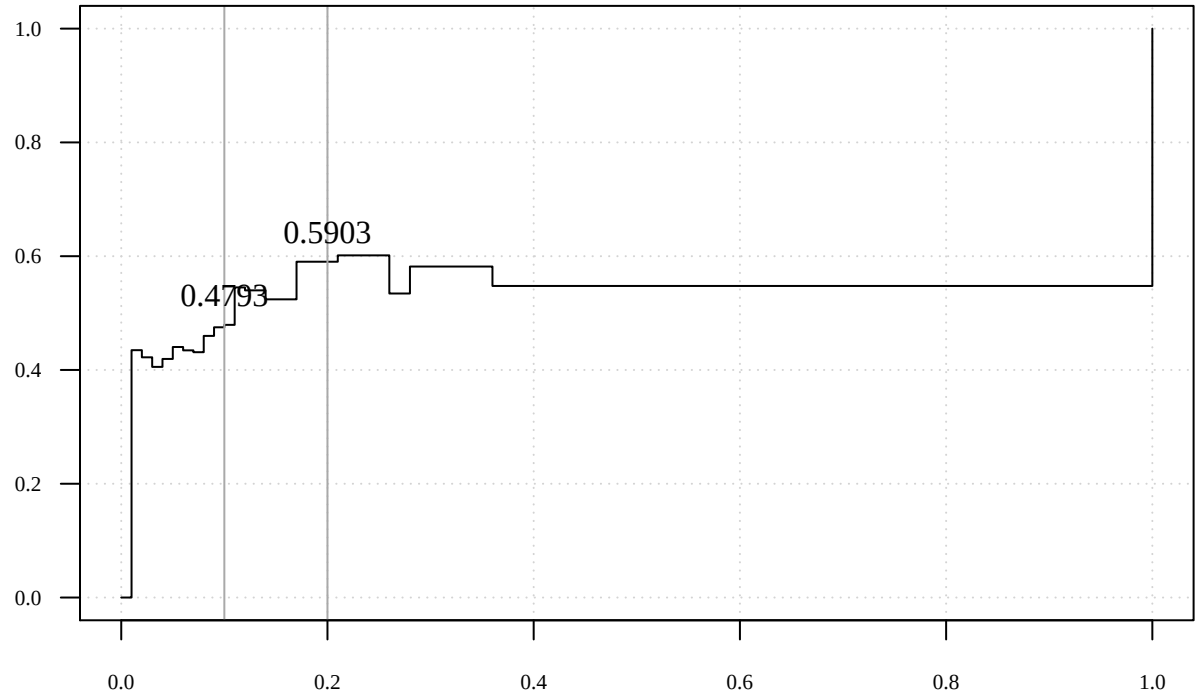

Figure 2.4: Jaccard Index for weekdays and weekend sets

3. Predictability of Crash Data

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 given different time periods, weekdays and weekend, are primarily showed a process of measuring data. Results are also presented that weekdays and weekend may consider both predictive, i.e., a model can be built using one of both sets, 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 are 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.

The magnitude of differences in distributions is generally unobservable, but subtle diversities in months become detected from the KS statistics (Table 3.1). 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 2: Table 3.1: KS tests using level set for months of year

Reference	Compare	D	p.value
JAN	JUN	0.1906	0.0509
FEB	APR	0.1630	0.1365
FEB	JUL	0.0622	1.0000
FEB	NOV	0.0602	1.0000
MAR	APR	0.0395	1.0000
MAR	MAY	0.0798	1.0000
MAR	SEP	0.0280	1.0000
MAR	NOV	0.1600	0.1506
APR	MAY	0.1193	0.4750
APR	SEP	0.0675	1.0000
APR	NOV	0.1292	0.3709
MAY	SEP	0.0518	1.0000
MAY	NOV	0.1881	0.0562
JUL	NOV	0.0748	1.0000
AUG	OCT	0.0530	1.0000
SEP	NOV	0.1794	0.0776

Classifications based on semester and off-semestere choice using distributions of months are shown in Figure 3.1. For semester class curves appear to be more consistent and slightly diverse as the proportion of segments increases whereas curves for off-semester class are more diverse across a sequence of α values. It is possible that using any one of semester months to construct the baseline is likely to have similar predictability in the same class. For off-semester months, choice of months as the baseline results in comparatively lower predictability.

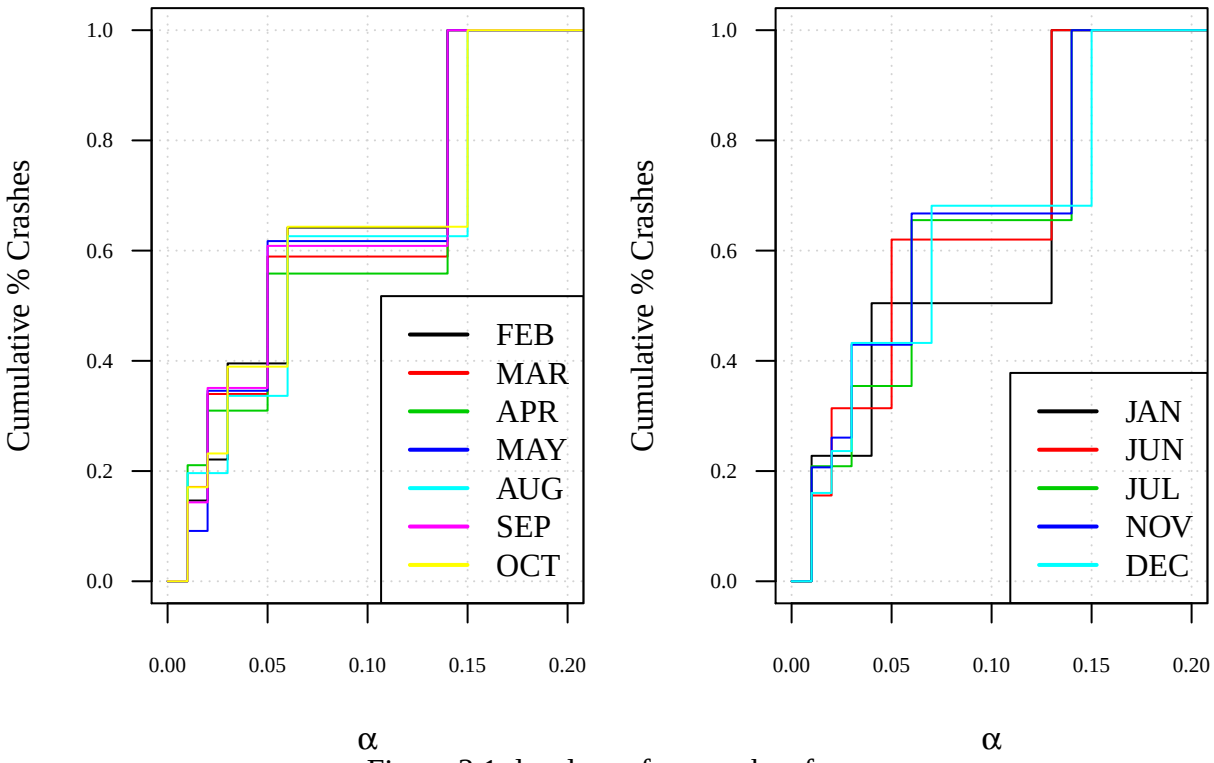

Figure 3.1: level sets for months of year

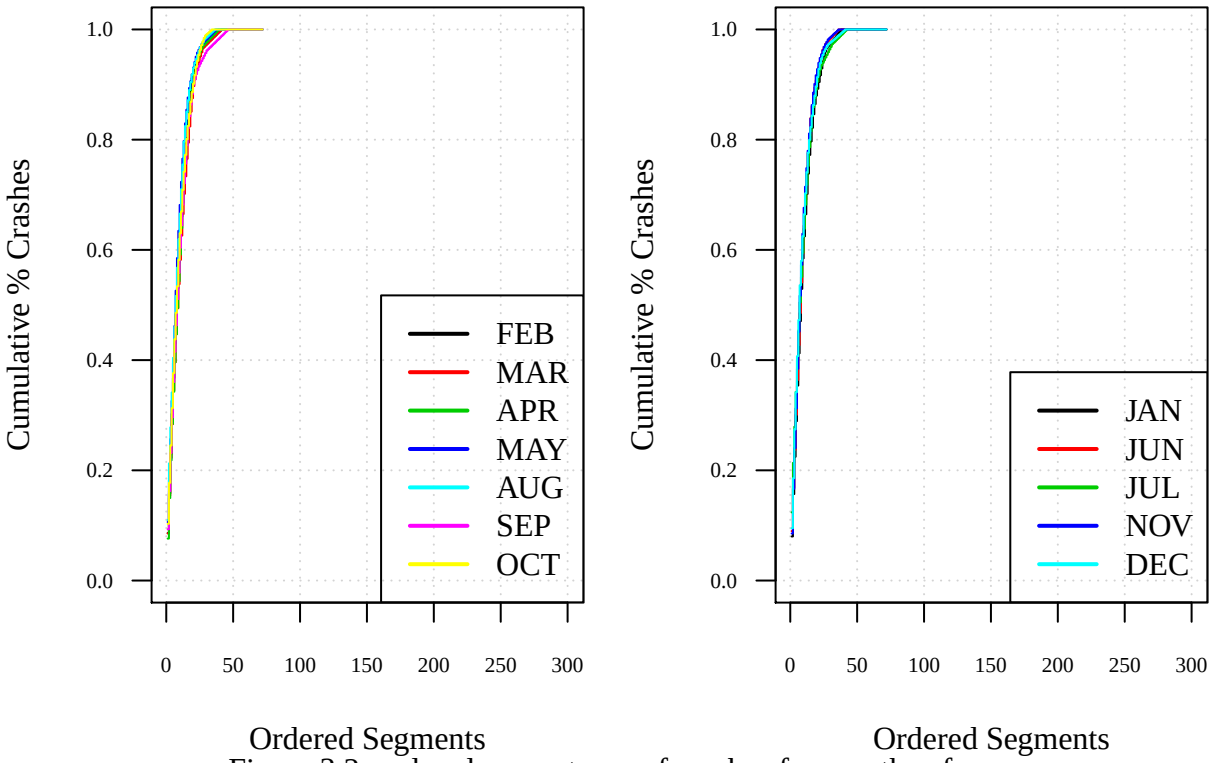

Figure 3.2: ordered percentages of crashes for months of year

Days of week are also of interest to be compared in the following (Figure 3.3) along with KS results from Table 3.2. Monday through Wednesday are comparatively stable and almost as same as each other in patterns; Thursday through Sunday are distinct from each other. A possible classification may be before-midweek and after-midweek patterns. A before-midweek based structure appears to be more representive and stable than the other group under this circumstance. Likewise, each of before-midweek based structure is expected to have higher predictability than of after-midweek based structure.

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

Reference Compare D p.value ———- ——— ——- ——— MON TUE 0.1137 0.5421 MON WED 0.0419 1.0000 TUE WED 0.0941 0.8184

<———>

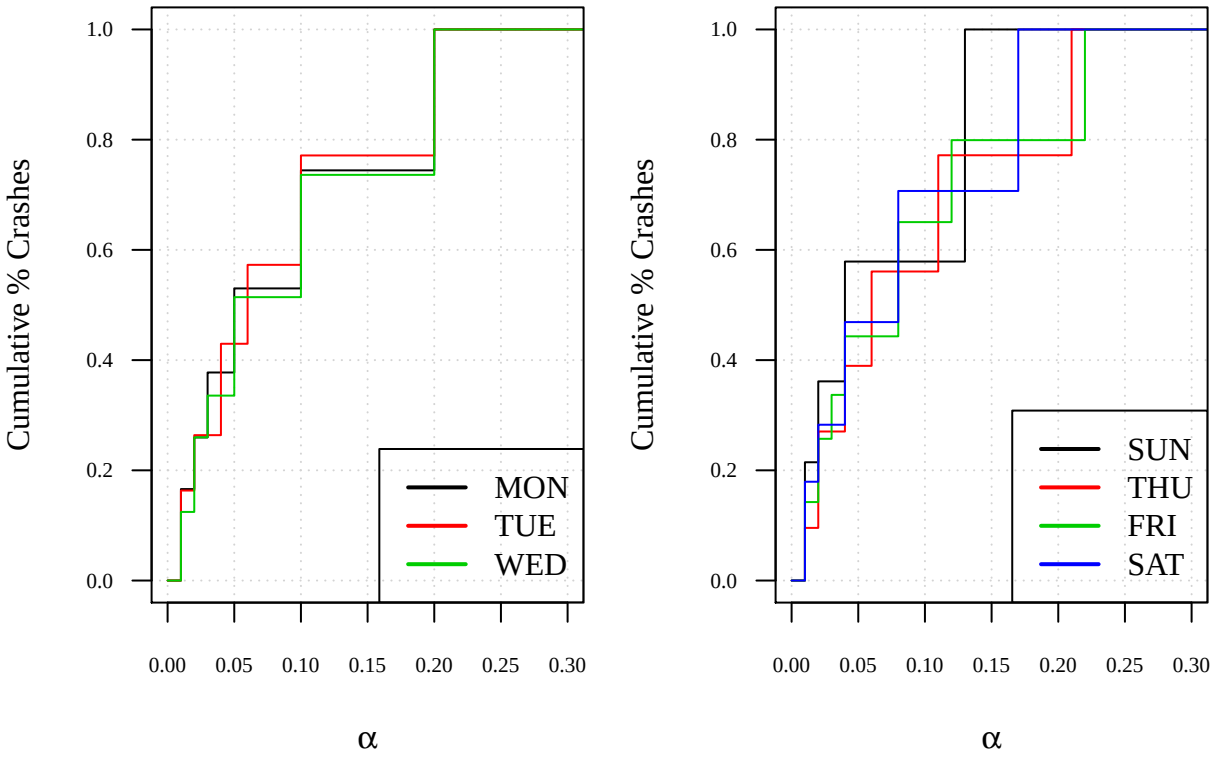

Figure 3.3: level sets for days of week

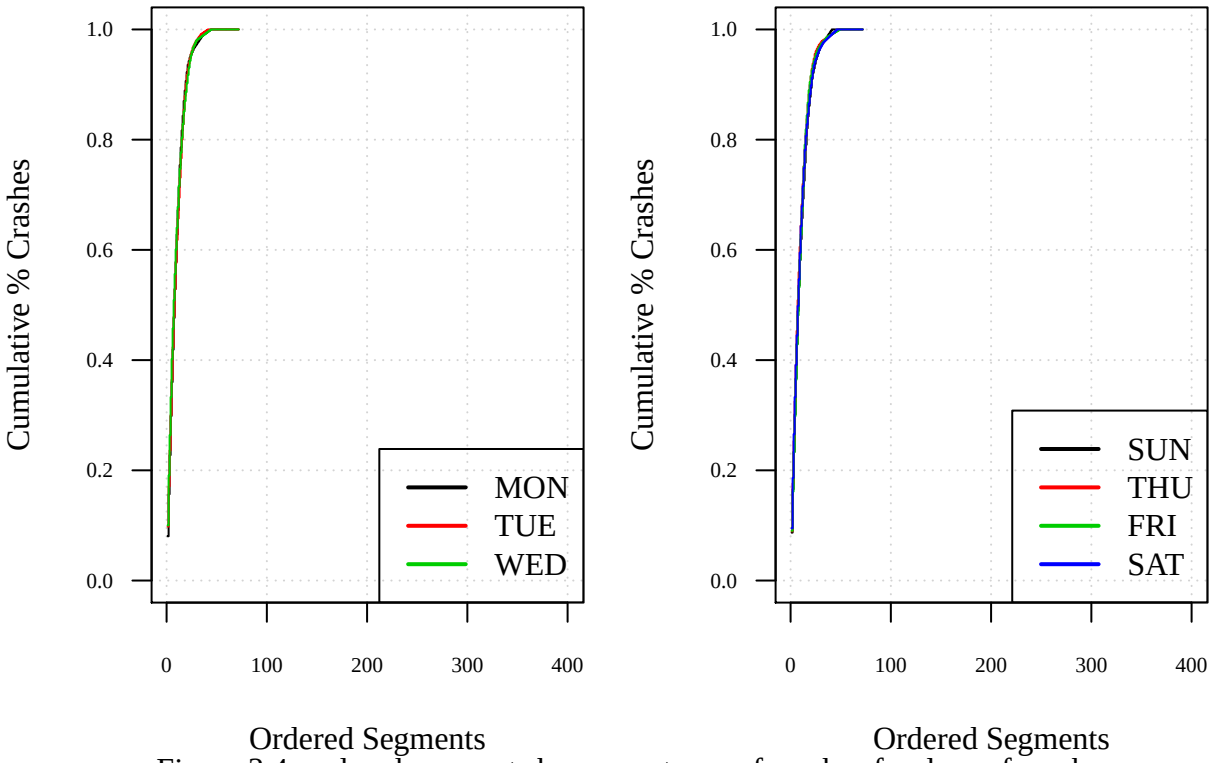

Figure 3.4: ordered segments by percentages of crashes for days of week

Two primary time separations, days of week and months of year, are ways of using one data to make predictions of the other data. One other possibility is to separate time of day into different time periods. The purpose of screening data beforehand is to ensure the quality of data used is predictive to capture the true pattern of crashes based on certain characteristics.

Considering time of day divided into 2-hour, 4-hour, 8-hour and 12-hour periods starting from 00:00 for each category, it is more likely to go extreme condition while more categories are used, i.e., slightly difference in a specific time period and a highway segments tends to be significant. The boundary and range for different division for time of day are arbituary, other reasonable choices of time period are also applicable. For 2-hour periods, from 12:00 to 20:00, are showed to be insignificant but at the boundary of 0.05; for 4-hour periods, from 04:00 to 20:00 tend to have similar patterns; for 8-hour periods, 08:00 to 24:00, are also insignificant but close to the significant boundary; 12-hour periods are similar to each other.

Table 3: Table 3.3: KS tests using level set for time of day in 12 shifts

Reference	Compare	D	p.value
s7	s9	0.1863	0.0601
s8	s9	0.1863	0.0601
s9	s10	0.1887	0.0549

Table 4: Table 3.4: KS tests using level set for time of day in 6 shifts

Reference	Compare	D	p.value
s2	s5	0.1717	0.1017
s3	s4	0.1207	0.4595
s3	s5	0.1313	0.3504
s4	s5	0.1123	0.5593

Table 5: Table 3.5: KS tests using level set for time of day in 3 shifts

Reference	Compare	D	p.value
s2	s3	0.0798	1

Table: Table 3.6: KS tests using level set for time of day in 2 shifts

Reference Compare D p.value ———- ——— ——— ——— s1 s2 0.083 0.9966

<———>

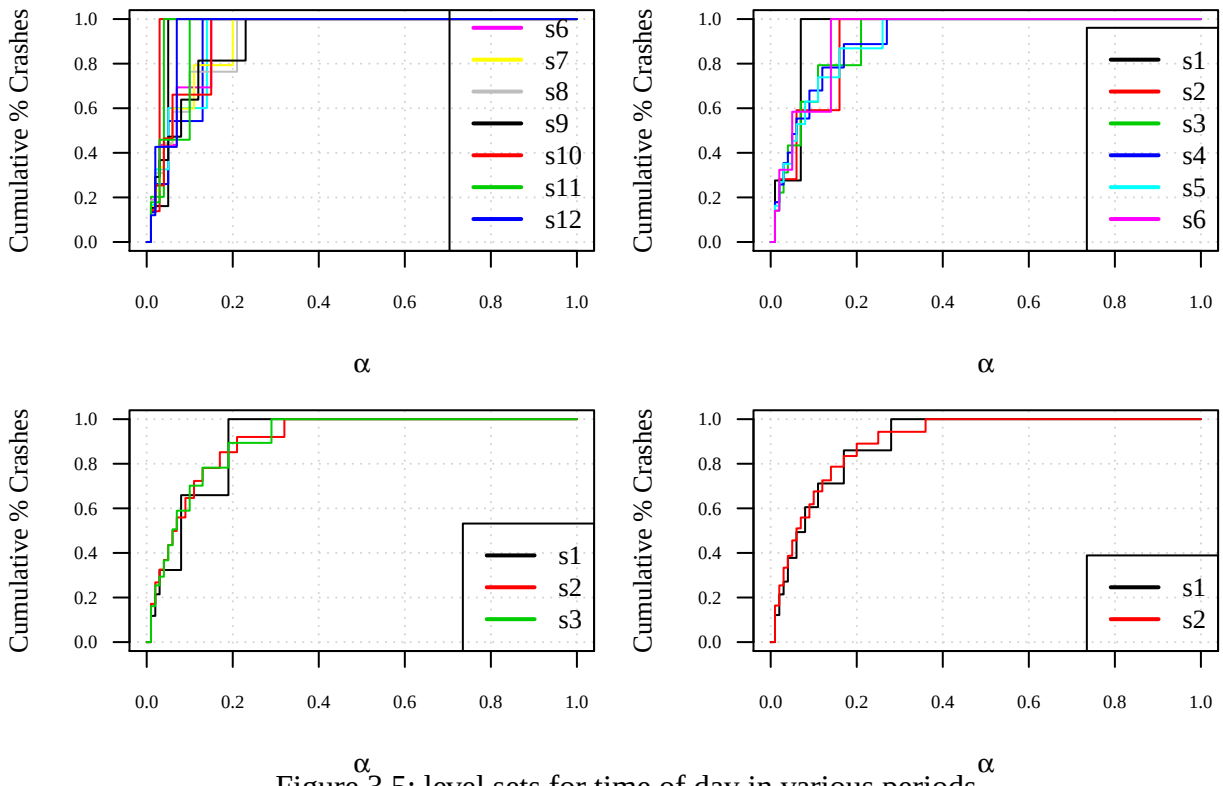

Figure 3.5: level sets for time of day in various periods

3.2.2 Segment variation

In time variations and previous descriptions, the distance between each primitive segments is fixed to show a thorough picture of distributions. Binning can determine a variety of distances to provide a less complex and flexible method of measurement. In most cases, binning is arduous to determine a suitable bin size because of its classification of observations at the boundaries; however, binning is simple to implement and adjust to requirements. Couple choices of binning size will be showed to observe the influences of predictions.

Recall from the comparison between weekdays and weekend data, bin sizes of 1 and 0.5 for weekdays and weekend are compared under the ordered bins and level sets (Figure 3.6 and 3.7). In the demonstration of KS test previously (Section 2.3.1), the statistic is $D_{m,n} = 0.085$, the maximal differences between distributions for two bin sizes is 0.0177 and 0.0153 repectively, which are insigificant and expecxted, because binning will result data being smoother.

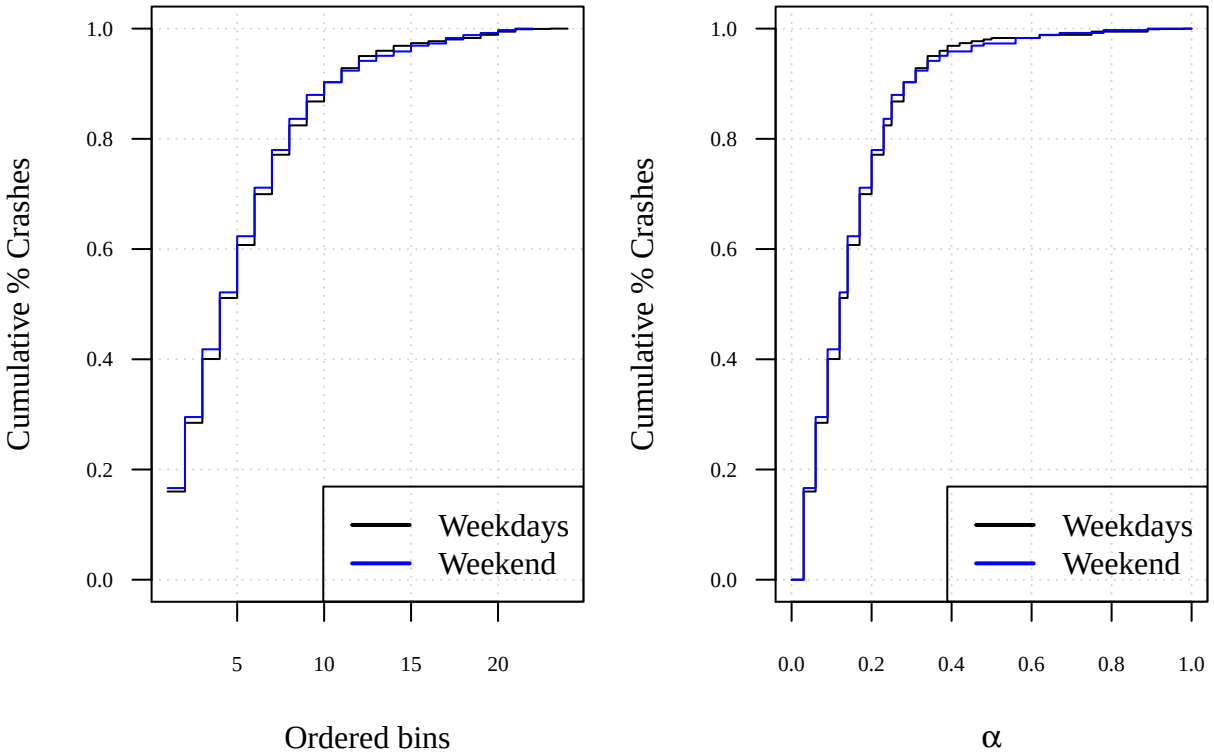

Figure 3.6: ordered bins with bin size = 1 for weekdays and weekend

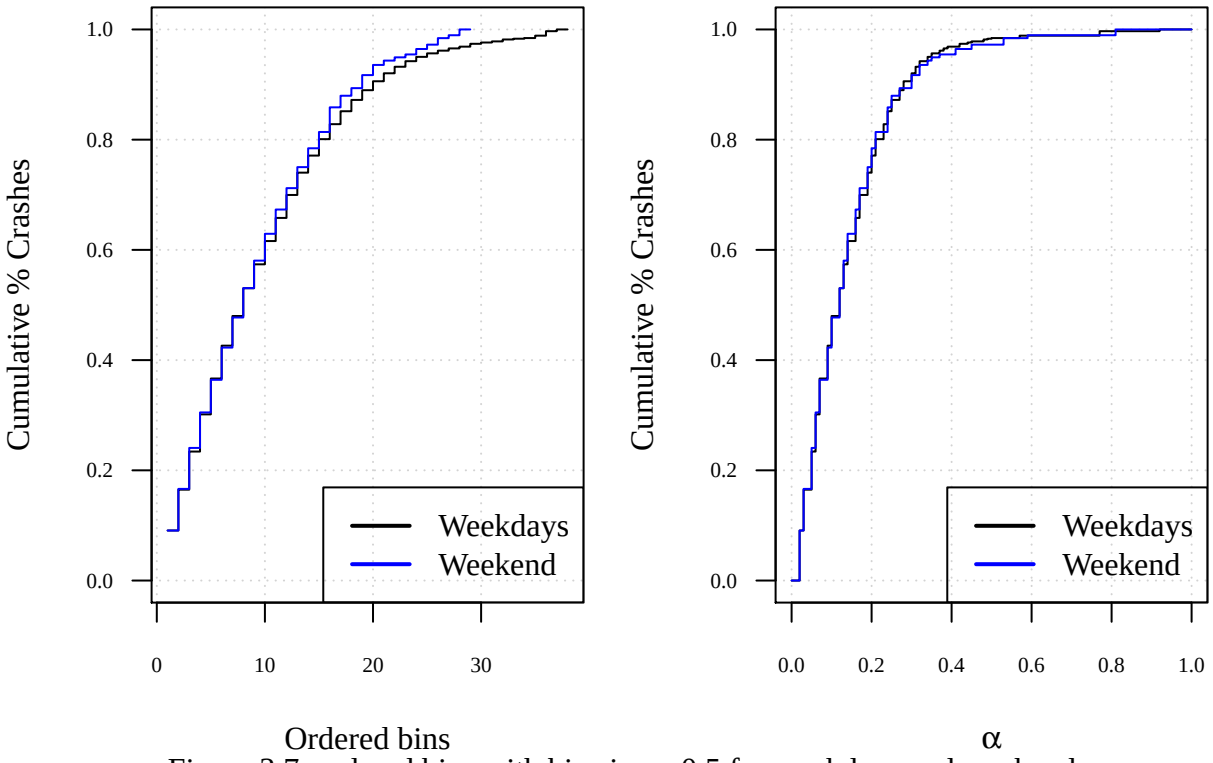

Figure 3.7: ordered bins with bin size = 0.5 for weekdays and weekend

Bin size of 1 and 0.5 for days of week are showed in Figure 3.8 and 3.9 along with KS results in table 3.7 and 3.8, respectively. Both results show insignificance in distributions whether in ordered bins or level sets fashion. With bin size of 1, Sunday appears to slightly stand out compared to the others although it shows insignificant. For the rest of the days, each of them can be used as reference and may still have the smiliar predictablity.

Table 6: Table 3.7: KS tests using level set with bin size 1 for days of week

Reference	Compare	D	p.value
SUN	MON	0.1513	0.1983
SUN	TUE	0.1621	0.1409
SUN	WED	0.1691	0.1112
SUN	THU	0.1638	0.1332
SUN	FRI	0.1541	0.1818
SUN	SAT	0.1862	0.0603
MON	TUE	0.0295	1.0000
MON	WED	0.0622	1.0000
MON	THU	0.0471	1.0000
MON	FRI	0.0273	1.0000
MON	SAT	0.0465	1.0000
TUE	WED	0.0571	1.0000
TUE	THU	0.0423	1.0000
TUE	FRI	0.0266	1.0000
TUE	SAT	0.0533	1.0000
WED	THU	0.0649	1.0000

Reference	Compare	D	p.value
WED	FRI	0.0735	1.0000
WED	SAT	0.0478	1.0000
THU	FRI	0.0277	1.0000
THU	SAT	0.0456	1.0000
FRI	SAT	0.0469	1.0000

Table: Table 3.8: KS tests using level set with bin size 0.5 for days of week

Reference Compare D p.value ———- ——— ——- ——— SUN MON 0.0929 0.8369 SUN TUE 0.0734 1.0000 SUN WED 0.0981 0.7565 SUN THU 0.0899 0.8834 SUN FRI 0.0977 0.7625 SUN SAT 0.0998 0.7320 MON TUE 0.0796 1.0000 MON WED 0.1072 0.6260 MON THU 0.0965 0.7812 MON FRI 0.0761 1.0000 MON SAT 0.1089 0.6039 TUE WED 0.0376 1.0000 TUE THU 0.0889 0.9006 TUE FRI 0.0805 1.0000 TUE SAT 0.0966 0.7797 WED THU 0.0612 1.0000 WED FRI 0.1005 0.7216 WED SAT 0.1025 0.6916 THU FRI 0.0923 0.8455 THU SAT 0.1072 0.6268 FRI SAT 0.1000 0.7280

<———————>

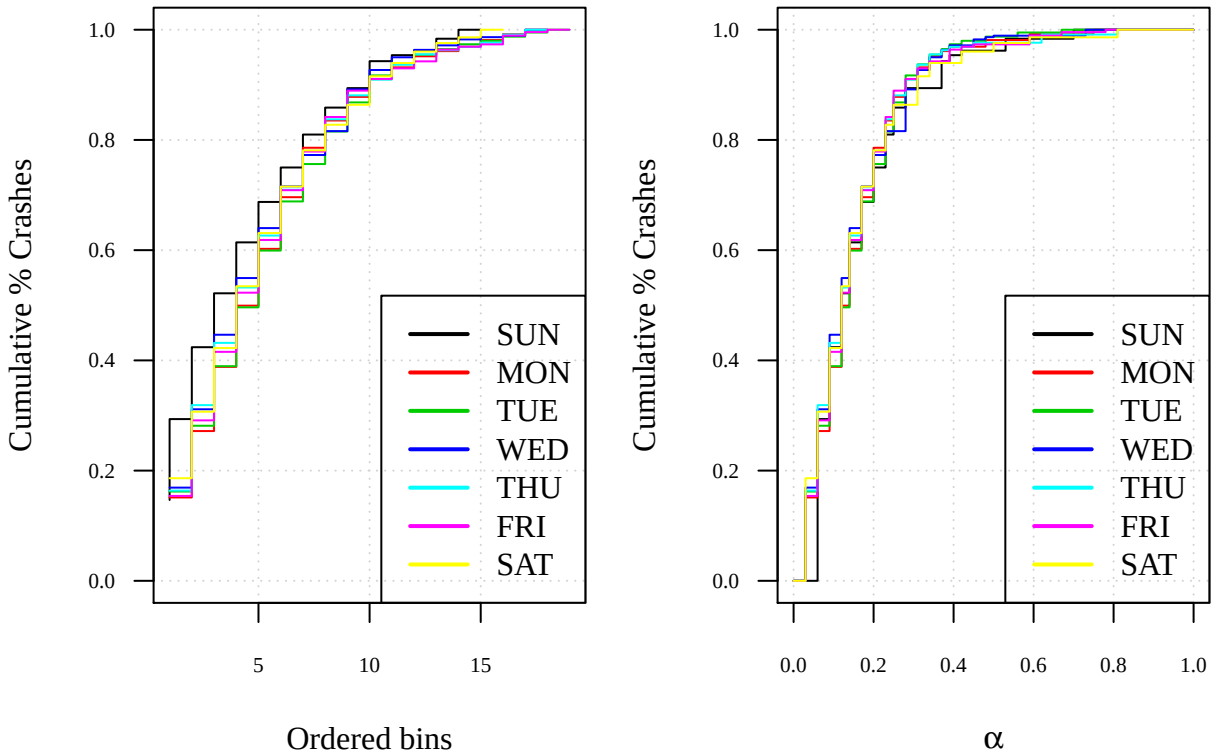

Figure 3.8: ordered bins with bin size = 1 for days of week

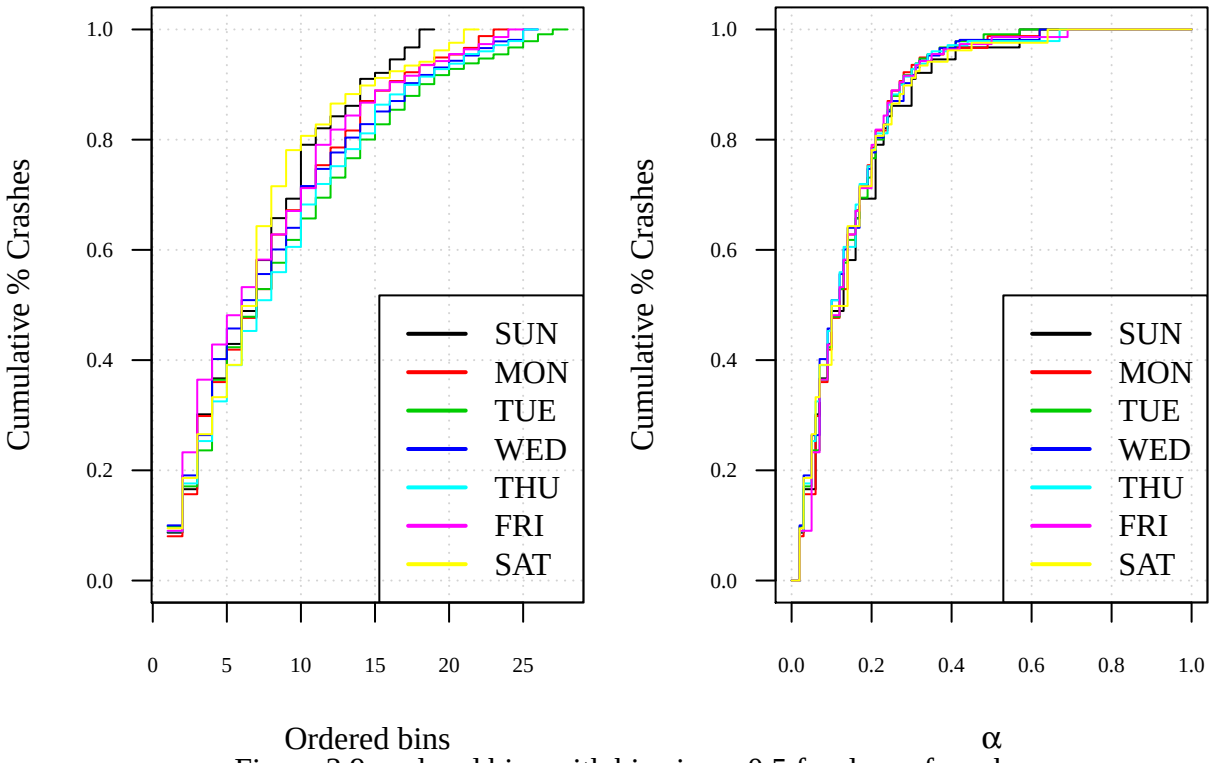

Figure 3.9: ordered bins with bin size = 0.5 for days of week

Considering different bin size 1 and 0.5 for distributions of crashes for months of year from Table 3.9 and 3.10 along with Figure 3.10 and 3.11, differences between months become smaller. For bin size 1, only June and August still have apparent gap in distributions. For bin size 0.5, February has the similar pattern as April, July and November before binning; here it results in difference between February and July after binning. This may account for classifications for observations at the boundaries that sometimes cause inconsistent results. January, March, September and December are not showed on these table, which may suggest a close performance on predictions using any of them because of the similarities with other months.

Table 7: Table 3.9: KS tests using level set with bin size 1 for months of year

Reference	Compare	D	p.value
JUN	AUG	0.2126	0.0208

Table: Table 3.10: KS tests using level set with bin size 0.5 for months of year

Reference Compare D p.value ———- ——— ——- ——— FEB JUL 0.1931 0.0463 FEB AUG 0.1920 0.0483 APR MAY 0.2348 0.0076 APR JUN 0.1971 0.0395 APR AUG 0.2192 0.0156 APR NOV 0.2297 0.0097 APR DEC 0.2179 0.0166 MAY OCT 0.1969 0.0398 OCT NOV 0.1916 0.0491

<———>

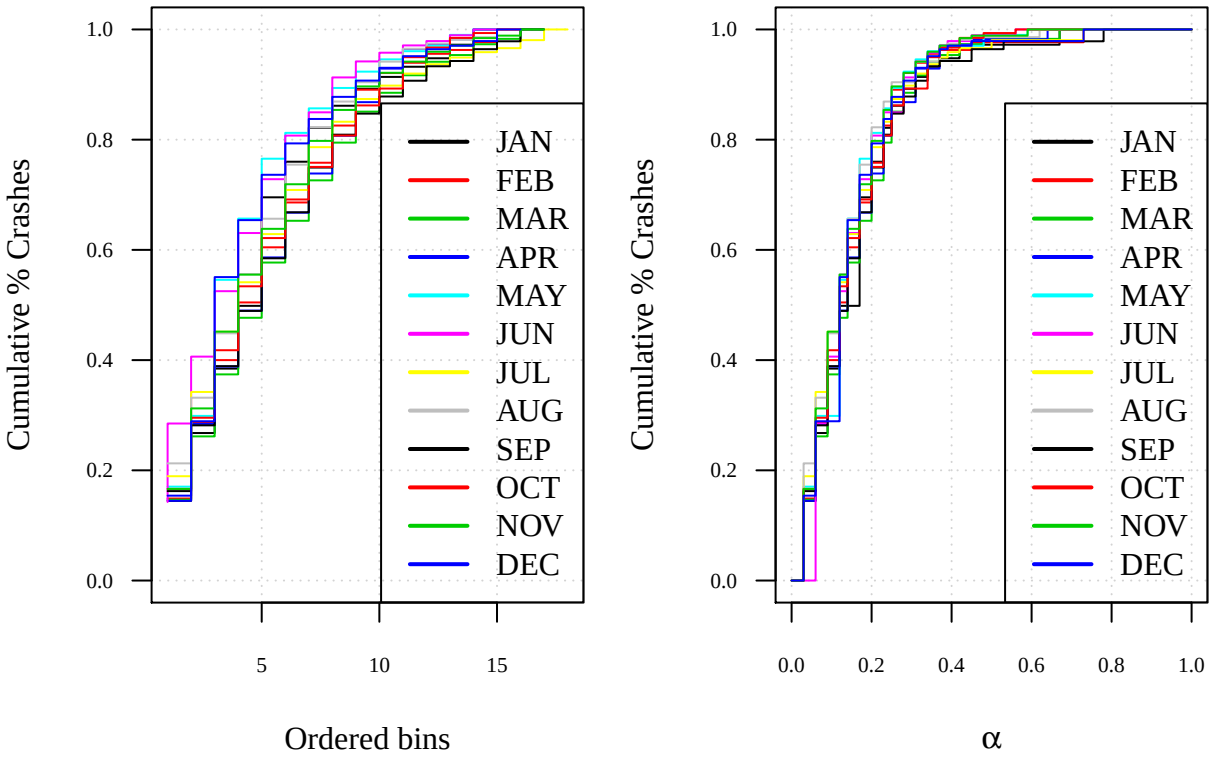

Figure 3.10: ordered bins with bin size = 1 for months of year

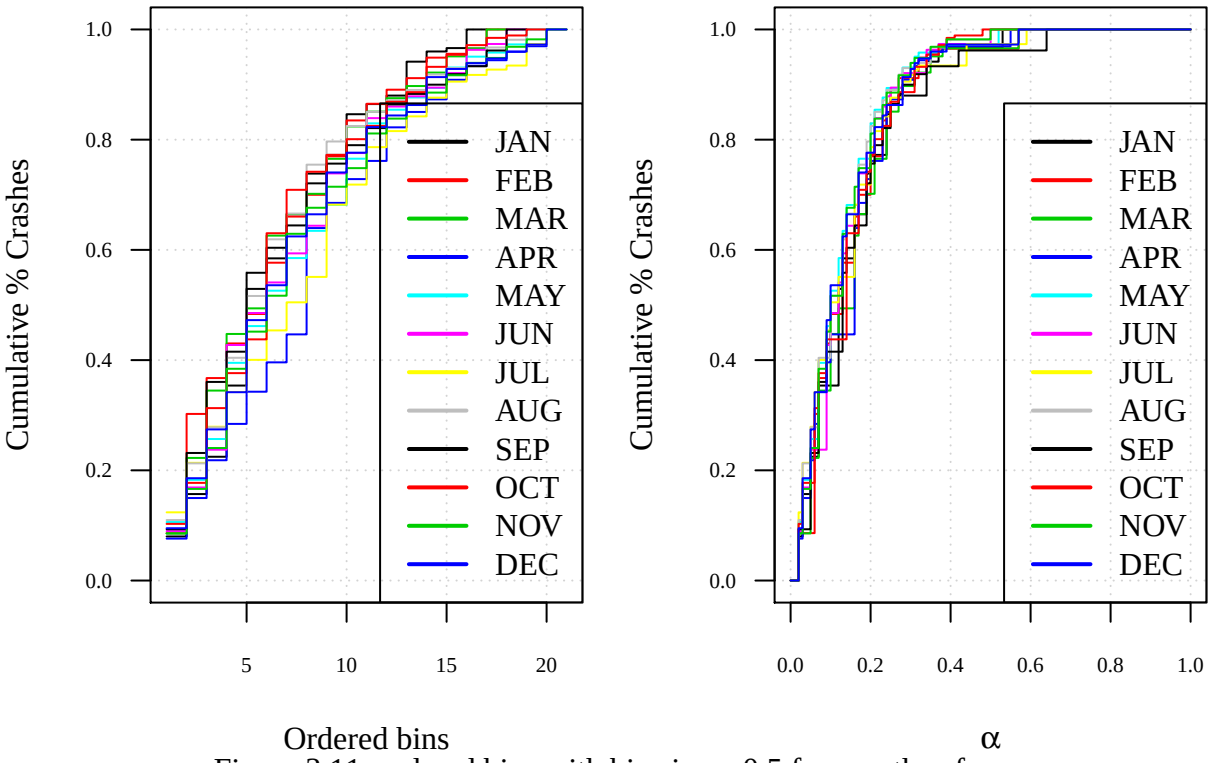

Figure 3.11: ordered bins with bin size = 0.5 for months of year

Considering different bin size 1 and 0.5 for distributions of crashes for time of day from Table 3.11-17 along with Figure 3.11-14. For 12 shifts, 20:00-22:00 and 02:00-06:00 with bin size 1, 00:00-04:00 with bine size 0.5, are apparently different in distributions from rest of the groups individually; for 6 shifts, only 04:00-08:00 and 20:00-24:00 with bin size 1 is different than the rest whereas all are similar for bin size 0.5; for 3 shifts and 2 shifts are individually showed similar in its own separations.

Table: Table 3.11: KS tests using level set for time of day bin size 1 in 12 shifts

Reference Compare D p.value ———- ——— — ———

Table 8: Table 3.12: KS tests using level set for time of day bin size 0.5 in 12 shifts

Reference	Compare	D	p.value
s1	s2	0.2586	0.0023

Table: Table 3.13: KS tests using level set for time of day bin size 1 in 6 shifts

Reference Compare D p.value ———- ——— — ———

Table 9: Table 3.14: KS tests using level set for time of day bin size 1 in 3 shifts

Reference	Compare	D	p.value
s1	s2	0.083	0.9972

Table 10: Table 3.15: KS tests using level set for time of day bin size 0.5 in 3 shifts

Reference	Compare	D	p.value
s1	s2	0.1577	0.162

Table 11: Table 3.16: KS tests using level set for time of day bin size 1 in 2 shifts

Reference	Compare	D	p.value
s1	s2	0.1273	0.3891

Table: Table 3.17: KS tests using level set for time of day bin size 0.5 in 2 shifts

Reference Compare D p.value ———- ——— ———- ——— s1 s2 0.0669 1

<———>

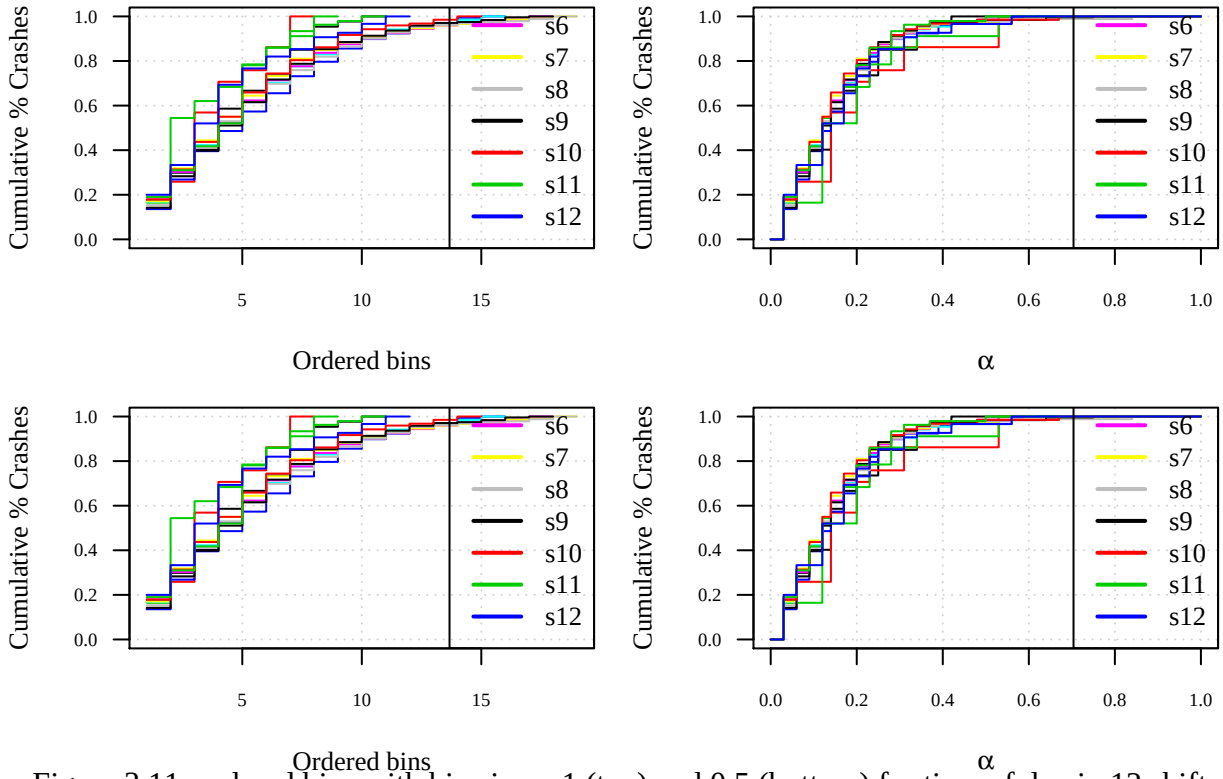

Figure 3.11: ordered bins with bin size = 1 (top) and 0.5 (bottom) for time of day in 12 shifts

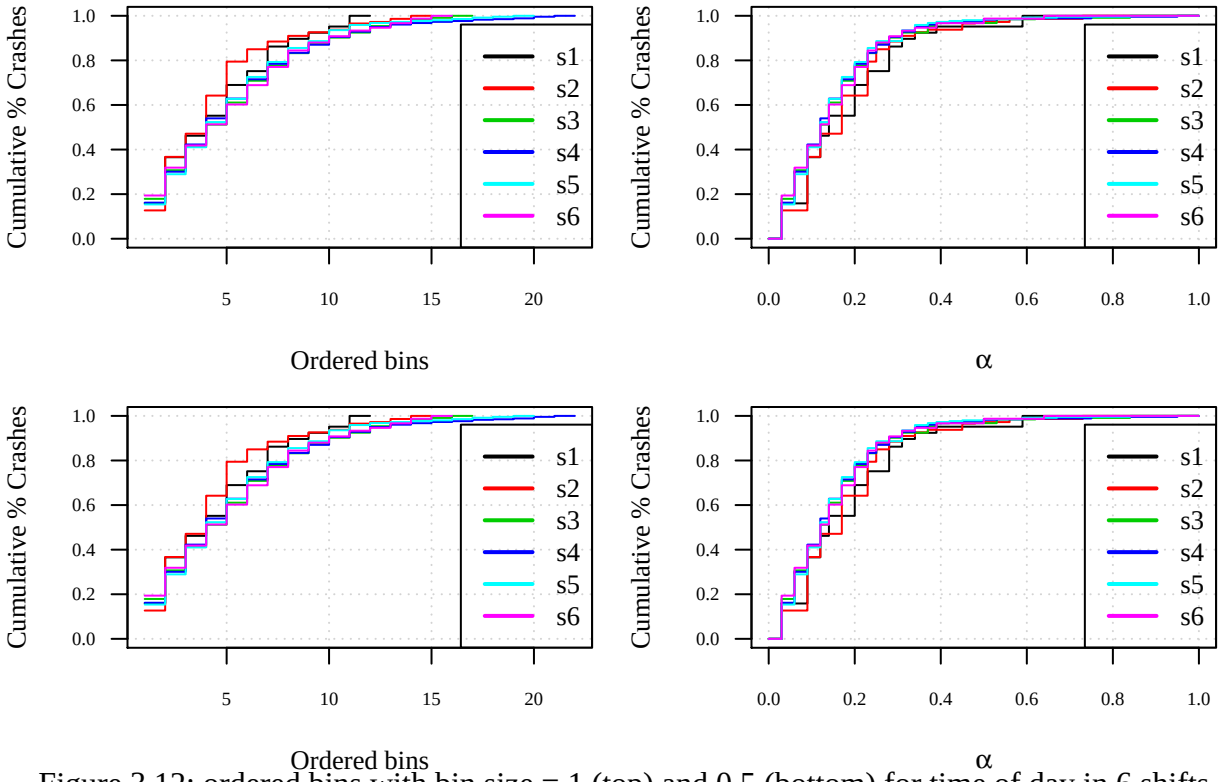

Figure 3.12: ordered bins with bin size = 1 (top) and 0.5 (bottom) for time of day in 6 shifts

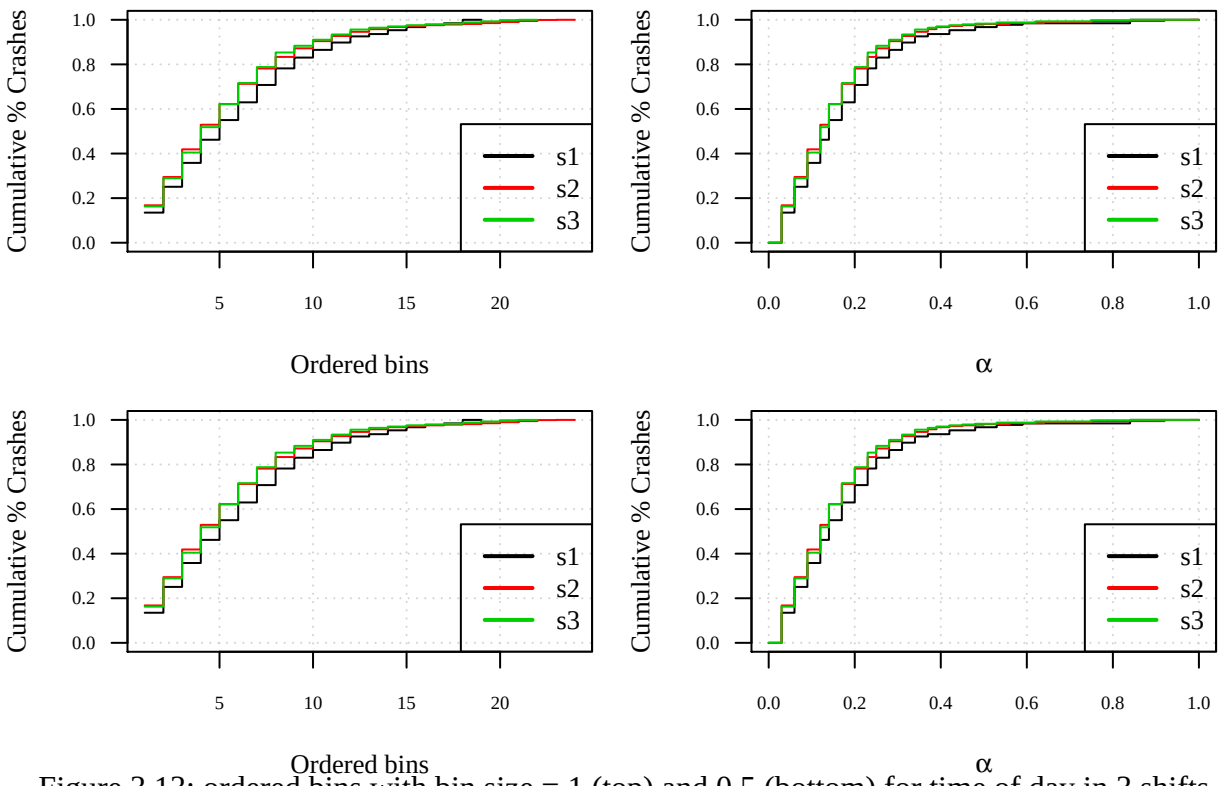

Figure 3.13: ordered bins with bin size = 1 (top) and 0.5 (bottom) for time of day in 3 shifts

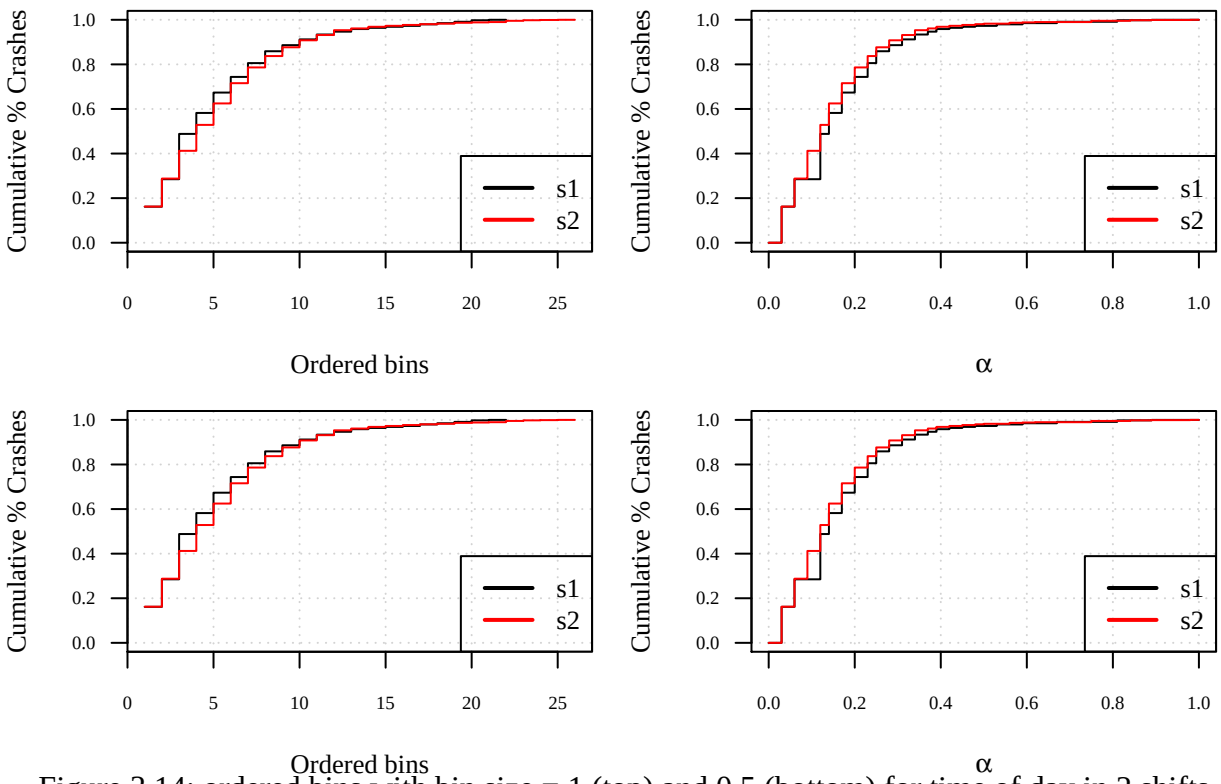

Figure 3.14: ordered bins with bin size = 1 (top) and 0.5 (bottom) for time of day in 2 shifts

3.3 Predictive performance

After a variety of scenarios are compared and tools of surveillance plots and level levets are implemented, similarities and predictability that are discovered during the evaluation process will be tested in this section. A set of data, e.g. weekdays, which is considered more predctive, will be established as the baseline of a pseudo model, then the other set, e.g. weekend, will be tested based on this model.

3.3.1 Predictability on time variation

Figure 3.15 shows the performance using one of the datasets, weekdays and weekend, as model, and the other as testing data. Using weekdays as model (black) has constantly higher predictability than using weekend as model (blue).

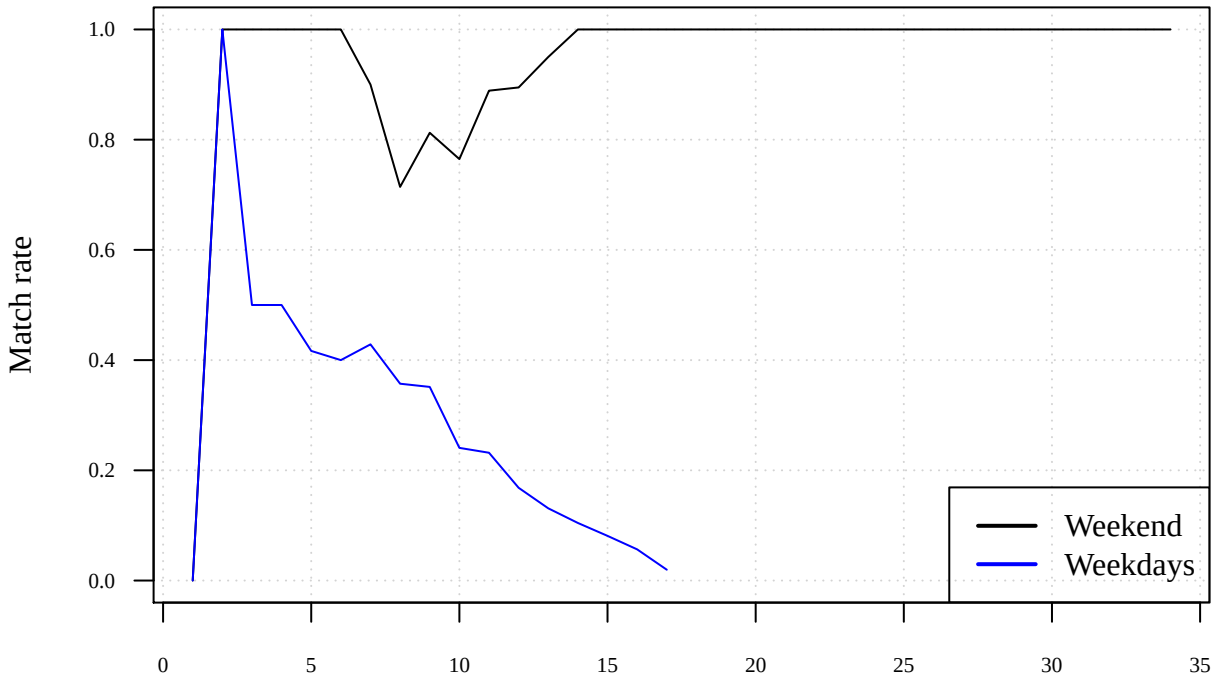

Figure 3.15: Match rates for weekdays and weekend

For days of week (Figure 3.16 and Table 3.18), Monday through Wednesday are combined and used as a model based on the level sets surveillance and KS results; Thursday through Sunday are compared to the reference model. For each testing datasets, Sunday is the best under Mon-Wed model, 10 segments with highest crash rates can be captured; Friday is the worst but still can be all captured using Mon-Wed model while up to 22 segments with highest crash rates on Friday are compared.

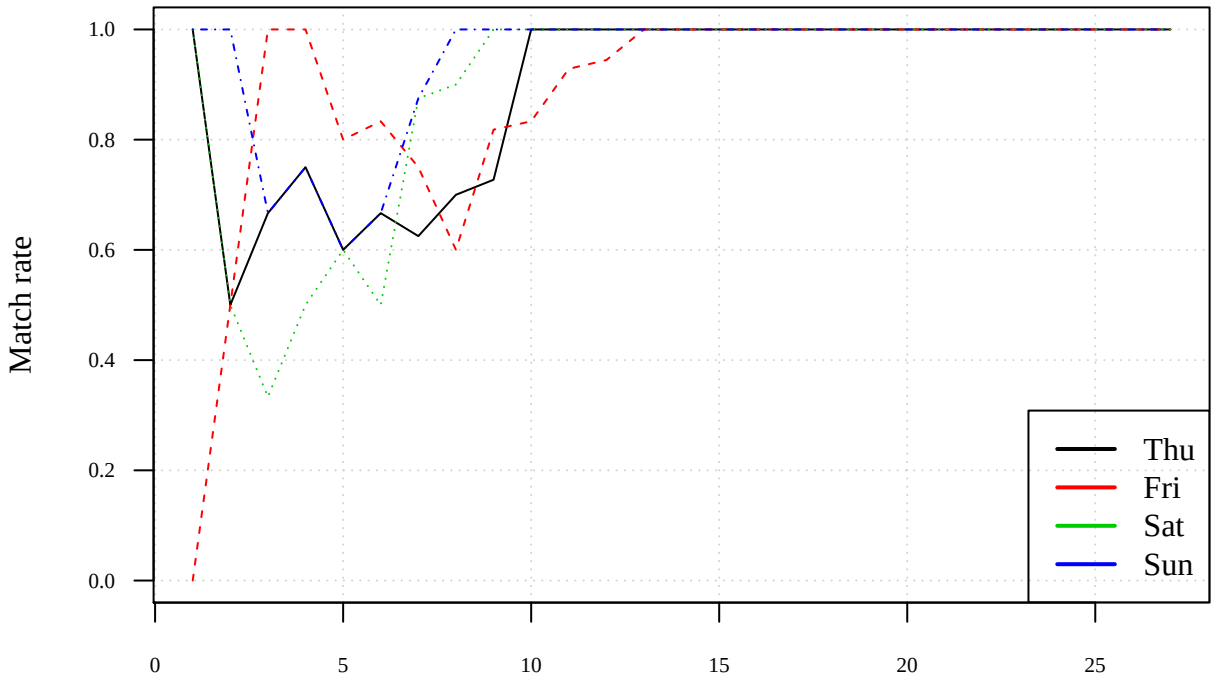

Figure 3.16: Predictive performace using Mon−Wed model

Table: Table 3.18: match rate table for Mon-Wed model

ref.rank thu thu.n fri fri.n sat sat.n sun sun.n ——— ———- —— ———- —— ———- —— ———- —— 1 1.0000000 1 0.0000000 0 1.0000000 1 1.0000000 1 2 0.5000000 1 0.5000000 1 0.5000000 1 1.0000000 2 3 0.6666667 2 1.0000000 3 0.3333333 1 0.6666667 2 4 0.7500000 3 1.0000000 4 0.5000000 2 0.7500000 3 5 0.6000000 3 0.8000000 4 0.6000000 3 0.6000000 3 6 0.6666667 4 0.8333333 5 0.5000000 3 0.6666667 4 7 0.6250000 5 0.7500000 6 0.8750000 7 0.8750000 7 8 0.7000000 7 0.6000000 6 0.9000000 9 1.0000000 10 9 0.7272727 8 0.8181818 9 1.0000000 11 1.0000000 11 10 1.0000000 12 0.8333333 10 1.0000000 12 1.0000000 12 11 1.0000000 14 0.9285714 13 1.0000000 14 1.0000000 14 12 1.0000000 18 0.9444444 17 1.0000000 18 1.0000000 18 13 1.0000000 22 1.0000000 22 1.0000000 22 1.0000000 22 14 1.0000000 25 1.0000000 25 1.0000000 25 1.0000000 25 15 1.0000000 31 1.0000000 31 1.0000000 31 1.0000000 31

For months of year (Figure 3.17 and Table 3.19), semester months (February through May and August through October) are combined and used as a model based on the analyses; rest of the months are compared to the model. January is the best using this model, 7 segments with highest crash rates on January can be captured; June, July, November and December are relatively predictive. Top 10 segments on June and July, 9 and 8 segments on November and December respectively, can be all captured.

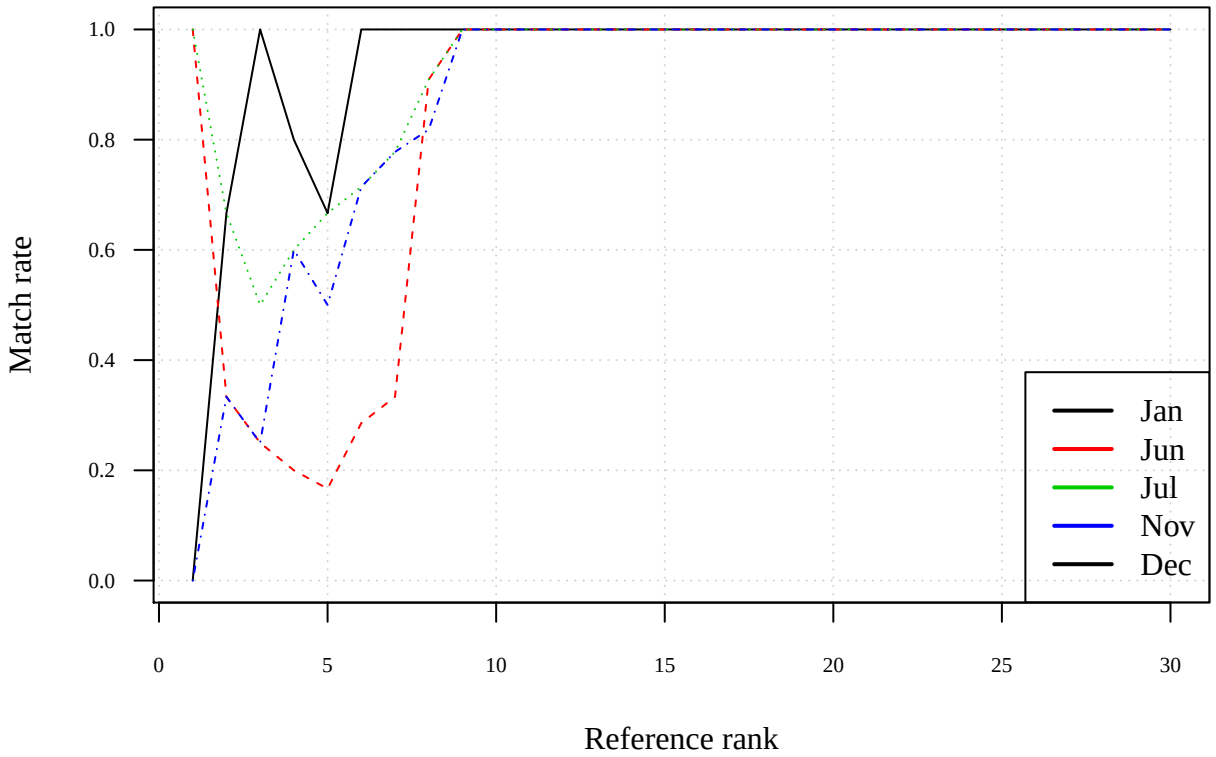

Figure 3.17: Predictive performace using Feb−May and Aug−Oct model

Table: Table 3.19: match rate table for Feb-May and Aug-Oct model

ref.rank jan jan.n jun jun.n jul jul.n nov nov.n dec dec.n ———— ————- ———— ————- ———— ————- ———— ————- ———— ————- 1 0.0000000 0 1.0000000 1 1.0000000 1 0.0000000 0 0.0000000 0 2 0.6666667 2 0.3333333 1 0.6666667 2 0.3333333 1 0.0000000 0 3 1.0000000 4 0.2500000 1 0.5000000 2 0.2500000 1 0.2500000 1 4 0.8000000 4 0.2000000 1 0.6000000 3 0.6000000 3 0.6000000 3 5 0.6666667 4 0.1666667 1 0.6666667 4 0.5000000 3 0.6666667 4 6 1.0000000 7 0.2857143 2 0.7142857 5 0.7142857 5 0.7142857 5 7 1.0000000 9 0.3333333 3 0.7777778 7 0.7777778 7 0.7777778 7 8 1.0000000 11 0.9090909 10 0.9090909 10 0.8181818 9 0.7272727 8 9 1.0000000 12 1.0000000 12 1.0000000 12 1.0000000 12 1.0000000 12 10 1.0000000 15 1.0000000 15 1.0000000 15 1.0000000 15 1.0000000 15

Time of day can subset into 2-hour, 4-hour and 8-hour periods (Figure 3.18, Table 3.20-22). The base models are established based on level sets and KS results described in previous sections. In 2-hour periods, 12:00-18:00 are used for model. Except for 00:00-06:00 period with rarity of crashes that can be all captured as expected, up to top 11 crash rate segments can be all inclusive for the other 2-hour periods. In 4-hour periods, 04:00-20:00 periods are used as a model, and the other 2 4-hour periods will be compared to reference. Similar to 2-hour periods, top 14 segments from 20:00-24:00 will match model predictions. In 8-hour periods, 08:00-24:00 are used as a model and all 9 segments are matched using this model.

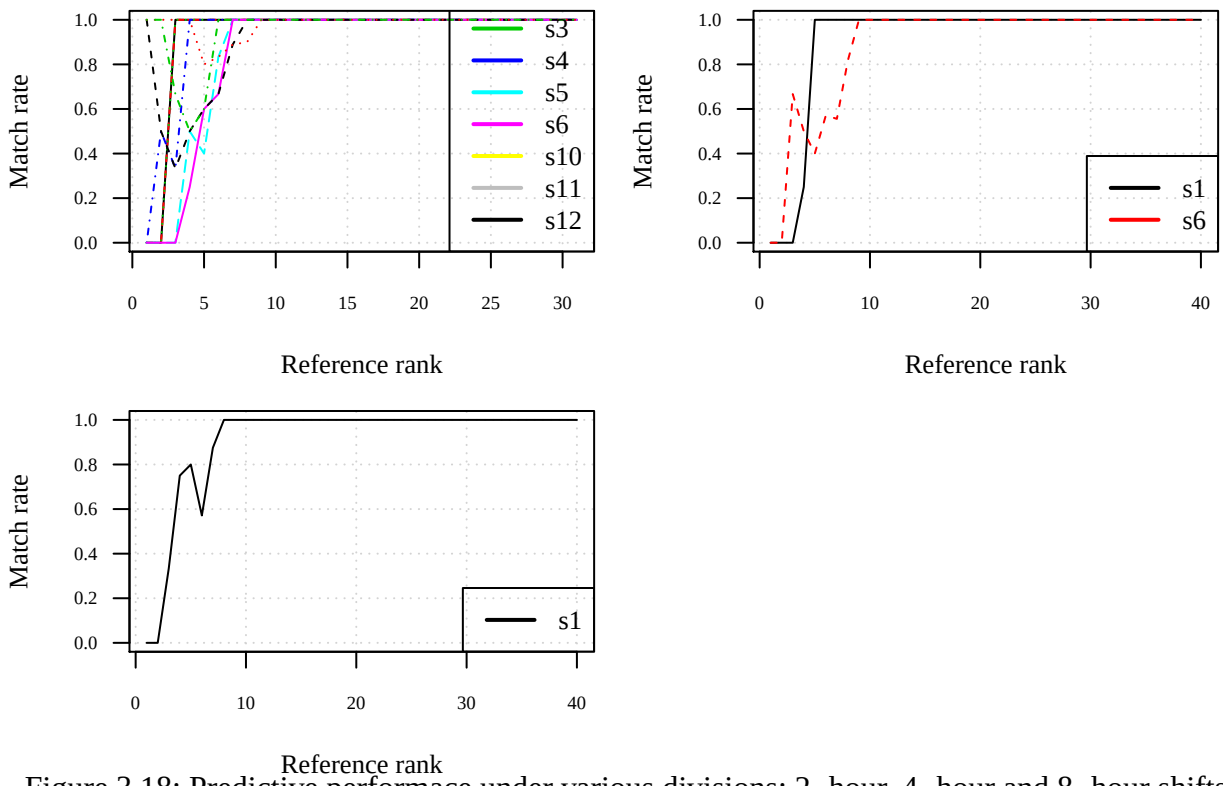

Figure 3.18: Predictive performace under various divisions: 2−hour, 4−hour and 8−hour shifts

<———>

Table 12: Table 3.20: 2-hour match rate table for 12:00-18:00 model

ref.rank	s1	s1.n	s2	s2.n	s3	s3.n	s4	s4.n	s5	s5.n	s6	s6.n	s10	s10.n	
1	0	0	0	0	0	0	0.0000000	0	0.0000000	0	0.0000000	0	1.0000000	1	1
2	0	0	0	0	0	0	0.5000000	1	0.0000000	0	0.0000000	0	0.5000000	1	1
3	1	3	1	3	1	3	0.3333333	1	0.0000000	0	0.0000000	0	0.3333333	1	1
4	1	4	1	4	1	4	1.0000000	4	0.5000000	2	0.2500000	1	0.5000000	2	1
5	1	5	1	5	1	5	1.0000000	5	0.4000000	2	0.6000000	3	0.6000000	3	0
6	1	6	1	6	1	6	1.0000000	6	0.8333333	5	0.6666667	4	0.6666667	4	0
7	1	9	1	9	1	9	1.0000000	9	1.0000000	9	1.0000000	9	0.8888889	8	0
8	1	10	1	10	1	10	1.0000000	10	1.0000000	10	1.0000000	10	1.0000000	10	0
9	1	11	1	11	1	11	1.0000000	11	1.0000000	11	1.0000000	11	1.0000000	11	1
10	1	12	1	12	1	12	1.0000000	12	1.0000000	12	1.0000000	12	1.0000000	12	1

Table 13: Table 3.21: 4-hour match rate table for 04:00-20:00 model

ref.rank	s1	s1.n	s6	s6.n
1	0.00	0	0.0000000	0
2	0.00	0	0.0000000	0
3	0.00	0	0.6666667	2
4	0.25	1	0.5000000	2
5	1.00	5	0.4000000	2
6	1.00	7	0.5714286	4

ref.rank	s1	s1.n	s6	s6.n
7	1.00	9	0.5555556	5
8	1.00	11	0.8181818	9
9	1.00	14	1.0000000	14
10	1.00	15	1.0000000	15

Table: Table 3.22: 8-hour match rate table for 08:00-24:00 model

ref.rank s1 s1.n ——— ———- —— 1 0.0000000 0 2 0.0000000 0 3 0.3333333 1 4 0.7500000 3 5 0.8000000 4 6 0.5714286 4 7 0.8750000 7 8 1.0000000 9 9 1.0000000 10 10 1.0000000 12

3.3.2 Predictability on segment variation

Different bin sizes of Logmile, 1 and 0.5, are compared to the model used similar distributions given different time periods. For the KS results within all groups are smiliar to each other, a model used all data is constructed and all datasets separated by different periods (e.g. Sunday through Saturday) are used as testing sets. For the other manipulation is that similar groups are combined and referenced as model, and the other groups will be testing using reference model.

For weekdays and weekend (Figure 3.19), weekdays are first as model, and weekend is tested (black); then weekend is used as model and then testing weekdays (blue). Two bin sizes, 1 (left) and 0.5 (right) are implemented. It appears that both sets are fairly predictive except for weekend model with bin size 0.5 for the reasons of the classifications of observations at boundaries and rarity of crashes on weekend.

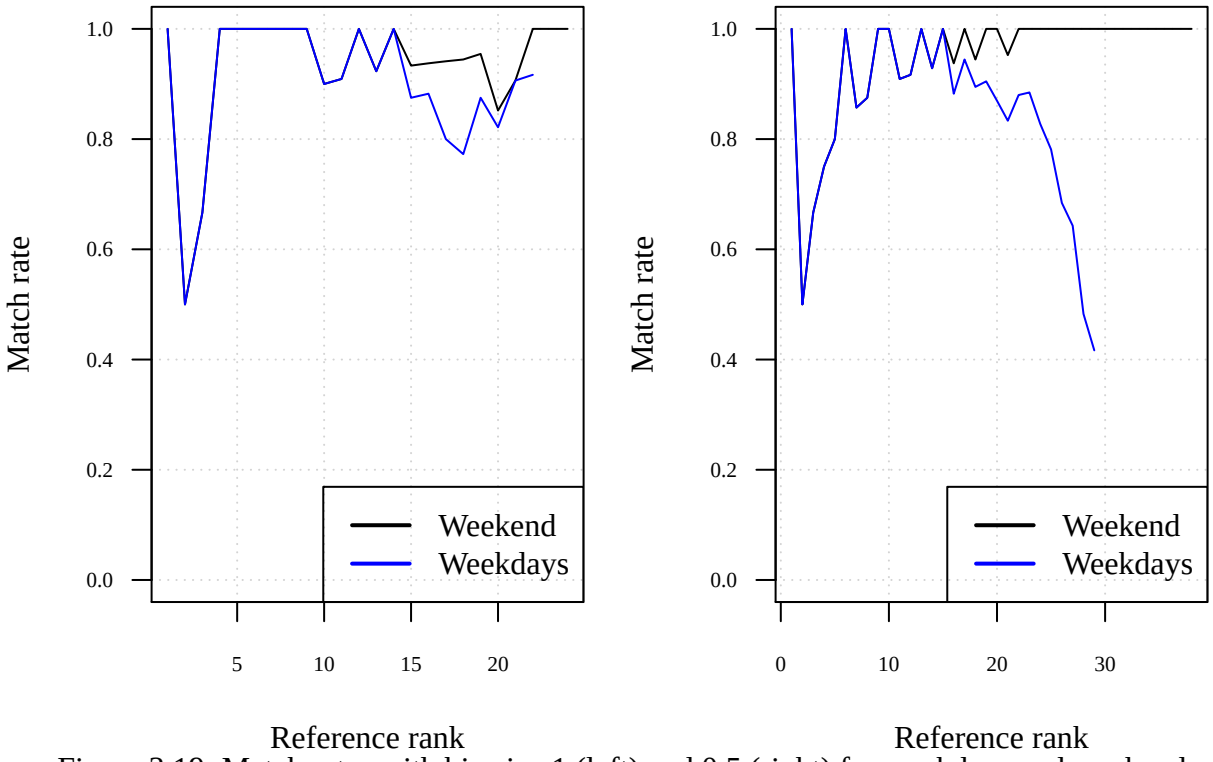

Figure 3.19: Match rates with bin size 1 (left) and 0.5 (right) for weekdays and weekend

Days of week (Figure 3.20) with bin sizes 1 (left) and 0.5 (right) are manipulated. For size 1, Monday through

Saturday are insignificant compared to Sunday in accordance with KS results in Section 3.2.2 and thus combined together and built as model. Sunday is then tested. For size 0.5, days of week are all comparatively insignificant and thus whole week is used as a model and each day is then tested. For Sunday, it appears to have similar predictability for both sizes.

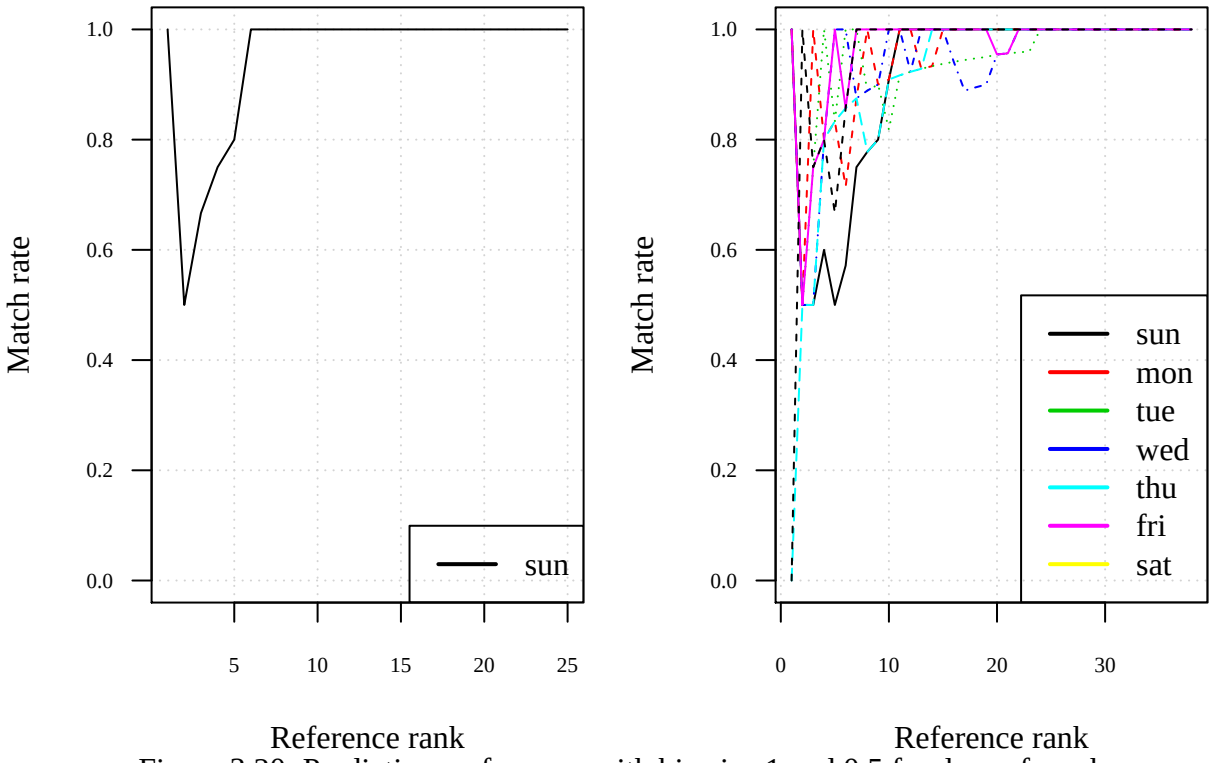

Figure 3.20: Predictive performace with bin size 1 and 0.5 for days of week

Months of year (Figure 3.21) with bin size 1 (left) and 0.5 (right) are compared. For size 1, only February and April are distinct from each other and therefore are used as testing data. For size 0.5, January, March, September and December are combined and served as a reference, the rest of months are compared to it.

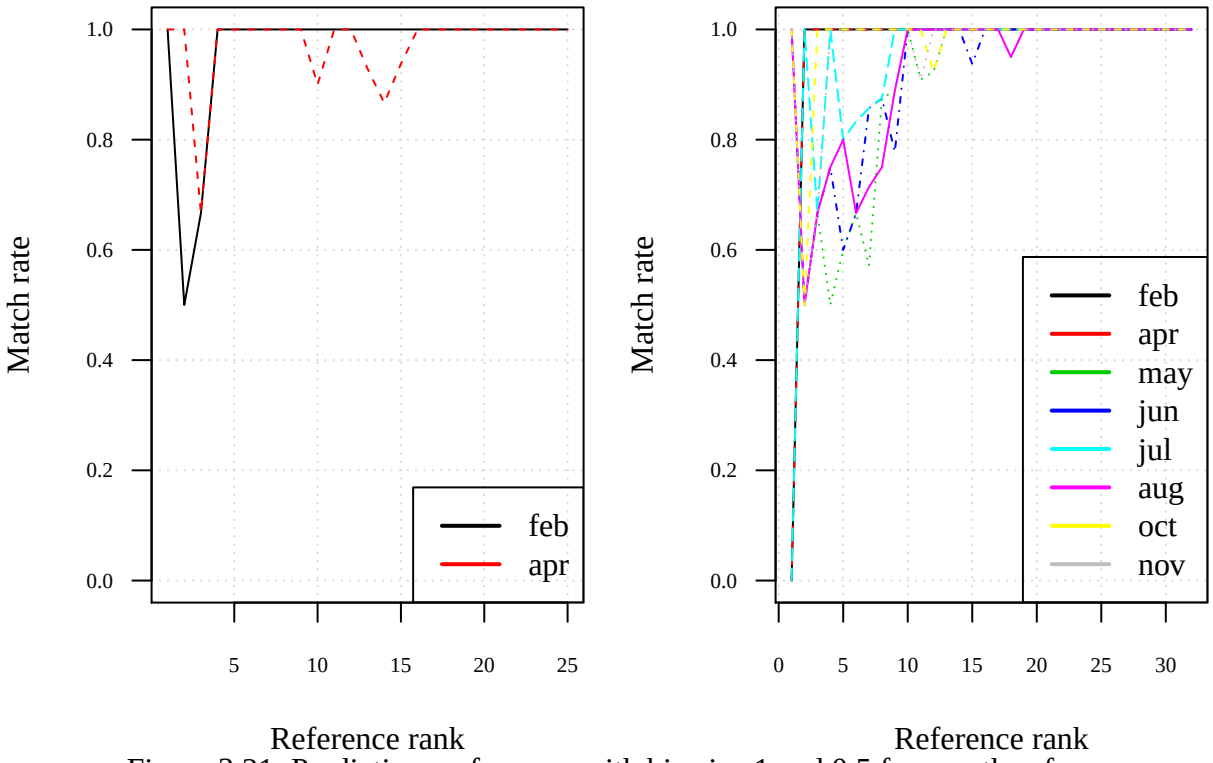

Figure 3.21: Predictive performace with bin size 1 and 0.5 for months of year

Time of day with bin size 1 and 0.5 for 2-hour, 4-hour, 8-hour divisions are showed in Figure 3.22. For 2-hour division (top), 00:00-02:00, 06:00-18:00 and 22:00-24:00 with size 1 are built as model whereas every periods with size 0.5 except for 00:00-04:00 are used. For 4-hour division (middle), 00:00-04:00 and 08:00-20:00 with size 1 are combined whereas all with size 0.5 are included and retested. For 8-hour division, all periods with each size are combined then retested based on the combined model.

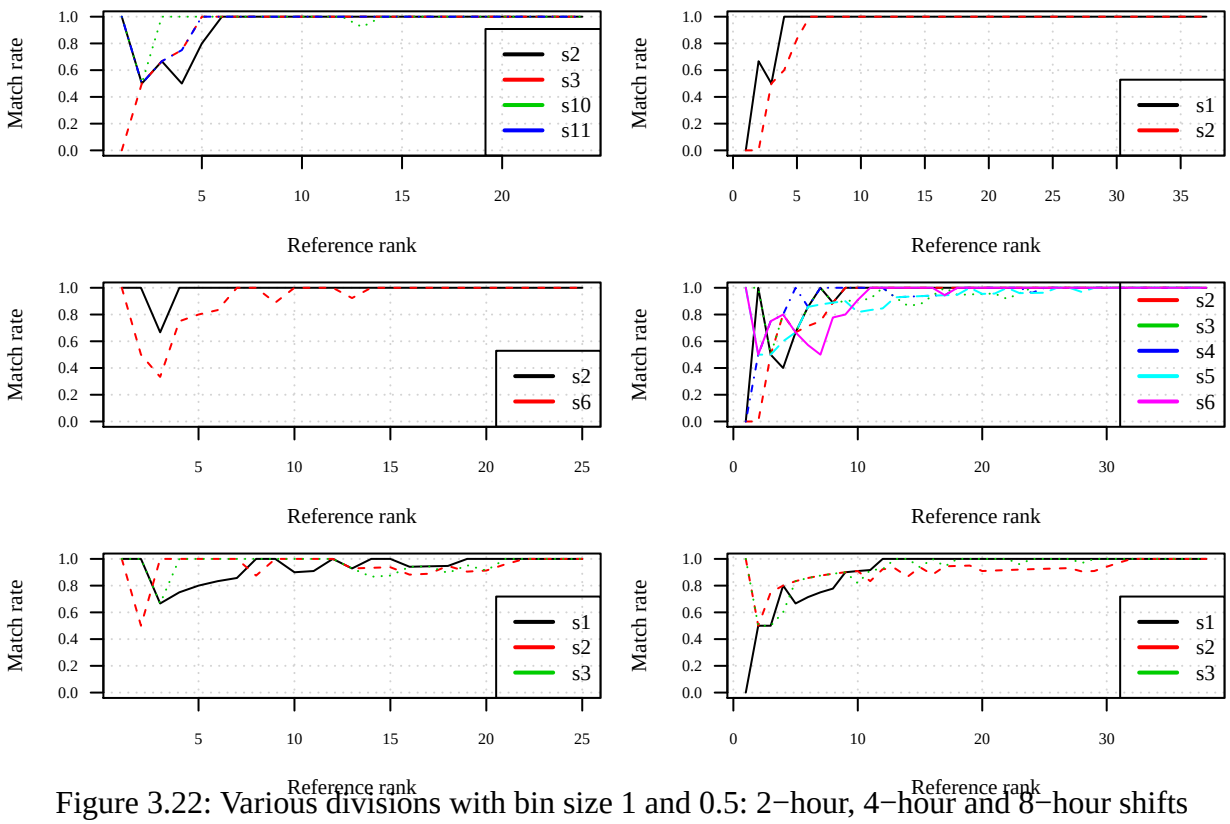

Figure 3.22: Various divisions with bin size 1 and 0.5: 2−hour, 4−hour and 8−hour shifts