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

1. Introduction

2. Evaluation Methodology

Instead of going through traditional procedure of building statistical models such as negative-binomial and Poisson regression, couple useful concepts and tools of evaluating data before model building are introduced and modified to meet the need for the mothodology in this paper. These concepts include the level sets (Section 2.1), surveillance plots (Section 2.2), the Jaccard Index (Section 2.3) and Kolmogorov-Smirnov test (Section 2.4) and will be illustrated with a dataset from highway data in this order of comcepts.

2.1 Level sets

Level sets 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 sets is used as a measure of pattern recognitions for discrete space crashes analysis in this study ([1] and [2]).

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:

$$\gamma = \{x \in \Omega | f(x) \geq c\} \tag{1}$$

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

Some drawbacks of using level sets 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 1 demonstrates a comparison of crashes given different time periods under level sets 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, (gray 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.

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 ([3] and [4]).

In conjuction with level sets, 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

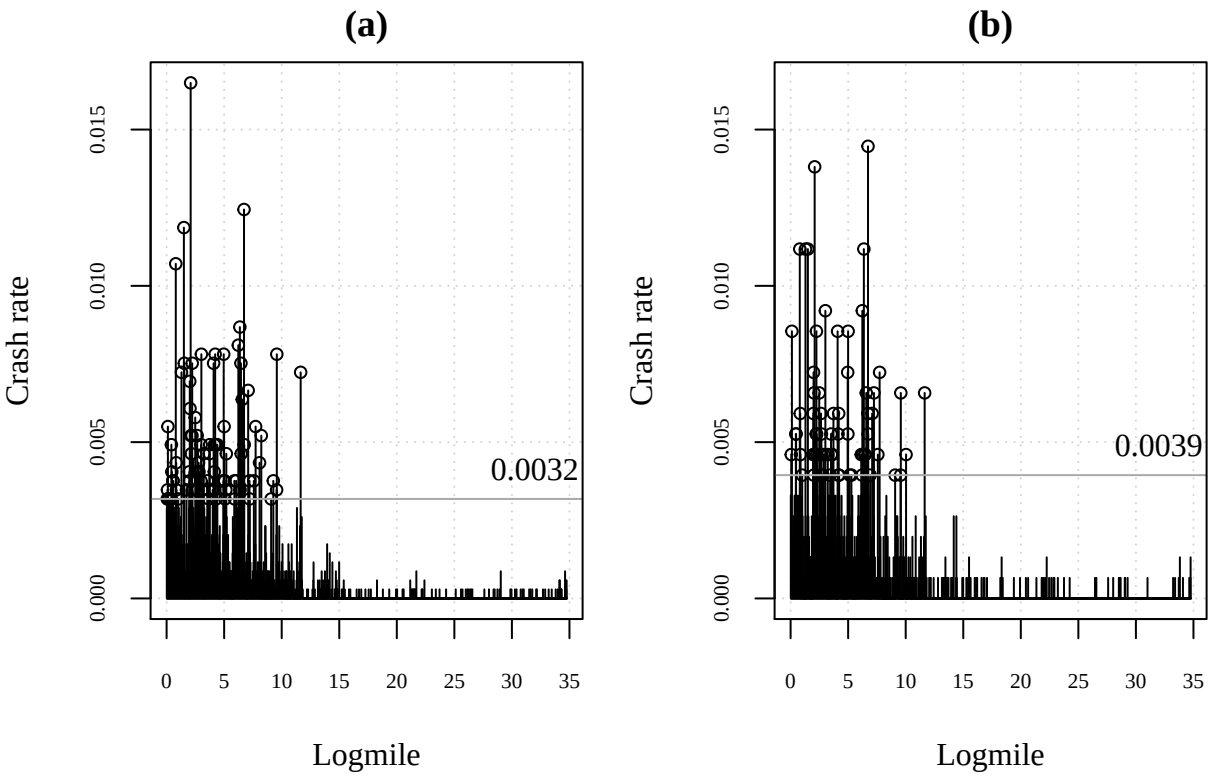

Figure 1: weekdays (a) and weekend (b) with $\alpha = 0.05$

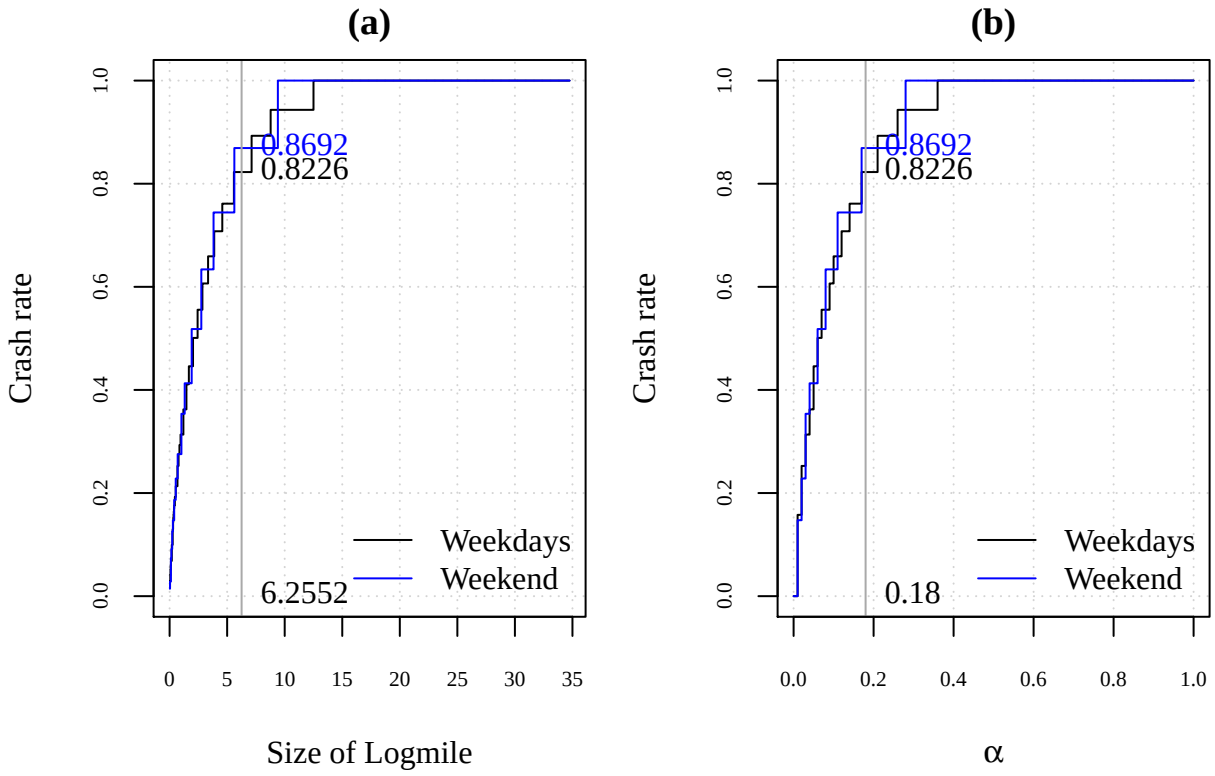

Figure 2: surveillance plots for weekdays and weekend under distinct perspectives α levels

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.

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 1 is also implemented in this example. Instead of single α level is observed in Figure 1, a sequence of α levels (from 0 to 1 with a increment of 0.01) are now concerned and showed in Figure 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 figure in detail.

Table 1 is obtained with arrangement by order and shows the composition for 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 (b), in analog of Figure 2 (a), will be showed simultaneously in the parentheses. The gray 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: 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
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 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. Gray 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.

2.3 Jaccard Index plot

The Jaccard Index, or Jaccard similarity coefficient was introduced by Paul Jaccard and use as a measure of similarity of data [5]. The Jaccard Index can be also extended to a Jaccard distance for measure of similarity in multidimensional space. Suppose two samples are drawn, it is defined as a fraction of intersection and union sets of two samples.

$$J(A,B) = \frac{\|A \cap B\|}{\|A \cup B\|} \tag{2}$$

where A and B represent two samples, $\|\cdot\|$ is the size of a sample.

In this paper similar idea of Jaccard Index is used but in a conditional structure. It is defined as the following:

$$J^*(A,B) = \frac{\|A \cap B\|}{\|A\|} \tag{3}$$

where $J^*(A,B)$ is conditioning on A. The diversity of sets A and B can be observed by comparing $J^*(A,B)$ and $J^*(B,A)$. The indexes should be close if both sets are similar, and vice versa.

Figure 4 shows the cumulative crash rate in original segments using (3). In 4 on the left, the black line is the exact cumulative crash rate on weekdays whereas the blue line is the cumulative crash rate with matched segments using reference weekdays, and right side shows the opposite condition.

Based on the adjustment (3), J^* can be calculated under a sequence of α levels is showed in Figure 5. On the right of Figure 5, under the same α level, the black line is the true cumulative crash rate in weekdays whereas the blue line

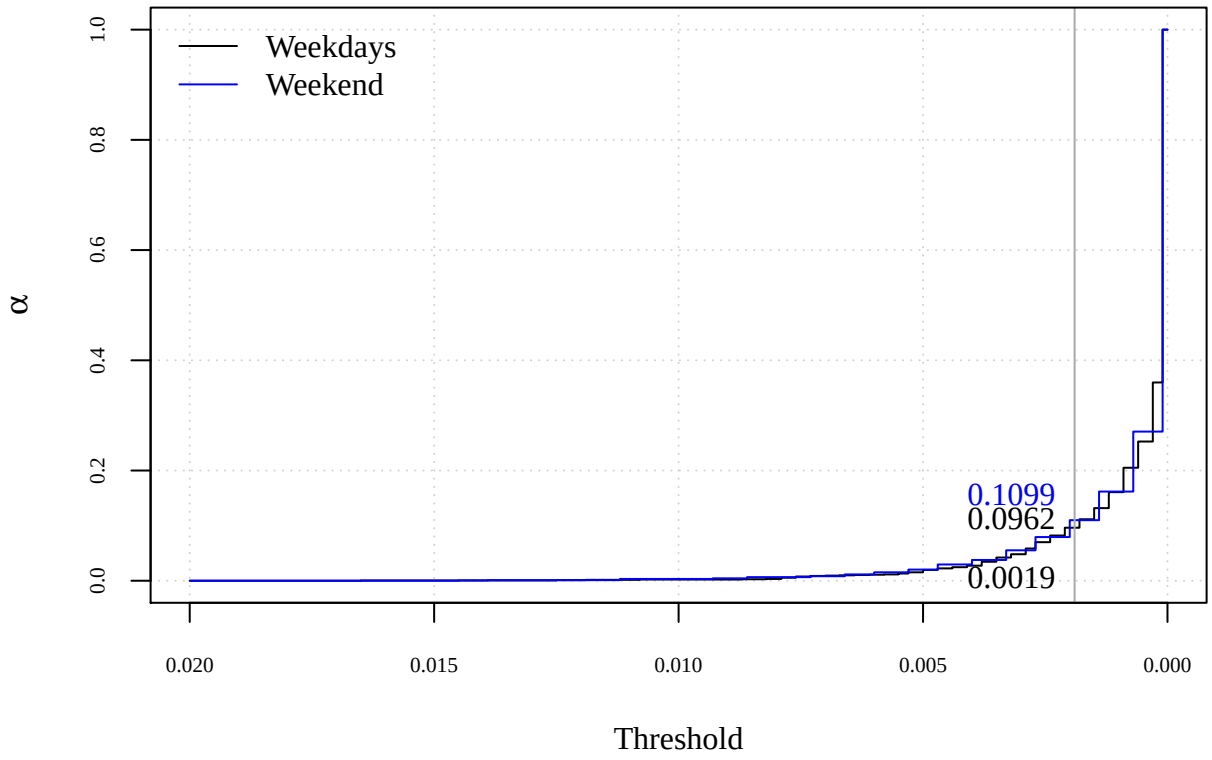

Figure 3: threshold based surveillance for weekdays and weekend

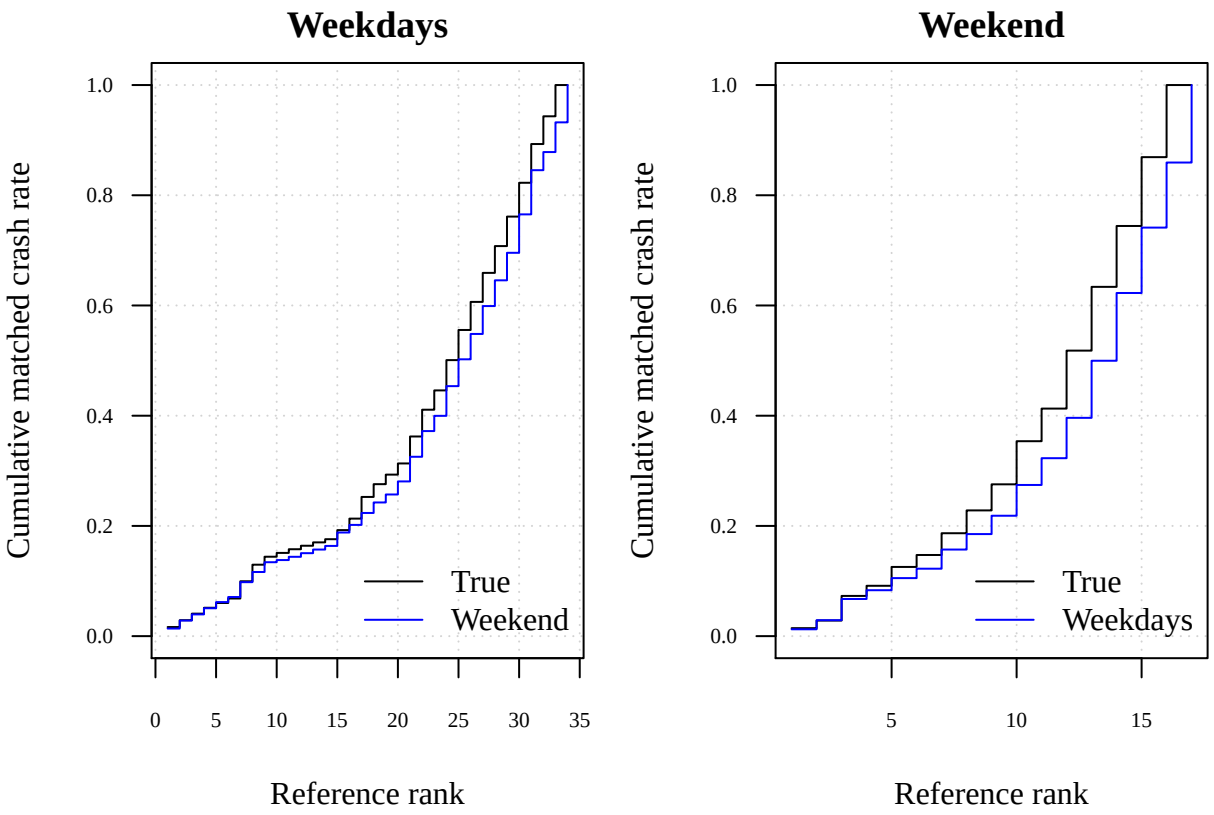

Figure 4: Comparison for weekdays and weekend using J^* Index

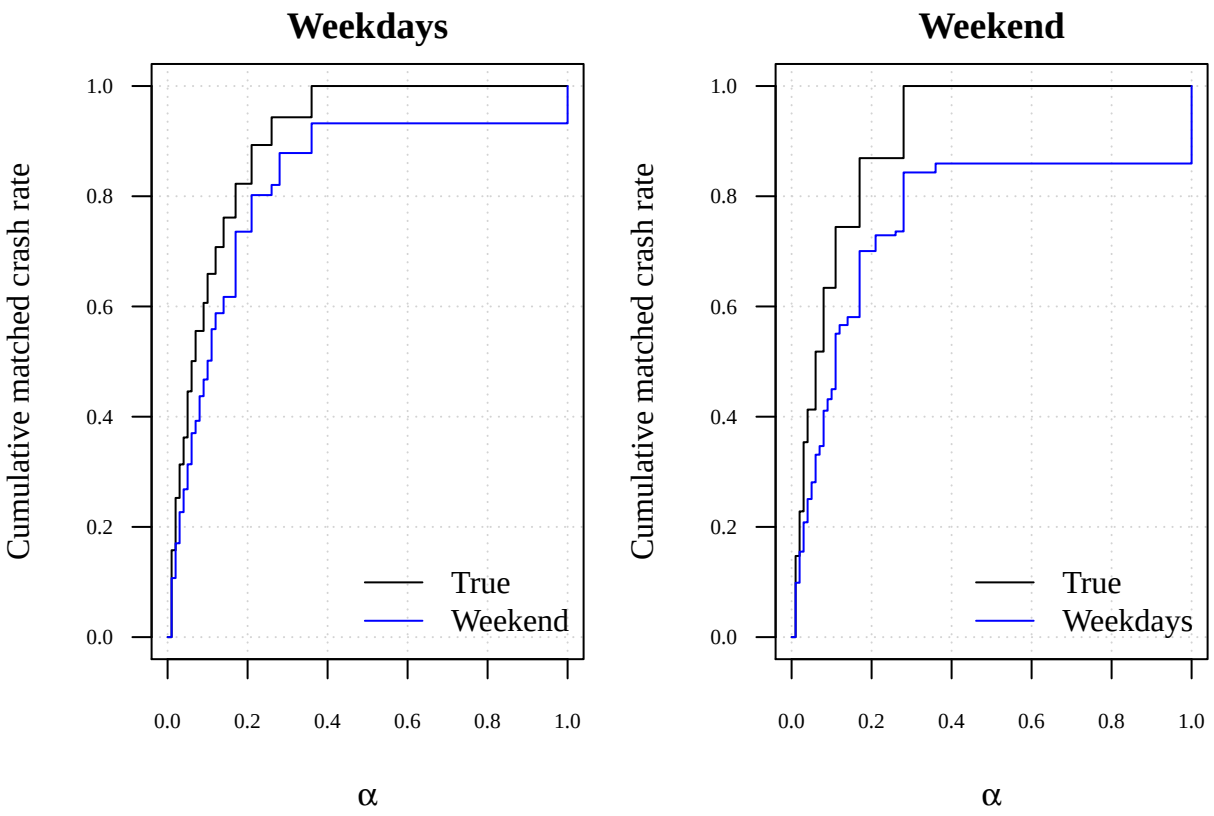

Figure 5: Comparison for weekdays and weekend under the level sets using J^* Index

shows the cumulative crash rate on weekend with the matched segments in weekdays, and the condition is opposite for the left side of Figure 5.

Figure 4 and Figure 5 demonstrate the relationships of ordered distributions between weekdays and weekend under distinct aspects. Under the level sets setting, it is comparatively objective for comparisons for the reason that sizes of reference and subject are approximately equivalent. Great dissimilarities in distributions of ordered segments may be clearly showed in Figure 5. In the next section, a metric to describe this similarity will be introduced and also served as a measure of quality of data.

2.4 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 ([6]). 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:

$$S_n(x) = \sum_{i=1}^{n} \mathbf{I}(x_i \leq x)/n \tag{4}$$

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:

$$D_n = \sup_x |S_n(x) - F_X(x)| \tag{5}$$

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

$$D_{m,n} = \max_x |S_m(x) - S_n(x)| \tag{6}$$

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. Instead of calculating the *cdf* in using a raw data, it is done in a ranking way and *cdf* is not based on 4 but summing up the crash rates with the same rank. It is defined as follow:

$$S_k^*(y) = \sum_{i=1}^{k} p(Y \geq y_{(i)}) \tag{7}$$

where $S_k^*(y)$ is the cumulative crash rate, $0 \leq S_k^*(y) \leq 1$; k represents the rank; Y is the ordered segments and $y_{(i)}$ is the i^{th} segments with the highest crash rate (ties are also considered). Since $y_{(1)}$ determines the maximum, $S_k^*(y)$ is showed as a way of survival function. (7) is viewed as foundation of the level sets applied in KS test.

A modified KS test using (7) is contructed based on the snenarios used from a surveillance perspective using the level sets; therefore, this is viewed as a solid metric to determine the diversity in distributions under different scenarios.

2.4.1 Demonstration of conducting KS test

Recall from Section 2.2, surveillance using level sets 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 sets repectively (see Figure 5). Aussume significance level is 0.05, then, $D_{week} = 0.1631$ and $D_{weekend} = 0.2229$ ($d_0 \approx 0.1911$) with *p-value* 0.1364 and 0.0132 (Figure 6)

Therefore, it is significant in ordered distributions given reference weekend (Table 2). This suggests that under the same size of segments, weekdays as reference to test weekend tends to be similar to true pattern. Also, weekdays as reference is more predictive than weekend as reference.

Table 2: KS results under the level sets in Figure 5

	Week	Weekend
max.d	0.1630506	0.2229139
p.value	0.1364238	0.0132255

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 as a predictive reference (say weekdays instead of weekend in the previous demonstration) to foresee patterns with better performance and precision.

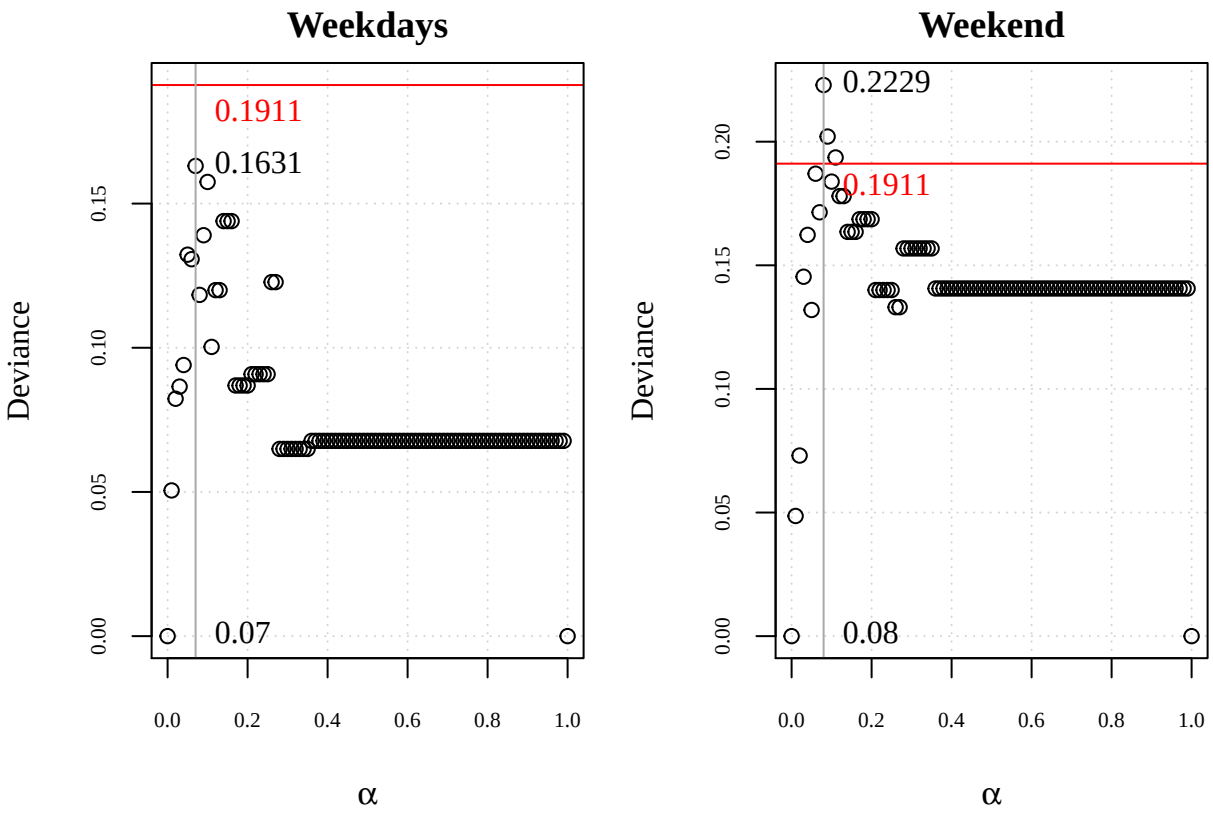

Figure 6: KS test results for weekdays and weekend under the level sets using J^* Index

3. Predictability of Crash Data

In this section, the methods mentioned in previous section are implemented on a variety of scenarios, e.g. time variation and segments varation. These scenarios will be built under level sets, compared using modified Jaccard Index and evaluated with KS test. The visualization of prediction will be showed lastly.

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 ([7]).

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.

3.2 Concentration of crashes: empirical performance measures

In previous demonstrations, crashes on the highway segments given different time periods, weekdays and weekend, are primarily showed a process of measuring data. KS test results are presented that weekdays may consider more predictive than weekend, i.e., a model can be built using weekdays set, and this model may also capture a similar pattern for weekend set. Other characteristrics of causes of crashes, e.g. time variation (Section 3.2.1) and segment variation (Section 3.2.2), will be compared through the same process as demonstrations earlier to uncover potential relationships in ordered distributions.

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 at the various significance level. The reason to the various significance level is that crashes distributed on an entire route are rare and more sparse with the consideration of time. Small change in distribution results in tremendous or negligible extremes. Therefore, the choice of significance level is arbituary to an observable extent of relationships in time variation.

Note that the approximation of KS test and rare events condition result in a inconsistent in maximal distance and in *p-value*, so the magnitude of differences for both measures in distributions for months of year can be seen in Table 3 and 4. The month used as reference is showed on the row whereas the compared groups are showed on the column. Cells in the table highlights in red to show the behavior under the significance level of 0.3. If only consider maximal distance in distributions ($D_{0.3} = 0.1722$ for this setting), May, June and July (see from row vectors) which only have one significant pair under level sets individually, are appear to have better performance than any other months as reference.

Days of week are also of interest to be compared and showed from Table 5 using maximal distance and 6 using *p-value* with a significance level of 0.3. The day used as a reference is showed on the row whereas compared days are showed on columns. Each day except for Sunday has one pair with a significant result. In a similar manner, Sunday appears to be more predictve while using as a reference. If only to predict from Monday to Saturday, each day during this period may also share the similar performance.

Two primary time separations, days of week and months of year, are ways of using one data to make predictions of the other data under the level sets. One other possibility is to separate time of day into different time periods.

Table 3: KS test: max.d using level sets for months of year with significance level = 0.3

	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec
1	0	0.4123	0.4185	0.4031	0.4492	0.4031	0.4492	0.3754	0.4092	0.3292	0.4154	0.3723
2	0.4512	0	0.4186	0.4395	0.4302	0.4209	0.3977	0.4326	0.4302	0.4023	0.4209	0.4326
3	0.4548	0.401	0	0.3863	0.4303	0.4205	0.4523	0.3961	0.4621	0.4254	0.4401	0.4499
4	0.4442	0.4264	0.3731	0	0.3756	0.4213	0.4594	0.3959	0.4365	0.3858	0.4239	0.3756
5	0.4716	0.4074	0.3901	0.3704	0	0.4123	0.4222	0.4025	0.3975	0.3778	0.3877	0.363
6	0.4037	0.3958	0.4116	0.3826	0.4037	0	0.4591	0.4274	0.3879	0.3483	0.3879	0.3931
7	0.4539	0.4078	0.4248	0.4078	0.4029	0.3981	0	0.3908	0.4029	0.3932	0.4126	0.3956
8	0.4089	0.4089	0.4416	0.3855	0.4206	0.4322	0.4533	0	0.4182	0.3598	0.4206	0.3692
9	0.4368	0.4296	0.4439	0.3604	0.3795	0.4057	0.4248	0.3986	0	0.4081	0.3723	0.3914
10	0.4201	0.4289	0.4858	0.4179	0.3982	0.4201	0.4376	0.442	0.4551	0	0.442	0.3589
11	0.409	0.4202	0.4427	0.3573	0.3955	0.3551	0.4472	0.364	0.4067	0.3551	0	0.3573
12	0.4367	0.4325	0.4367	0.3987	0.3671	0.4008	0.4262	0.3734	0.4135	0.3397	0.4114	0

Table 4: KS test: p.value using level sets for months of year with significance level = 0.3

	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec
1	1	0.5133	0.4136	0.5452	0.398	0.4634	0.3253	0.6478	0.443	0.8403	0.3562	0.5001
2	0.4811	1	0.4132	0.4264	0.4549	0.406	0.4818	0.4477	0.378	0.5478	0.34	0.3079
3	0.4702	0.5526	1	0.6061	0.4546	0.4072	0.3172	0.5701	0.2927	0.4701	0.2883	0.2643
4	0.5027	0.467	0.5714	1	0.6468	0.4048	0.2993	0.5706	0.3599	0.608	0.3317	0.4878
5	0.4216	0.5301	0.5083	0.6675	1	0.433	0.402	0.5473	0.4823	0.6385	0.445	0.5357
6	0.6391	0.5712	0.4353	0.6201	0.543	1	0.3	0.4637	0.5164	0.7578	0.4443	0.4264
7	0.4729	0.5289	0.3943	0.5289	0.5458	0.4805	1	0.5895	0.464	0.5806	0.3644	0.4181
8	0.6206	0.525	0.3458	0.6091	0.4859	0.3722	0.3148	1	0.4143	0.7099	0.3411	0.5119
9	0.5262	0.4569	0.3395	0.7076	0.632	0.4546	0.3941	0.5612	1	0.5277	0.5001	0.4322
10	0.5813	0.4591	0.2391	0.4945	0.5623	0.4084	0.3568	0.419	0.31	1	0.2835	0.5517
11	0.6202	0.487	0.3428	0.7202	0.5722	0.6431	0.3307	0.6928	0.4512	0.7295	1	0.5579
12	0.5263	0.4479	0.3594	0.5606	0.6805	0.471	0.3901	0.6555	0.4293	0.7947	0.3681	1

Table 5: KS test: max.d using level sets for days of week with significance level = 0.3

	sun	mon	tue	wed	thu	fri	sat
1	0	0.2527	0.2473	0.288	0.2473	0.2011	0.3125
2	0.4565	0	0.2436	0.2664	0.2195	0.2195	0.3052
3	0.4585	0.2349	0	0.2525	0.2123	0.2161	0.3128
4	0.4668	0.2544	0.2517	0	0.2273	0.2016	0.3139
5	0.4839	0.2481	0.2481	0.2556	0	0.2323	0.3337
6	0.4601	0.2508	0.2327	0.255	0.2412	0	0.322
7	0.4897	0.2293	0.269	0.2621	0.2552	0.2293	0

Table 6: KS test: p.value using level sets for days of week with significance level = 0.3

	sun	mon	tue	wed	thu	fri	sat
1	1	0.9907	0.9032	0.739	0.9602	1	0.7532
2	0.3066	1	0.9245	0.8534	1	0.9706	0.7878
3	0.3014	1	1	0.9305	1	0.9928	0.7517
4	0.2813	0.9814	0.8778	1	1	1	0.7465
5	0.2432	1	0.8983	0.9133	1	0.8905	0.6566
6	0.2975	1	0.9891	0.9163	0.9948	1	0.7092
7	0.2311	1	0.7809	0.8772	0.9156	0.9088	1

Table 7: KS test: max.d using level sets for 2-hour periods with significance level = 0.3

	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12
1	0	0.8506	0.9195	0.5862	0.5747	0.4828	0.3449	0.3908	0.3539	0.4598	0.6322	0.7241
2	0.7759	0	0.8448	0.5862	0.6034	0.4828	0.3794	0.4138	0.3482	0.4196	0.5517	0.7586
3	0.9241	0.8861	0	0.5443	0.4937	0.4304	0.4177	0.3714	0.2967	0.3797	0.5696	0.7342
4	0.8644	0.8955	0.8616	0	0.4322	0.3983	0.3249	0.2779	0.277	0.4435	0.596	0.726
5	0.8677	0.8931	0.8321	0.43	0	0.3919	0.2723	0.2595	0.2081	0.3868	0.542	0.715
6	0.8407	0.8952	0.8387	0.3911	0.4052	0	0.2278	0.2177	0.2129	0.3508	0.5605	0.7077
7	0.8358	0.9015	0.8601	0.4538	0.3954	0.337	0	0.2032	0.1825	0.3577	0.5499	0.7117
8	0.8633	0.9089	0.8441	0.4376	0.4161	0.3441	0.229	0	0.2166	0.3933	0.5444	0.7014
9	0.8419	0.8902	0.8479	0.4522	0.4139	0.3766	0.2659	0.2165	0	0.3444	0.566	0.6858
10	0.823	0.8806	0.7974	0.4264	0.3646	0.3731	0.2175	0.2239	0.1812	0	0.5416	0.678
11	0.8512	0.876	0.7934	0.4545	0.438	0.3843	0.3099	0.281	0.2517	0.3967	0	0.6983
12	0.8067	0.9067	0.82	0.4467	0.44	0.4733	0.3	0.2438	0.2067	0.3933	0.5733	0

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/insignificant. 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 (Table 7 and Table 8), all periods have poor performance in predicting from 00:00 to 06:00, which is expected since it is during normal bed time. From 08:00 to 16:00 is comparatively similar, each 2-hour period during this time gap should have close performance, and so is the time from 18:00 to 22:00. 06:00-08:00 and 16:00-18:00 these two periods can be refered as rush hours so that they are diverse to other day time periods. 22:00 until 06:00 should be the night time period that distinct patterns exist in their distributions.

For 4-hour periods, only 00:00-04:00 has better predictability than others based on Table 9. Recall from the inconsistency between *p-value* and maximal distance measures, however, any period between 08:00-24:00 on the other hand has better performance from Table 10. In accordance with the previous 2-hour groups, the result from Table 10 appears to be reasonable.

For 8-hour and 12-hour periods, the diversities between them in each separation are more similar from Table 11 to Table 14.

It can be approximately seen that two measures are actually presented similarly for 2-hour periods, inconsistent in 4-hour periods, negligible for few groups. Hypothetically, the separation for time of day may work better with groups somewhere near 4-hour gap with proper start and end points.

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

Table 8: KS test: p.value using level sets for 2-hour periods with significance level = 0.3

	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12
1	1	0.2283	0.0679	0.128	0.1424	0.1945	0.3782	0.2357	0.3055	0.2984	0.1818	0.086
2	0.3287	1	0.1151	0.128	0.1086	0.1945	0.2666	0.182	0.3246	0.4101	0.322	0.0633
3	0.1544	0.1897	1	0.1869	0.2846	0.3138	0.1738	0.29	0.534	0.5462	0.2855	0.0788
4	0.2126	0.1804	0.1027	1	0.4488	0.4093	0.4564	0.6783	0.6328	0.3406	0.2373	0.0847
5	0.209	0.1827	0.1254	0.4556	1	0.4307	0.7085	0.7789	1	0.5204	0.3432	0.0931
6	0.24	0.1807	0.12	0.5882	0.5376	1	0.9671	1	1	0.6607	0.3037	0.0991
7	0.246	0.1747	0.1037	0.3852	0.5727	0.6425	1	1	1	0.6324	0.3259	0.0958
8	0.2138	0.1678	0.1157	0.4321	0.5007	0.612	0.9597	1	0.9896	0.4971	0.338	0.1045
9	0.2385	0.1856	0.1127	0.3897	0.508	0.4841	0.7435	1	1	0.6877	0.2927	0.119
10	0.2621	0.1953	0.1572	0.4669	0.6905	0.497	1	0.9915	1	1	0.3441	0.1268
11	0.2275	0.2001	0.1613	0.383	0.431	0.4567	0.5212	0.6622	0.7732	0.4852	1	0.1072
12	0.2839	0.1698	0.1358	0.4054	0.425	0.2128	0.5673	0.8703	1	0.497	0.2783	1

Table 9: KS test: max.d using level sets for 4-hour periods with significance level = 0.3

	s1	s2	s3	s4	s5	s6
1	0	0.5241	0.3586	0.3345	0.2736	0.4552
2	0.7945	0	0.261	0.2198	0.229	0.485
3	0.7683	0.3465	0	0.1417	0.1611	0.4308
4	0.7742	0.3889	0.1999	0	0.1501	0.4269
5	0.7524	0.3892	0.2237	0.1387	0	0.4193
6	0.75	0.3827	0.2526	0.1862	0.1876	0

Table 10: KS test: p.value using level sets for 4-hour periods with significance level = 0.3

	s1	s2	s3	s4	s5	s6
1	1	0.1687	0.3758	0.2133	0.52	0.3099
2	0.0852	1	0.8251	0.7613	0.7785	0.2408
3	0.1045	0.679	1	1	1	0.3763
4	0.0999	0.5128	1	1	1	0.3878
5	0.118	0.5117	1	1	1	0.411
6	0.1201	0.5354	0.8728	1	1	1

Table 11: KS test: max.d using level sets for 8-hour periods with significance level = 0.3

	s1	s2	s3
1	0	0.2052	0.1968
2	0.3084	0	0.1104
3	0.3134	0.1005	0

Table 12: KS test: p.value using level sets for 8-hour periods with significance level = 0.3

	s1	s2	s3
1	1	0.5656	0.853
2	0.8495	1	1
3	0.8264	1	1

Table 13: KS test: max.d using level sets for 12-hour periods with significance level = 0.3

	s1	s2
1	0	0.104
2	0.1342	0

Table 14: KS test: p.value using level sets for 12-hour periods with significance level = 0.3

	s1	s2
1	1	1
2	1	1

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 meet the needs. Two choices of binning size 1 and 0.5 will be showed to observe the influences of predictions.

Figure 7 shows the comparisons of weekdays and weekend with different bin sizes under level sets. On the top left, black line is the true crash pattern on weekdays with bin size 1 under level sets whereas blue line is predictive pattern for weekend used weekdays as reference, and so forth. Recall from 5, the diversity between two sets are obvious, yet here it is showed similar patterns between weekdays and weekend with binning. The KS test results for weekdays and weekend with binning are also supportive and showed from Table 15 Table to 18.

Bin size of 1 and 0.5 for days of week are showed from Table 19 to Table 22, respectively. Both results provide the same information in distributions using level sets. With bin size of 1, Wednesday and Thursday have better performance than other days although both have poor prediction on Sunday. With bin size of 0.5, Tuesday along with Wednesday and Thursday are more predictive. *P.value* and *max.d* measures are also consistent under both bin sizes individually.

Considering different bin size 1 and 0.5 for distributions of crashes for months of year from Table 23 to Table 26, differences between months become distinct. For bin size 1, only August appears to have comparatively better performance in predicting crashes in January, April, May, June, September, October, November and December. For bin size 0.5, August still have better prediction compared to others, and have similar pattern with January, February, March, April, June, September and November. August stands out in both sizes but has different predictability in other months.

Considering different bin size 1 and 0.5 for distributions of crashes for time of day from Table 27 to Table 30 for 2-hour periods. 18:00-20:00 and 12:00-14:00 with bin size 1 and 0.5, are comparatively predictive in 00:00-02:00 and 08:00-22:00, and 10:00-16:00 and 18:00-20:00, respectively.

Bin size 1 and 0.5 for distributions of crashes for time of day from Table 31 to Table 34 for 4-hour periods. 08:00-12:00 and 08:00-12:00 or 20:00-24:00 with bin size 1 and 0.5, are comparatively predictive in whole periods except 20:00-24:00, and 12:00-16:00 or 04:00-08:00, respectively.

Bin size 1 and 0.5 for distributions of crashes for time of day from Table 35 to Table 38 for 8-hour periods. 08:00-16:00 or 16:00-24:00 and 16:00-24:00 with bin size 1 and 0.5, are comparatively predictive in whole periods, and 00:00-08:00, respectively.

Lastly, bin size 1 and 0.5 for distributions of crashes for time of day from Table 39 to Table 42 for 12-hour periods.

Table 15: KS test: max.d using level sets for weekdays and weekend with size 1 and significance level = 0.3

	wk	wkd
1	0	0.1352
2	0.1184	0

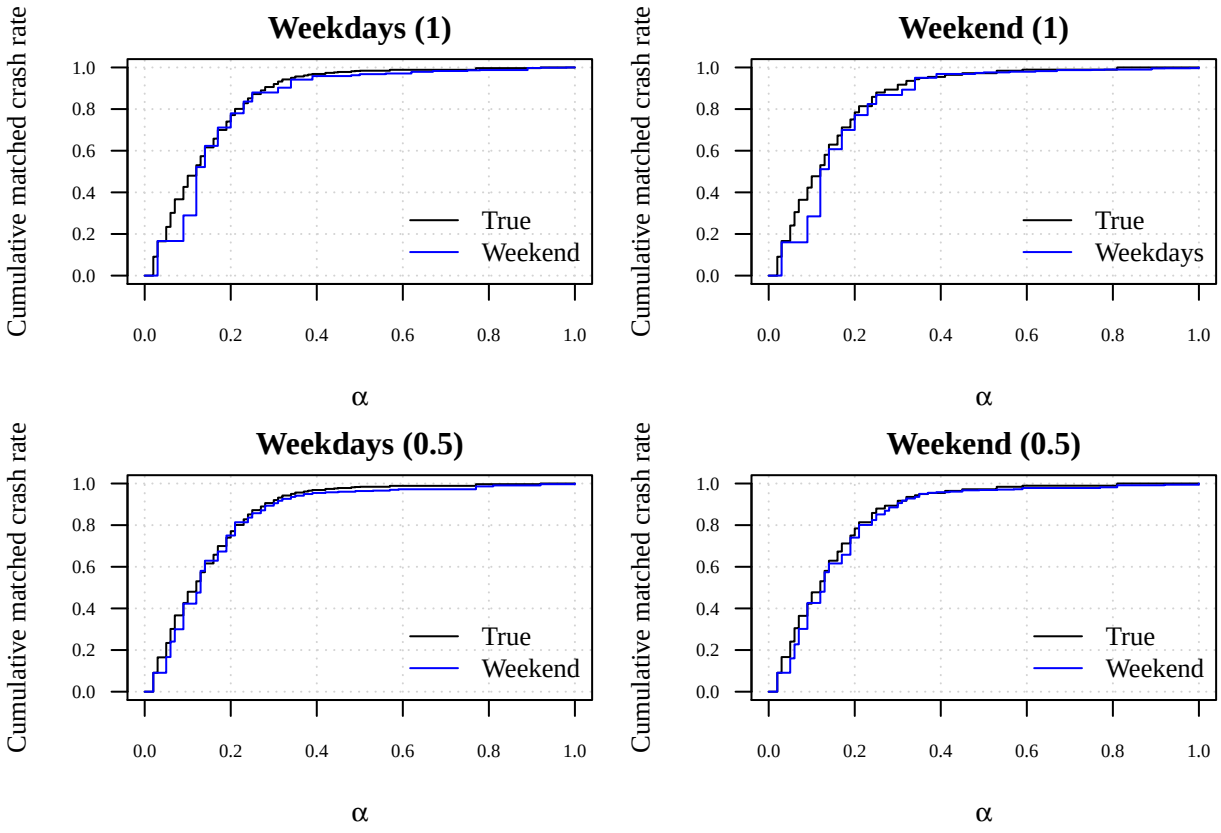

Figure 7: Comparison for weekdays and weekend with bin sizes 1 (top) and 0.5 (bottom) under the level sets using J^* Index

Table 16: KS test: p.value using level sets for weekdays and weekend with size 1 and significance level = 0.3

	wk	wkd
1	1	0.3158
2	0.4856	1

Table 17: KS test: max.d using level sets for weekdays and weekend with size 0.5 and significance level = 0.3

	wk	wkd
1	0	0.0806
2	0.0742	0

Table 18: KS test: p.value using level sets for weekdays and weekend with size 0.5 and significance level = 0.3

	wk	wkd
1	1	1
2	1	1

Table 19: KS test: max.d using level sets for days of week with size 1 and significance level = 0.3

	sun	mon	tue	wed	thu	fri	sat
1	0	0.1513	0.2517	0.2998	0.285	0.1541	0.241
2	0.1522	0	0.1301	0.1788	0.1676	0.1438	0.1556
3	0.2619	0.1181	0	0.1599	0.1428	0.1341	0.1463
4	0.2548	0.1026	0.0632	0	0.0649	0.122	0.1204
5	0.2601	0.1115	0.0645	0.0215	0	0.1274	0.1151
6	0.1327	0.097	0.1273	0.1681	0.1648	0	0.1528
7	0.22	0.0865	0.0826	0.1397	0.1249	0.105	0

Table 20: KS test: p.value using level sets for days of week with size 1 and significance level = 0.3

	sun	mon	tue	wed	thu	fri	sat
1	1	0.1983	0.0033	2e-04	5e-04	0.1818	0.0057
2	0.193	1	0.3616	0.0792	0.1172	0.248	0.1733
3	0.002	0.4887	1	0.151	0.2549	0.3252	0.2303
4	0.0028	0.6906	1	1	1	0.4444	0.4625
5	0.0022	0.5692	1	1	1	0.3882	0.5244
6	0.3375	0.7726	0.3891	0.1151	0.1289	1	0.1892
7	0.0151	0.9394	1	0.279	0.4141	0.6572	1

Table 21: KS test: max.d using level sets for days of week with size 0.5 and significance level = 0.3

	sun	mon	tue	wed	thu	fri	sat
1	0	0.1701	0.1924	0.2497	0.2688	0.1577	0.2727
2	0.1719	0	0.1248	0.1885	0.1728	0.1524	0.2145
3	0.1932	0.121	0	0.1222	0.1617	0.1283	0.1694
4	0.2043	0.1159	0.109	0	0.1519	0.1326	0.1654
5	0.2196	0.113	0.1493	0.1039	0	0.1459	0.1017
6	0.1539	0.137	0.1369	0.1745	0.1635	0	0.2298
7	0.2587	0.1576	0.098	0.1725	0.1606	0.1938	0

Table 22: KS test: p.value using level sets for days of week with size 0.5 and significance level = 0.3

	sun	mon	tue	wed	thu	fri	sat
1	1	0.1077	0.0476	0.0037	0.0014	0.1624	0.0011
2	0.1012	1	0.4144	0.0553	0.0981	0.1915	0.0192
3	0.0462	0.4553	1	0.443	0.1424	0.3791	0.1101
4	0.0295	0.5154	0.6024	1	0.1948	0.3387	0.1263
5	0.0153	0.5501	0.2103	0.6715	1	0.2331	0.703
6	0.1829	0.3005	0.3012	0.0924	0.1343	1	0.0096
7	0.0023	0.1627	0.7583	0.0991	0.1477	0.045	1

Table 23: KS test: max.d using level sets for months of year with size 1 and significance level = 0.3

	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec
1	0	0.1488	0.2091	0.2437	0.1583	0.2617	0.244	0.1809	0.0863	0.123	0.1397	0.1555
2	0.1446	0	0.1907	0.1471	0.1704	0.1552	0.205	0.2126	0.1623	0.1663	0.1663	0.154
3	0.2368	0.243	0	0.2416	0.2841	0.2596	0.2546	0.2945	0.2276	0.2723	0.2934	0.289
4	0.2399	0.158	0.2294	0	0.1954	0.1403	0.3068	0.2058	0.2444	0.2733	0.307	0.2073
5	0.0858	0.1488	0.178	0.218	0	0.236	0.2064	0.1498	0.0903	0.1192	0.1529	0.1187
6	0.2421	0.1488	0.2316	0.1447	0.1704	0	0.309	0.2126	0.2465	0.2755	0.3092	0.154
7	0.0784	0.1488	0.1887	0.199	0.2053	0.2265	0	0.2051	0.1333	0.17	0.1867	0.1149
8	0.068	0.1488	0.145	0.0839	0.0543	0.0769	0.1894	0	0.073	0.0862	0.1199	0.0979
9	0.1107	0.1488	0.1951	0.226	0.1807	0.244	0.2396	0.1699	0	0.1067	0.1344	0.1647
10	0.071	0.1488	0.1814	0.222	0.1182	0.24	0.2107	0.1532	0.0585	0	0.1034	0.1227
11	0.0671	0.1488	0.1644	0.222	0.0463	0.24	0.177	0.1362	0.0658	0.0472	0	0.1227
12	0.1137	0.246	0.2201	0.2343	0.1448	0.2523	0.1624	0.2946	0.1276	0.1414	0.1627	0

Table 24: KS test: p.value using level sets for months of year with size 1 and significance level = 0.3

	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec
1	1	0.2135	0.0242	0.005	0.159	0.002	0.0049	0.0734	0.9432	0.4341	0.2783	0.1737
2	0.2419	1	0.0507	0.2252	0.1066	0.1757	0.0287	0.0208	0.1399	0.1224	0.1225	0.1822
3	0.0069	0.0051	1	0.0055	6e-04	0.0022	0.0029	3e-04	0.0107	0.0011	3e-04	4e-04
4	0.006	0.161	0.0098	1	0.0423	0.274	1e-04	0.0277	0.0048	0.0011	1e-04	0.0261
5	0.95	0.2135	0.0815	0.0165	1	0.0072	0.027	0.2072	0.8784	0.4765	0.1885	0.4824
6	0.0054	0.2135	0.0089	0.2415	0.1066	1	1e-04	0.0208	0.0043	9e-04	1e-04	0.1822
7	1	0.2135	0.0549	0.0366	0.0283	0.0112	1	0.0286	0.3326	0.108	0.0591	0.5268
8	1	0.2135	0.2392	0.9832	1	1	0.0534	1	1	0.9448	0.4681	0.7592
9	0.5806	0.2135	0.0427	0.0115	0.0739	0.0049	0.0061	0.1082	1	0.6328	0.3228	0.129
10	1	0.2135	0.0721	0.0138	0.4877	0.0059	0.0226	0.1869	1	1	0.6795	0.4369
11	1	0.2135	0.1304	0.0138	1	0.0059	0.0846	0.3068	1	1	1	0.4368
12	0.5422	0.0044	0.015	0.0078	0.2409	0.0032	0.1393	3e-04	0.3861	0.2655	0.1382	1

Table 25: KS test: max.d using level sets for months of year with size 0.5 and significance level = 0.3

	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec
1	0	0.2254	0.2225	0.2252	0.2525	0.1963	0.3093	0.2271	0.1515	0.1772	0.2394	0.2648
2	0.2246	0	0.1521	0.2183	0.2568	0.2375	0.2791	0.278	0.2315	0.1772	0.2404	0.2743
3	0.2246	0.2074	0	0.1987	0.315	0.2196	0.2897	0.2233	0.1459	0.2273	0.3023	0.222
4	0.2547	0.2468	0.1512	0	0.3515	0.2479	0.353	0.3359	0.2079	0.3079	0.3465	0.3346
5	0.1569	0.1107	0.1645	0.1781	0	0.0897	0.2712	0.1064	0.1259	0.1117	0.209	0.2158
6	0.185	0.13	0.1759	0.2051	0.1576	0	0.2527	0.1667	0.1229	0.1389	0.3051	0.1817
7	0.1645	0.1423	0.1999	0.1676	0.2481	0.2053	0	0.2314	0.124	0.1326	0.1716	0.167
8	0.0969	0.1271	0.1127	0.1232	0.1603	0.0897	0.1782	0	0.0766	0.1377	0.1133	0.1857
9	0.1784	0.1893	0.1563	0.1508	0.2185	0.135	0.2001	0.1962	0	0.1506	0.2063	0.1857
10	0.2246	0.1391	0.152	0.1814	0.1969	0.156	0.1812	0.1815	0.1208	0	0.1916	0.1868
11	0.1134	0.2214	0.1441	0.1763	0.24	0.1566	0.1759	0.1638	0.1222	0.1579	0	0.2609
12	0.1703	0.086	0.1339	0.1957	0.1875	0.129	0.1629	0.2126	0.0931	0.1028	0.2282	0

Table 26: KS test: p.value using level sets for months of year with size 0.5 and significance level = 0.3

	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec
1	1	0.0118	0.0135	0.0119	0.0032	0.0409	1e-04	0.0109	0.1969	0.0838	0.0061	0.0017
2	0.0122	1	0.1934	0.0163	0.0026	0.0067	8e-04	8e-04	0.0089	0.0838	0.0058	0.001
3	0.0122	0.0259	1	0.0371	1e-04	0.0153	4e-04	0.013	0.2328	0.0108	2e-04	0.0138
4	0.0029	0.0043	0.1984	1	0	0.004	0	0	0.0254	1e-04	0	0
5	0.1663	0.5807	0.1301	0.0813	1	0.8872	0.0012	0.6369	0.4037	0.5671	0.0243	0.0181
6	0.0631	0.3631	0.0879	0.0286	0.1628	1	0.0032	0.1207	0.4348	0.2849	2e-04	0.0712
7	0.1299	0.2589	0.0354	0.1173	0.004	0.0283	1	0.009	0.4232	0.3385	0.1024	0.1197
8	0.7754	0.3913	0.5547	0.4314	0.1491	0.8872	0.081	1	1	0.2948	0.5467	0.0615
9	0.0804	0.0537	0.1698	0.2014	0.0161	0.3177	0.035	0.041	1	0.2023	0.0272	0.0615
10	0.0122	0.2833	0.1937	0.072	0.0398	0.1711	0.0727	0.0717	0.4577	1	0.0491	0.0589
11	0.5458	0.0141	0.2459	0.0865	0.0059	0.1682	0.0877	0.1333	0.4428	0.1614	1	0.0021
12	0.1069	0.9468	0.3272	0.0419	0.0575	0.3722	0.1371	0.0208	0.8337	0.6872	0.0104	1

Table 27: KS test: max.d using level sets for 2-hour periods with size 1 and significance level = 0.3

	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12
1	0	0.2011	0.2455	0.2684	0.2505	0.262	0.2744	0.2541	0.2475	0.2742	0.1948	0.2221
2	0.2644	0	0.3215	0.3479	0.3862	0.3842	0.4119	0.392	0.3726	0.4175	0.3827	0.3313
3	0.2546	0.2586	0	0.3955	0.4224	0.3024	0.444	0.4126	0.4128	0.4371	0.4174	0.3333
4	0.2989	0.3063	0.3268	0	0.2615	0.1875	0.3187	0.283	0.1474	0.3134	0.1901	0.2006
5	0.1791	0.1466	0.3764	0.1729	0	0.1141	0.1508	0.2433	0.0958	0.1499	0.2077	0.3521
6	0.1288	0.2665	0.2439	0.1534	0.12	0	0.1416	0.2238	0.1052	0.1408	0.1224	0.2196
7	0.1517	0.2733	0.374	0.2684	0.1427	0.1321	0	0.0925	0.1127	0.0512	0.1396	0.2438
8	0.2488	0.2117	0.3908	0.2684	0.2689	0.2618	0.1611	0	0.1295	0.1541	0.1564	0.2382
9	0.1904	0.1813	0.4023	0.1356	0.1394	0.1323	0.1767	0.1545	0	0.1714	0.1679	0.2652
10	0.1258	0.2662	0.3652	0.2684	0.1339	0.1233	0.0453	0.0978	0.1039	0	0.1308	0.2279
11	0.146	0.2714	0.3269	0.1485	0.2267	0.193	0.2117	0.1978	0.198	0.2235	0	0.2801
12	0.2023	0.2756	0.2336	0.1468	0.3242	0.2288	0.244	0.2818	0.282	0.2371	0.101	0

Table 28: KS test: p.value using level sets for 2-hour periods with size 1 and significance level = 0.3

	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12
1	1	0.0336	0.0046	0.0014	0.0035	0.002	0.001	0.0029	0.0041	0.001	0.0433	0.0137
2	0.0017	1	1e-04	0	0	0	0	0	0	0	0	0
3	0.0029	0.0023	1	0	0	2e-04	0	0	0	0	0	0
4	2e-04	2e-04	0	1	0.002	0.0574	1e-04	6e-04	0.2229	1e-04	0.052	0.0344
5	0.0784	0.2284	0	0.0977	1	0.5373	0.2012	0.0051	0.7913	0.2065	0.0256	0
6	0.3746	0.0015	0.0049	0.1855	0.4674	1	0.2638	0.0127	0.6536	0.2698	0.4402	0.0153
7	0.1958	0.0011	0	0.0014	0.256	0.3432	1	0.8422	0.555	1	0.2794	0.0049
8	0.0038	0.0217	0	0.0014	0.0013	0.002	0.1456	1	0.3676	0.1816	0.1689	0.0065
9	0.0513	0.0724	0	0.3123	0.2809	0.341	0.0853	0.1796	1	0.1028	0.1159	0.0016
10	0.4042	0.0016	0	0.0014	0.3273	0.4305	1	0.7606	0.6726	1	0.3552	0.0105
11	0.2323	0.0012	0	0.2155	0.0112	0.0464	0.0216	0.0385	0.0382	0.0129	1	7e-04
12	0.0321	9e-04	0.0081	0.227	0	0.0101	0.0049	7e-04	7e-04	0.0068	0.7139	1

Table 29: KS test: max.d using level sets for 2-hour periods with size 0.5 and significance level = 0.3

	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12
1	0	0.2816	0.3019	0.3605	0.4139	0.2969	0.3395	0.3397	0.358	0.2945	0.3582	0.3986
2	0.3736	0	0.3684	0.3532	0.3976	0.3982	0.3683	0.3763	0.3914	0.4229	0.4674	0.3469
3	0.2751	0.4424	0	0.3983	0.3969	0.3696	0.4647	0.4281	0.3696	0.391	0.3471	0.3247
4	0.2905	0.3063	0.3792	0	0.2635	0.2085	0.2096	0.182	0.2304	0.2158	0.2711	0.3081
5	0.2473	0.2759	0.3165	0.126	0	0.2216	0.2019	0.205	0.2272	0.243	0.1942	0.2454
6	0.1628	0.2759	0.3165	0.1667	0.2094	0	0.1071	0.0969	0.1469	0.1385	0.2625	0.2791
7	0.2426	0.2759	0.336	0.1667	0.2157	0.1008	0	0.1156	0.1528	0.131	0.2173	0.2876
8	0.2542	0.2968	0.3529	0.1667	0.2443	0.1284	0.1389	0	0.1237	0.1559	0.2422	0.2922
9	0.2386	0.3211	0.3563	0.1667	0.2459	0.1654	0.1882	0.1124	0	0.1602	0.293	0.3548
10	0.1472	0.2759	0.3165	0.1667	0.2178	0.1455	0.1652	0.1472	0.2174	0	0.2671	0.2775
11	0.2275	0.3045	0.3165	0.0847	0.1885	0.2073	0.1959	0.1853	0.2651	0.2561	0	0.1936
12	0.2145	0.309	0.3165	0.2498	0.2757	0.2611	0.2271	0.2574	0.249	0.2333	0.2624	0

Table 30: KS test: p.value using level sets for 2-hour periods with size 0.5 and significance level = 0.3

	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12
1	1	7e-04	2e-04	0	0	3e-04	0	0	0	3e-04	0	0
2	0	1	0	0	0	0	0	0	0	0	0	0
3	0.001	0	1	0	0	0	0	0	0	0	0	0
4	4e-04	2e-04	0	1	0.0018	0.0247	0.0237	0.0704	0.0094	0.0181	0.0012	1e-04
5	0.0041	9e-04	1e-04	0.402	1	0.014	0.0325	0.0287	0.0109	0.0051	0.0443	0.0046
6	0.1376	9e-04	1e-04	0.1209	0.0238	1	0.6285	0.7742	0.226	0.288	0.0019	8e-04
7	0.0053	9e-04	0	0.1209	0.0182	0.7166	1	0.5184	0.1889	0.3537	0.017	5e-04
8	0.0029	3e-04	0	0.1209	0.0048	0.378	0.2853	1	0.4268	0.1719	0.0054	4e-04
9	0.0064	1e-04	0	0.1209	0.0045	0.1261	0.0559	0.558	1	0.1497	3e-04	0
10	0.2244	9e-04	1e-04	0.1209	0.0166	0.2358	0.1271	0.2242	0.0169	1	0.0015	8e-04
11	0.0108	2e-04	1e-04	0.9683	0.0553	0.026	0.0415	0.0624	0.0016	0.0027	1	0.0455
12	0.0192	1e-04	1e-04	0.0037	9e-04	0.002	0.011	0.0025	0.0038	0.0082	0.0019	1

Table 31: KS test: max.d using level sets for 4-hour periods with size 1 and significance level = 0.3

	s1	s2	s3	s4	s5	s6
1	0	0.18	0.2535	0.2778	0.2705	0.3233
2	0.2007	0	0.2598	0.2278	0.2362	0.2138
3	0.0802	0.0639	0	0.1282	0.1154	0.2395
4	0.1791	0.0834	0.1453	0	0.1275	0.2565
5	0.1727	0.0779	0.1014	0.1468	0	0.2645
6	0.164	0.0669	0.223	0.2288	0.2165	0

Table 32: KS test: p.value using level sets for 4-hour periods with size 1 and significance level = 0.3

	s1	s2	s3	s4	s5	s6
1	1	0.0759	0.003	8e-04	0.0012	1e-04
2	0.0342	1	0.0022	0.0106	0.0072	0.0197
3	1	1	1	0.3807	0.5209	0.0061
4	0.0782	0.9909	0.2375	1	0.3873	0.0026
5	0.0982	1	0.7078	0.2267	1	0.0017
6	0.1323	1	0.0131	0.0101	0.0176	1

Table 33: KS test: max.d using level sets for 4-hour periods with size 0.5 and significance level = 0.3

	s1	s2	s3	s4	s5	s6
1	0	0.2758	0.295	0.2916	0.3307	0.3413
2	0.275	0	0.2142	0.2646	0.2665	0.1833
3	0.2167	0.1524	0	0.1302	0.2187	0.2273
4	0.2323	0.1985	0.1456	0	0.1542	0.2523
5	0.2219	0.157	0.2109	0.163	0	0.2616
6	0.2405	0.1261	0.2512	0.294	0.2552	0

Table 34: KS test: p.value using level sets for 4-hour periods with size 0.5 and significance level = 0.3

	s1	s2	s3	s4	s5	s6
1	1	9e-04	3e-04	4e-04	0	0
2	0.001	1	0.0194	0.0017	0.0015	0.0671
3	0.0174	0.1914	1	0.3612	0.016	0.0108
4	0.0086	0.0373	0.2348	1	0.1811	0.0032
5	0.0138	0.1659	0.0224	0.1366	1	0.002
6	0.0058	0.4013	0.0034	3e-04	0.0028	1

Table 35: KS test: max.d using level sets for 8-hour periods with size 1 and significance level = 0.3

	s1	s2	s3
1	0	0.1712	0.1748
2	0.0831	0	0.1213
3	0.0713	0.1323	0

Table 36: KS test: p.value using level sets for 8-hour periods with size 1 and significance level = 0.3

	s1	s2	s3
1	1	0.1034	0.0915
2	0.9959	1	0.4522
3	1	0.341	1

Table 37: KS test: max.d using level sets for 8-hour periods with size 0.5 and significance level = 0.3

	s1	s2	s3
1	0	0.2581	0.2141
2	0.1419	0	0.1526
3	0.1344	0.1536	0

Table 38: KS test: p.value using level sets for 8-hour periods with size 0.5 and significance level = 0.3

	s1	s2	s3
1	1	0.0024	0.0195
2	0.2619	1	0.1901
3	0.3229	0.1848	1

Table 39: KS test: max.d using level sets for 12-hour periods with size 1 and significance level = 0.3

	s1	s2
1	0	0.142
2	0.1228	0

Table 40: KS test: p.value using level sets for 12-hour periods with size 1 and significance level = 0.3

	s1	s2
1	1	0.2611
2	0.4358	1

12:00-24:00 are slightly better with bin size 1 whereas no period fits properly with bin size 0.5.

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 model, then the other set, e.g. weekend, will be tested based on this model.

3.3.1 Predictability on time variation

Figure 8 shows the performance using weekdays as model based on the previous analysis, and the other as testing data.

For days of week (Figure 9), Sunday is used as a model based on KS results using level sets, and others are compared to the reference model.

For months of year (Figure 10), May, June and July are used as a model individually based on the analyses; rest of the months are compared to the model.

Time of day can subset into 2-hour, 4-hour and 8-hour periods (Figure 11 to Figure 14). The base models are established based on level sets and KS results described in previous sections. In 2-hour periods, 04:00-06:00, 08:00-16:00 and 18:00-22:00 are used as model respectively. In 4-hour periods, 00:00-04:00 is used as a model. In 8-hour periods, each is used as a model and the rest is compared.

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 specific group stands out from the KS results, a model used based on that group is constructed to the rest of datasets (e.g. Sunday is model, Monday through Saturday are testing data).

For weekdays and weekend (Figure 14), weekdays are first as model, and weekend is tested (left); then weekend is used as model and then testing weekdays (right). Two bin sizes, 1 (top) and 0.5 (bottom) are implemented.

Table 41: KS test: max.d using level sets for 12-hour periods with size 0.5 and significance level = 0.3

	s1	s2
1	0	0.1625
2	0.1396	0

Table 42: KS test: p.value using level sets for 12-hour periods with size 0.5 and significance level = 0.3

	s1	s2
1	1	0.139
2	0.2792	1

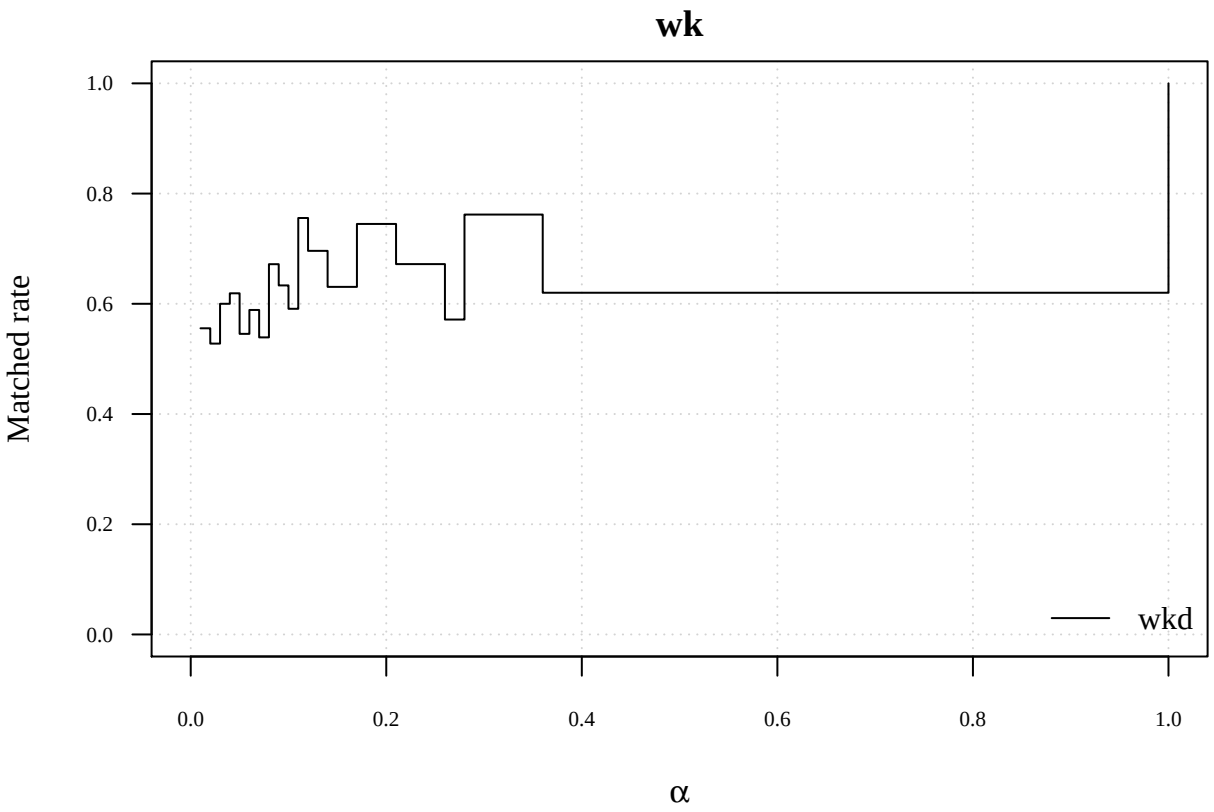

Figure 8: Matched rate for weekdays and weekend using level sets

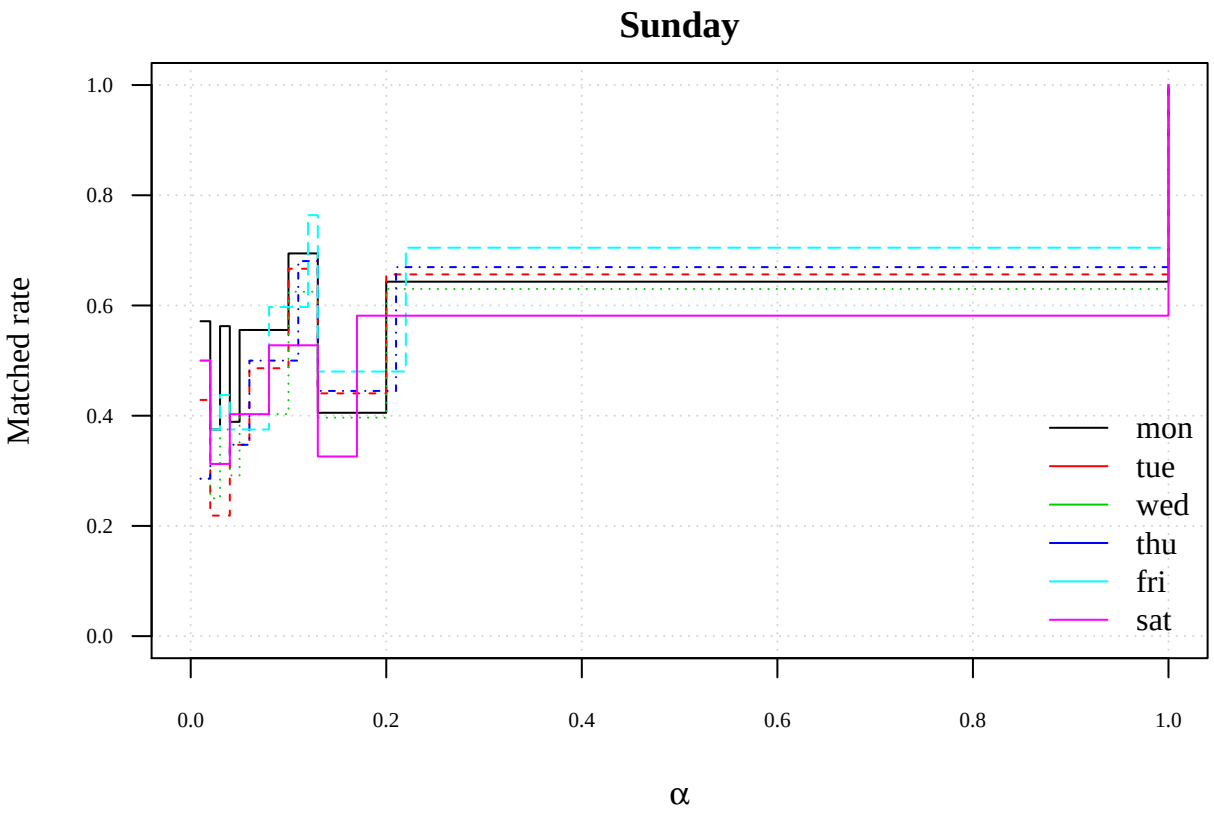

Figure 9: Matched rate for days of week using level sets

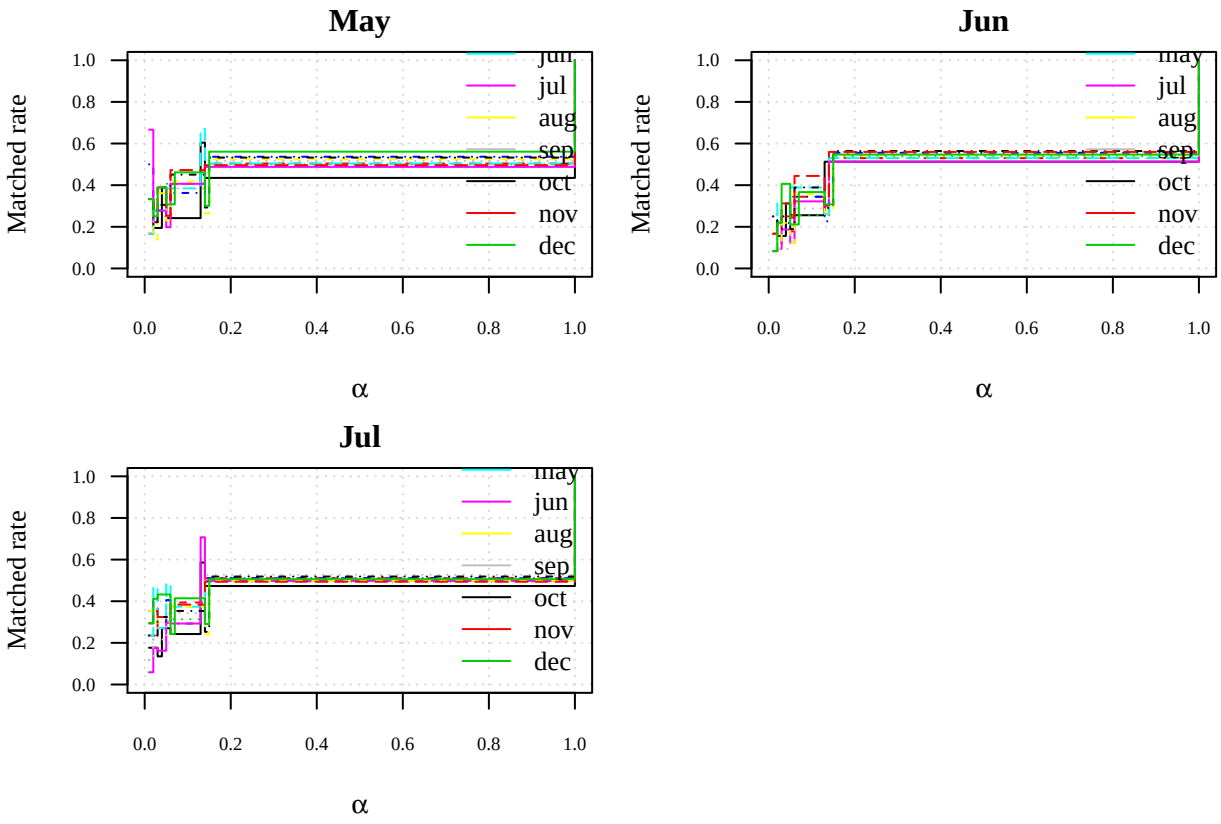

Figure 10: Matched rate for months of year using level sets

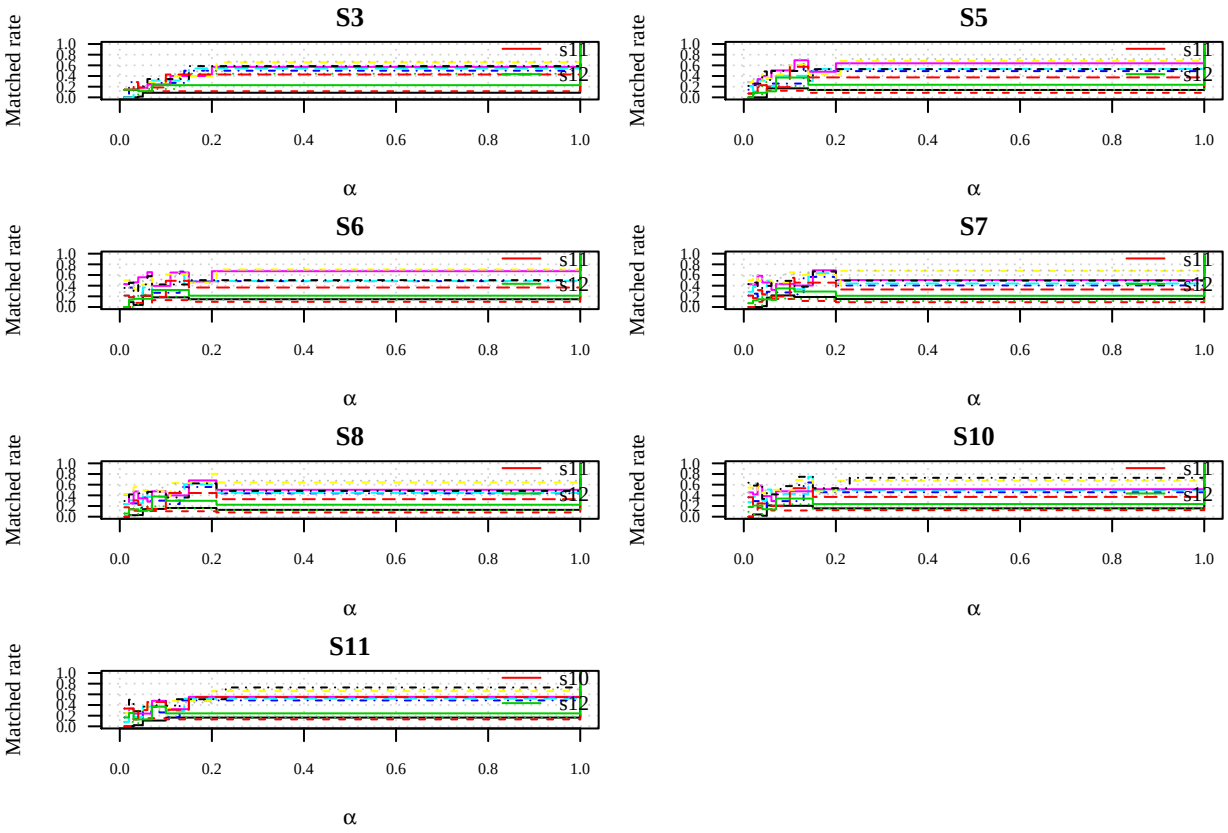

Figure 11: Matched rate for 2-hour periods using level sets

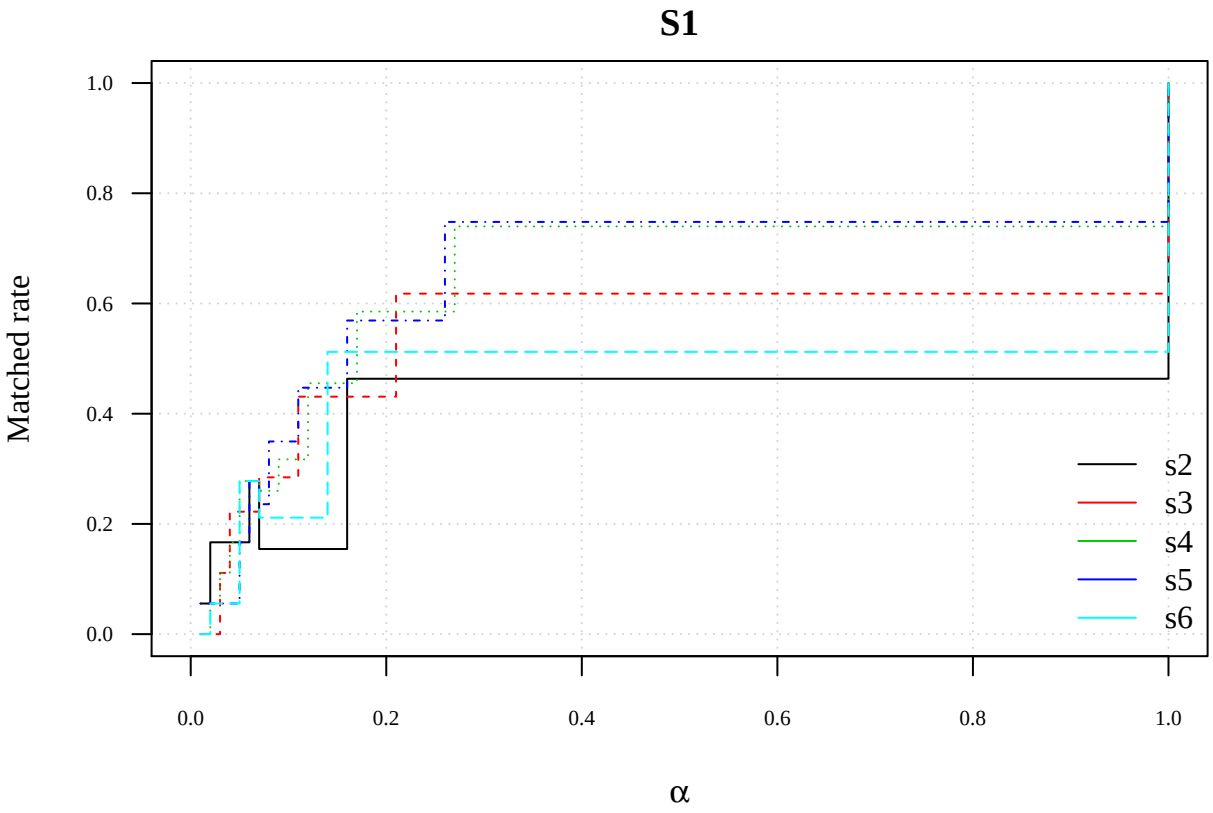

Figure 12: Matched rate for 4-hour periods using level sets

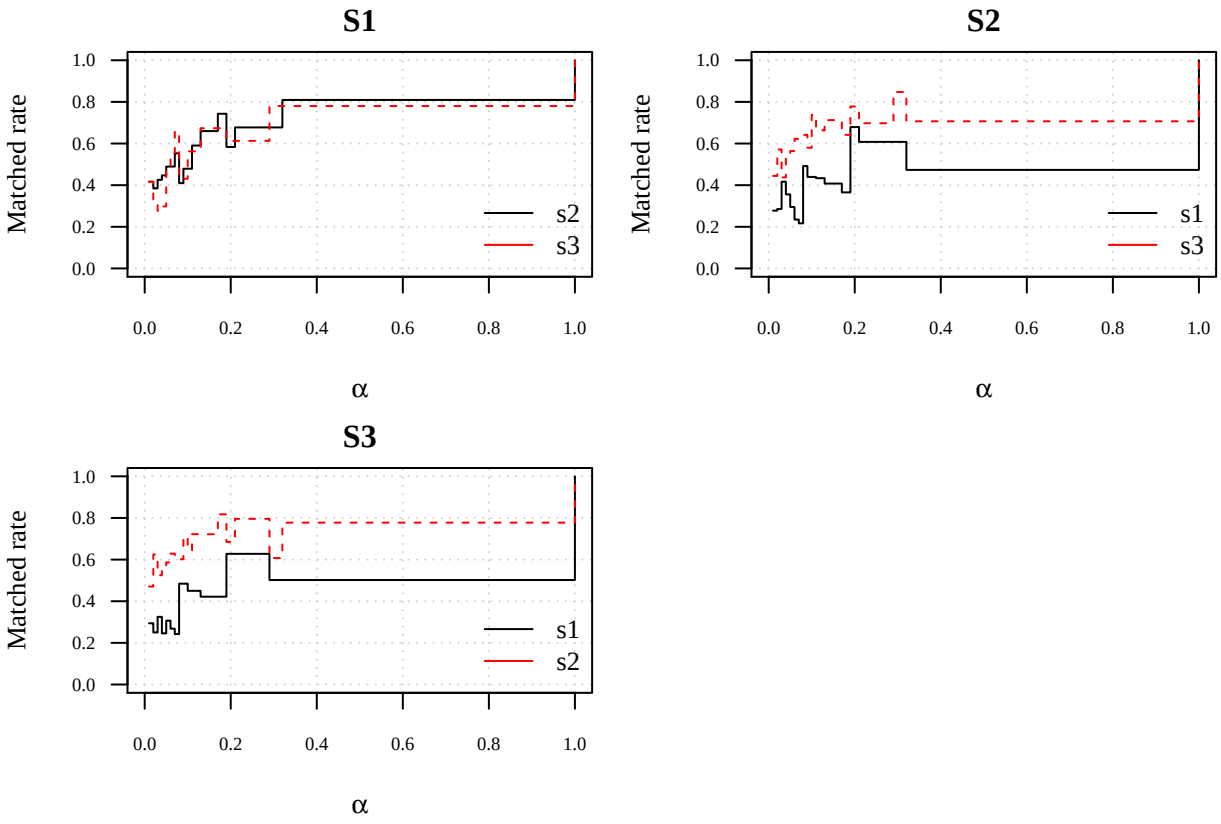

Figure 13: Matched rate for 8-hour periods using level sets

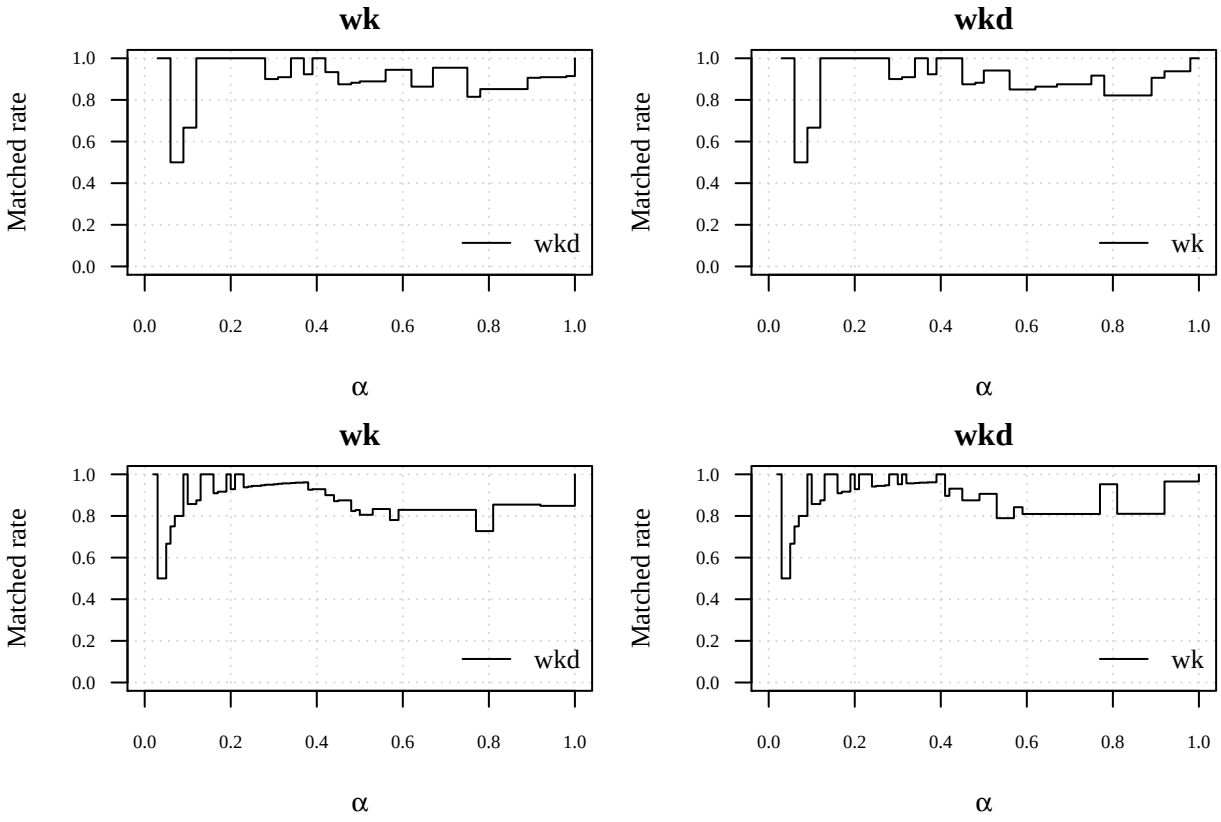

Figure 14: Matched rate for weekdays and weekend with size 1 and 0.5 using level sets

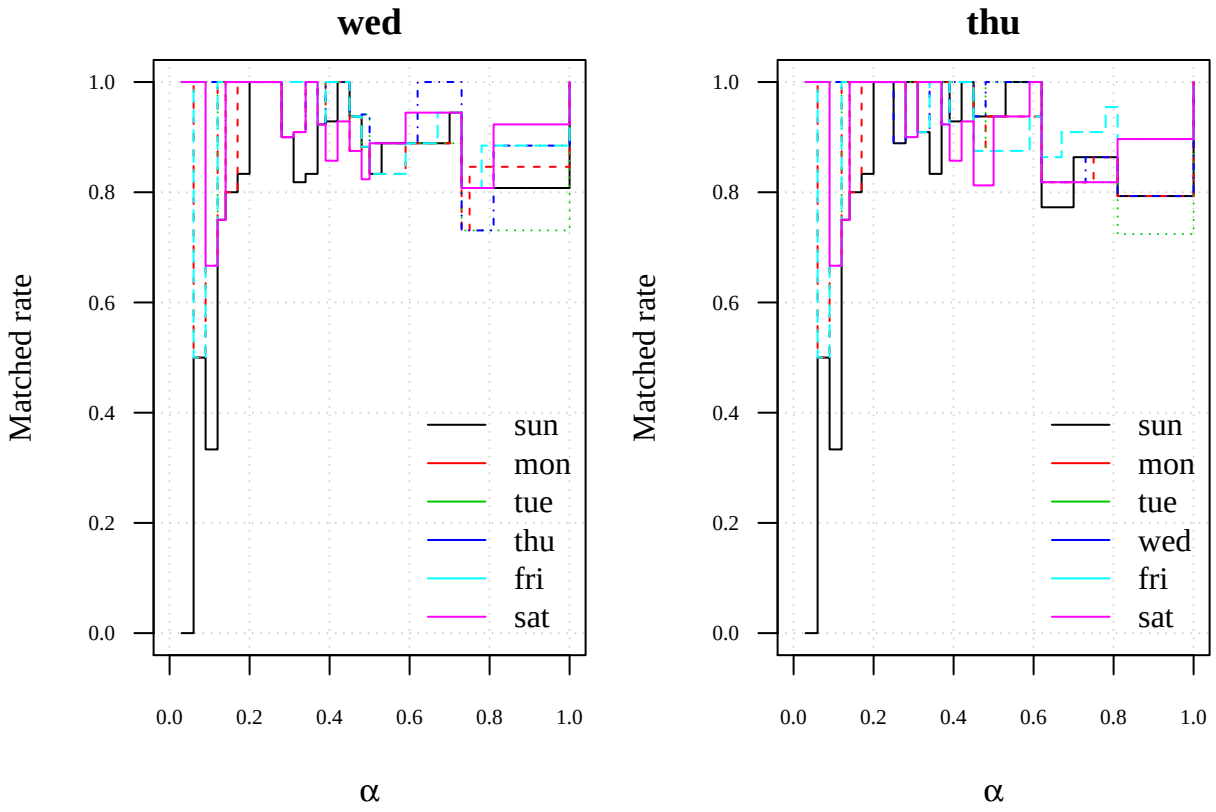

Figure 15: Matched rate for days of week with size 1 using level sets

Days of week with bin sizes 1 and 0.5 (Figure 15 and 16) are manipulated. For size 1, Wednesday and Thursday are used as model respectively in accordance with previous KS results. For size 0.5, Tuesday through Thursday are served as model individually.

Months of year (Figure 17) with bin size 1 (left) and 0.5 (right) are compared. For both sizes, only August is similar to more months than others and therefore, it is used as a model.

Time of day with only bin size 1 for 2-hour, 4-hour, 8-hour divisions are showed from Figure 18 to 21. For 2-hour division (Figure 18), 18:00-20:00 is built as model. For 4-hour division (Figure 19), 08:00-12:00 is used as reference. For 8-hour division (Figure 20), two periods, 08:00-16:00 and 16:00-24:00 are used as models individually. Lastly, For 12-hour division, 12:00-24:00 is refered as model.

References

[1] Ismail Ayed, Amar Mitiche, 2010. *Variational and Level Set Methods in Image Segmentation.* P61.

[2] Martin Burger, Andrea Mennucci, Stanley Osher, Martin Rumpf, 2008. *Level Set and PDE Based Reconstruction Methods in Imaging.* P15.

[3] Kumar S, et al. *Healthcare associated infections in neonatal intensive care unit and its correlation with environmental surveillance.* Journal of Infection and Public Health (2017).

[4] Melanie Carder et al. *The Health and Occupation Research Network: An Evolving Surveillance System.* Safety and Health at Work 8 (2017) 231-236.

[5] Paul Jaccard, 1912. *The distribution of the flora in the alpine zone.* The New Phytologist, 11 (2).

[6] Jean Gibbons, Subhabrata Chakraborti, 2003. *Nonparametric Statistical Inference.* 4e.

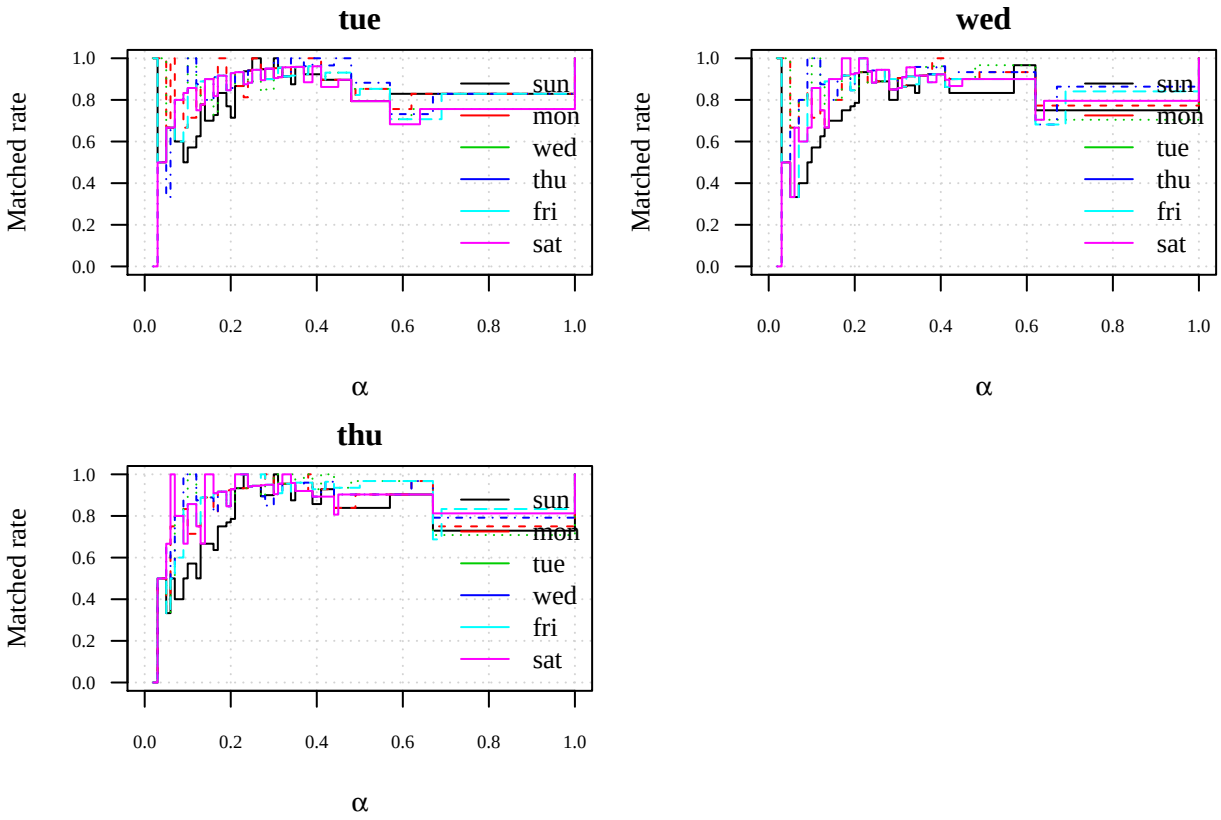

Figure 16: Matched rate for days of week with size 0.5 using level sets

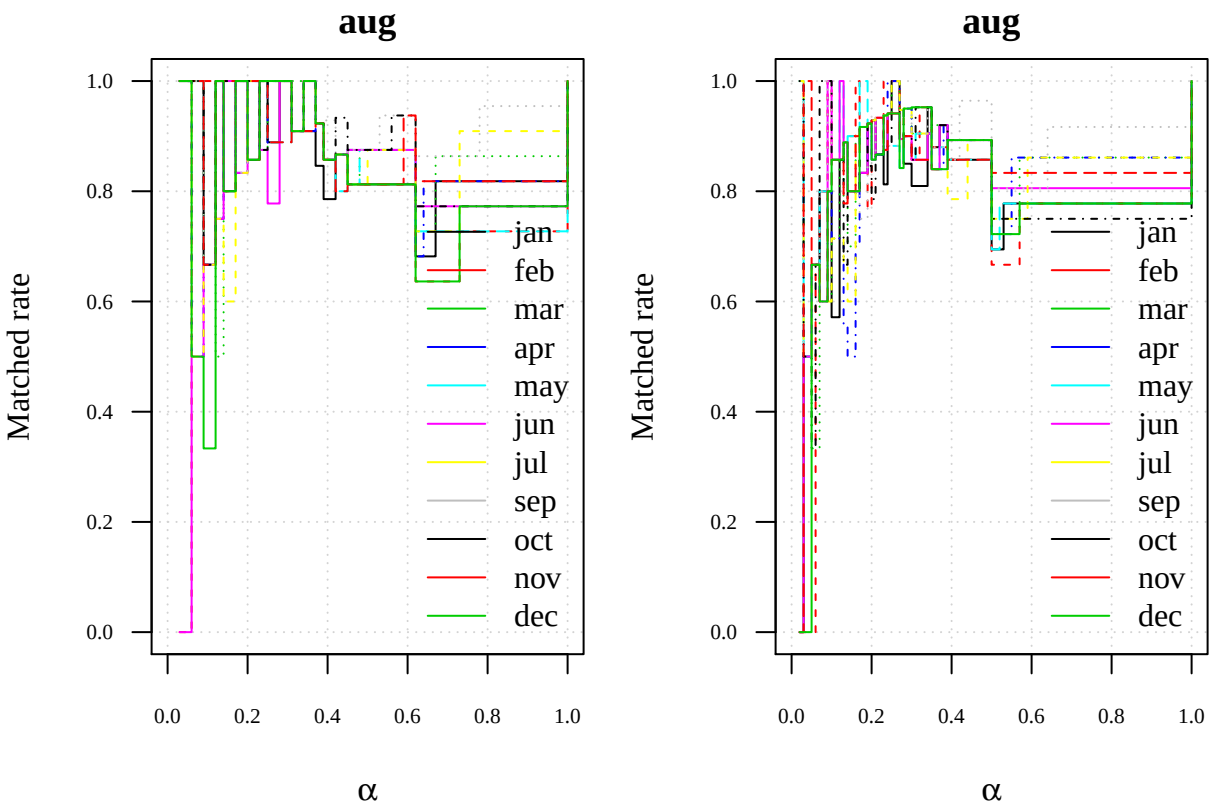

Figure 17: Matched rate for months of year with size 1 and 0.5 using level sets

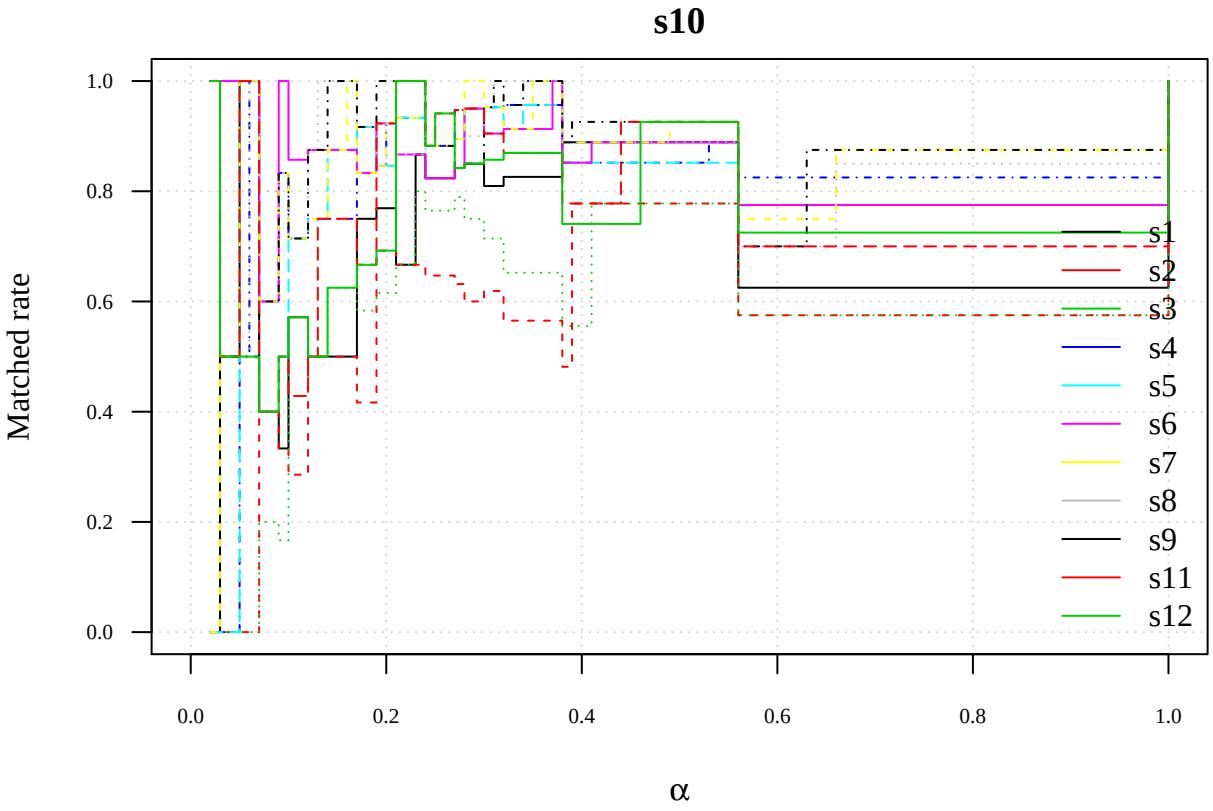

Figure 18: Matched rate for 2-hour periods with size 1 using level sets

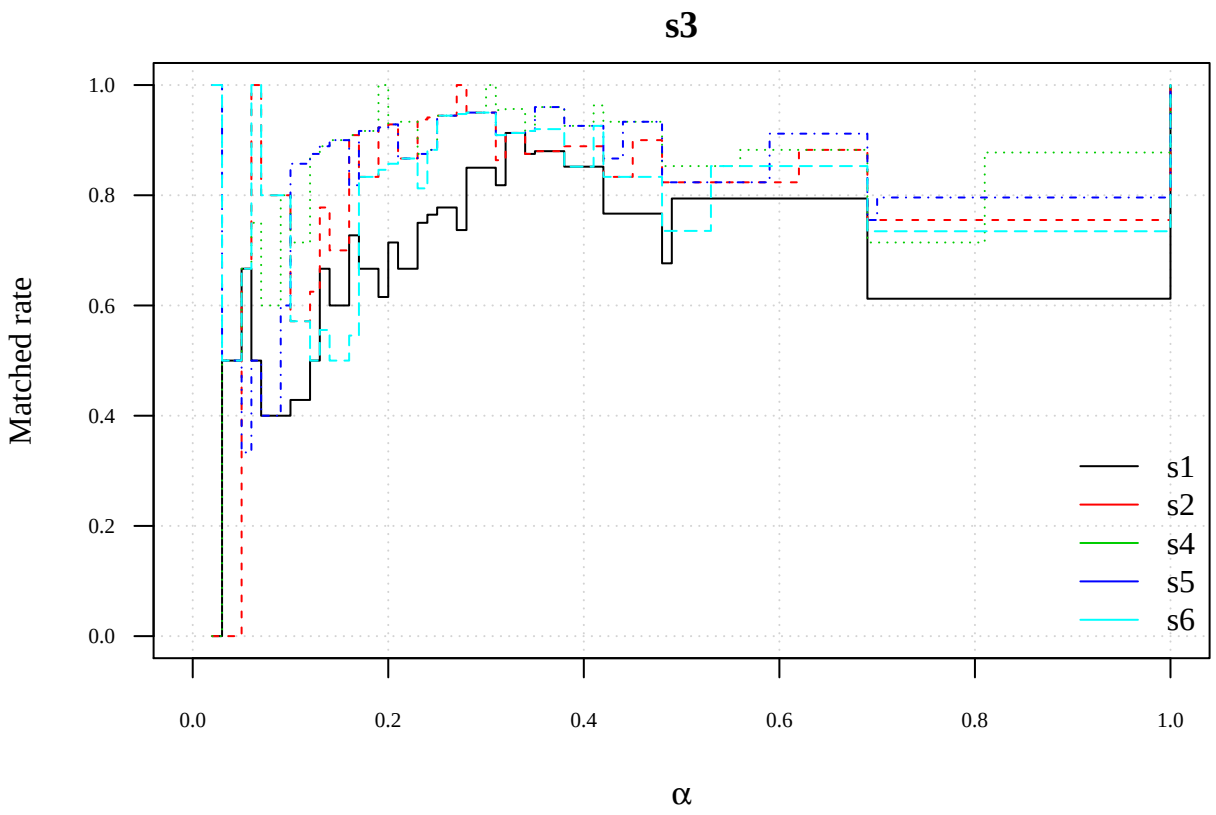

Figure 19: Matched rate for 4-hour periods with size 1 using level sets

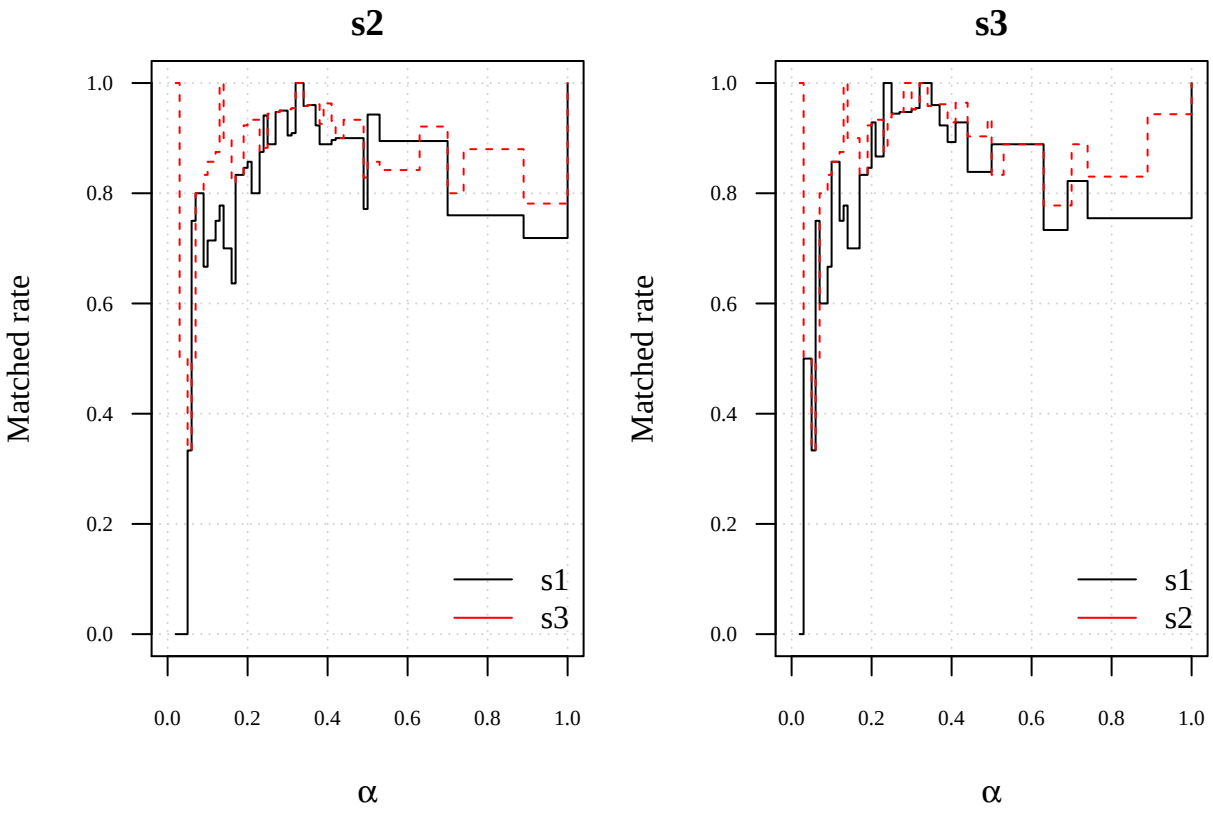

Figure 20: Matched rate for 8-hour periods with size 1 using level sets

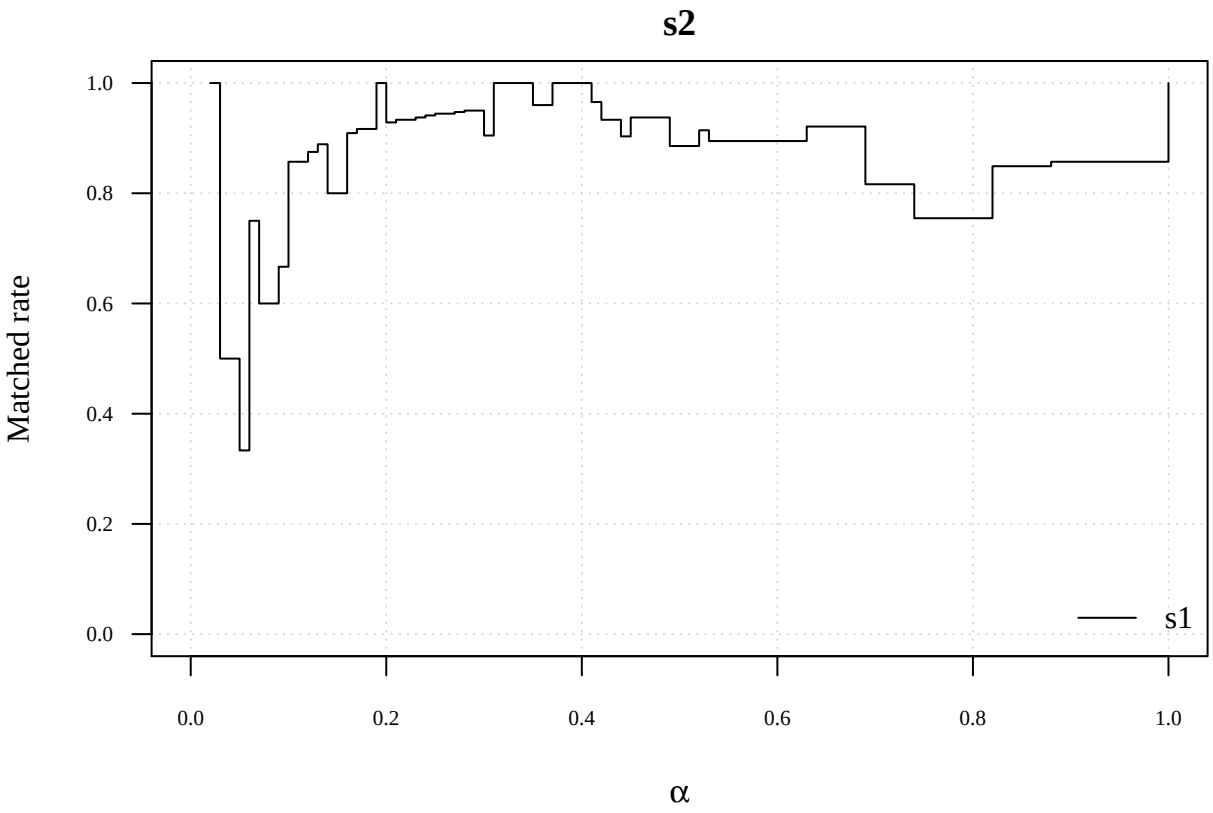

Figure 21: Matched rate for 12-hour periods with size 1 using level sets

[7] Grigorios Fountas, Panagiotis Anastasopoulos, 2017. *A random thresholds random parameters hierarchical ordered probit analysis of highway accident injury-severities.* Analytic Methods in Accident Research 15 1–16.