Emission Inspection Failure or Pass Rate Summary in PA

Independent Study Research

Lin Lyu

Carnegie Mellon University
Civil & Environmental Engineering

1 Time since last pass calculation for year 2015 and 2016 for PA

Bins' length 10 weeks

Penn(2015, 2016) A	
Bin weeks (time since last pass)	original records in each bin
0-10	109131
10 20	59192
20-30	55240
30-40	65785
40-50	666399
50-60	1756191
>60	310391
sum	3022329

- ***Total records in each bin***
- ***Calculation Method:***
  Got from dict, VIN as key, then calculate time since last inspection as value,
  count numbers in each bin
- ***90.4% records lie in more than 40 weeks***

1 Time since last pass calculation for year 2015 and 2016 for PA

Bins' length 10 weeks

Penn(2015, 2016)	A	B
Bin weeks (time since last pass)	original records in each bin	numbers in each bins all stations time since last pass(P-P)
0-10	109131	40520
10 20	59192	51216
20-30	55240	50322
30-40	65785	60910
40-50	666399	630261
50-60	1756191	1644653
>60	310391	203297
sum	3022329	2681179

- ***Time since last pass, P-P, in each bin***
- ***Calculation Method:***
  Got from dict, VIN as key, if first record is a 'Pass' and second one is also a 'Pass' , calculate the time since last inspection as a value, then count numbers

1 Time since last pass calculation for year 2015 and 2016 for PA

Bins' length 10 weeks

Penn(2015, 2016)	A	B	C
Bin weeks (time since last pass)	original records in each bin	numbers in each bins all stations time since last pass(P-P)	number in each bin/total number of last pass
0-10	109131	40520	0.0151
10 20	59192	51216	0.0191
20-30	55240	50322	0.0188
30-40	65785	60910	0.0227
40-50	666399	630261	0.2351
50-60	1756191	1644653	0.6134
>60	310391	203297	0.0758
sum	3022329	2681179	

- ***Calculation Method: # each bin / Total P-P, in each bin***
- **92.43% lie in more than 40 weeks**

1 Time since last pass calculation for year 2015 and 2016 for PA

Bins' length 10 weeks

Penn(2015, 2016)	A	B	C	D
Bin weeks (time since last pass)	original records in each bin	numbers in each bins all stations time since last pass(P-P)	number in each bin/total number of last pass	totally pass rate(P-P/all) all data (Col B / Col A)
0-10	109131	40520	0.0151	0.3713
10 20	59192	51216	0.0191	0.8653
20-30	55240	50322	0.0188	0.9110
30-40	65785	60910	0.0227	0.9259
40-50	666399	630261	0.2351	0.9458
50-60	1756191	1644653	0.6134	0.9365
>60	310391	203297	0.0758	0.6550
sum	3022329	2681179		

- ***Time since last pass***
- ***Calculation Method: totally pass rate(P-P/all) (Col B / Col A)***

1 Time since last pass calculation for year 2015 and 2016 for PA

Bins' length 10 weeks

Penn(2015, 2016)	A	B	C	D	E
Bin weeks (time since last pass)	original records in each bin	numbers in each bins all stations time since last pass(P-P)	number in each bin/total number of last pass	totally pass rate(P-P/all) all data (B/D)	pass rate(P-P/all & F+N-P/all) all data
0-10	109131	40520	0.0151	0.3713	0.8673
10 20	59192	51216	0.0191	0.8653	0.9408
20-30	55240	50322	0.0188	0.9110	0.9503
30-40	65785	60910	0.0227	0.9259	0.9541
40-50	666399	630261	0.2351	0.9458	0.9633
50-60	1756191	1644653	0.6134	0.9365	0.9566
>60	310391	203297	0.0758	0.6550	0.9294
sum	3022329	2681179			

- ***Time since last pass, pass rate(P-P/all & N-P/all) all data***
- ***Calculation Method: make dict, VIN as key, calculate time since last inspection if first one is P or N or F and second is P***
- ***pass means the second observation is pass, so it could be P-P, N-P, F-P***

1 Time since last pass calculation for year 2015 and 2016 for PA

Bins' length 10 weeks

Penn(2015, 2016)	A	B	C	D	E	F
Bin weeks (time since last pass)	original records in each bin	numbers in each bins all stations time since last pass(P-P)	number in each bin/total number of last pass	totally pass rate(P-P/all) all data (B/D)	pass rate(P-P/all & N-P/all) all data	records at the same station(after clean all data without same station)
0-10	109131	40520	0.0151	0.3713	0.8673	44705
10 20	59192	51216	0.0191	0.8653	0.9408	14050
20-30	55240	50322	0.0188	0.9110	0.9503	14270
30-40	65785	60910	0.0227	0.9259	0.9541	21500
40-50	666399	630261	0.2351	0.9458	0.9633	422817
50-60	1756191	1644653	0.6134	0.9365	0.9566	1058287
>60	310391	203297	0.0758	0.6550	0.9294	151296
sum	3022329	2681179				1726925

- ***Records at the same station***
- ***After clean all data without same station, then made dcit using VIN as key and calculate time since last inspection as value, count numbers***
- ***57% happening at the same station***

2 Time since last pass calculation for year 2015 and 2016 for PA

Bins' length 5 weeks VS 10 weeks

Penn	5 weeks				10 weeks			
Bin weeks (time since last pass)	numbers in each bins, all stations, time since last pass(P-P)	original records in each bin	# P-P / total in bin(5 weeks)	P-Pnum of diff stations/total in each bin	LARGE BIN (10 weeks)	all stations, time since last pass(P-P)	original records in each bin	# P-P / total in bin(10 weeks)
0-5	19240	78861	0.2440	0.7275	0-10	40520	109137	0.3713
5-10	21280	30276	0.7029	0.7410				
10-15	24835	29388	0.8451	0.7224	10-20	51216	59197	0.8652
15-20	26381	29809	0.8850	0.6963				
20-25	26130	28818	0.9067	0.6776	20-30	50322	55241	0.9110
25-30	24192	26423	0.9156	0.6929				
30-35	24734	26892	0.9198	0.6984	30-40	60910	65790	0.9258
35-40	36176	38898	0.9300	0.5843				
40-45	147769	156263	0.9456	0.3732	40-50	630261	666436	0.9457
45-50	482492	510173	0.9457	0.3271				
50-55	1221046	1297685	0.9409	0.3586	50-60	1644653	1756274	0.9364
55-60	423607	458589	0.9237	0.4351				
60-65	124278	136159	0.9127	0.4838	>60	203297	310394	0.6550
>65	79019	174235	0.4535	0.5769				

- ***Calculation Methos***
- ***Made dict use VIN as key, if the first one and the second one happening at different station, calculate time since last inspection as value, then count numbers***

Use Records at the diff station / total # in that Bin

2 Time since last pass calculation for year 2015 and 2016 for PA

Bins' length 5 weeks Code

1 Add conditions, such as P-P, P-F or N

```
'100000000010375WI': [(Timestamp('2015-02-26 00:00:00'), 'P', 'BN73'),
 (Timestamp('2016-02-11 00:00:00'), 'P', 'BN73')],
'101NDS2J42M677618': [(Timestamp('2015-01-12 00:00:00'), 'P', '9548')],
'101PC5SBXE7354343': [(Timestamp('2016-02-04 00:00:00'), 'P', 'EK45')],
'103HV13T095816523': [(Timestamp('2016-01-22 00:00:00'), 'P', '8678')],
'104GP21E968516129': [(Timestamp('2016-03-04 00:00:00'), 'P', 'K80')],
'104GP24R758421565': [(Timestamp('2016-06-13 00:00:00'), 'P', 'T563')],
'104GP25303B184442': [(Timestamp('2016-05-09 00:00:00'), 'P', 'M816')],
'104GP253438103314': [(Timestamp('2016-09-06 00:00:00'), 'P', 'E014')],
'104GP45393B247176': [(Timestamp('2015-03-13 00:00:00'), 'P', 'L66')],
```

```
def get_time_dict(x):
    time_dict = {}
    for key in tqdm(merge_dict):
        l = len(merge_dict[key])
        for i in range(l):
            if i+1 < l and merge_dict[key][i][1]==x and merge_dict[key][i+1][1]==y:
                date1 = merge_dict[key][i][0]
                date2 = merge_dict[key][i+1][0]
                interval = int((date2 - date1).days/7)
                if key not in time_dict:
                    time_dict[key] = [[interval,merge_dict[key][i][2],merge_dict[key][i+1][2]]]
                else:
                    time_dict[key].append([interval,merge_dict[key][i][2],merge_dict[key][i+1][2]])
    return time_dict
```

```
time_dict
```

```
'19UUA5641XA040857': [[52, 'DH38', 'DH38']],
'19UUA5641XA041975': [[53, 'U052', 'U052']],
'19UUA5641XA045525': [[51, 'B975', 'B975']],
'19UUA5641XA047484': [[61, 'T443', 'EN51']],
'19UUA5641XA049820': [[54, 'EX73', 'EY96']],
'19UUA5641XA051227': [[51, 'A278', 'A278']],
'19UUA5641XA051700': [[51, 'EG07', 'EG07']],
'19UUA5641XA052801': [[50, 'B839', 'B839']],
'19UUA5642XA004661': [[53, 'EB11', 'EB11']],
```

2 Count number in each bin based on diff Conditions, such as P-P

```
def count_nr(time_dict):
    count1=0
    count2=0
    count3=0
    count4=0
    count5=0
    count6=0
    count7=0
    count8=0
    count9=0
    count10=0
    count11=0
    count12=0
    count13=0
    count14=0
    dict_5={}
    dict_10={}
    dict_15={}
    dict_20={}
    dict_25={}
    dict_30={}
    dict_35={}
    dict_40={}
    dict_45={}
    dict_50={}
    dict_55={}
    dict_60={}
    dict_65={}
    dict_70={}
```

3 Return a list of num of diff stations / total number in each bin, based on diff conditions, such asP-P

```
def get_prop():
    prop_lst=[]
    for i in range(13):
        j = (i+1)*5
        d = globals()['dict_{}'.format(j)]
        num_dif_stations = num_dif(d,j-5,j)
        prop = num_dif_stations/globals()['count{}'.format(i+1)]
        prop_lst.append(prop)
    return prop_lst
```

```
def get_all_prop(dict_70, p_lst):
    num_dif_station_70 = 0
    for key in dict_70:
        for row in dict_70[key]:
            if 65<row[0]:
                if row[1]!=row[2]:
                    num_dif_station_70+=1
    #print(num_dif_station_70/count14)
    p_lst.append(num_dif_station_70/count14)

    return p_lst
```

2 Time since last pass calculation for year 2015 and 2016 for PA

Bins' length 5 weeks Code

P-P

In [49]:
```
# get time_dict with filter of 'F' &'P'
time_dict_f_p = get_time_dict('P','P')

# get numbers of count in each bin, for F & P
# and a dict for each bin for counting num of diff stations in each bin afterw
# update count and dict_{} for F & P
count1,count2,count3,count4,count5,count6,count7,count8,count9,count10,count11

# get percentage of num of diff stations in each bin as a list
prop_lst_f_p = get_prop()
all_lst = get_all_prop(dict_70, prop_lst_f_p)
```

```
100%|| 5597872/5597872 [01:55<00:00, 48435.01it/s]
```

In [50]:
```
print(all_lst)
print(count1,count2,count3,count4,count5,count6,count7,count8,count9,count10,c
```

```
[0.7275467775467775, 0.7410244360902256, 0.7224078920877793, 0.696334483150752
33, 0.6984717393062182, 0.5843653250773994, 0.37328533048203616, 0.32711215937
6321095, 0.48383462881604145, 0.576924537136638]
19240 21280 24835 26381 26130 24192 24734 36176 147769 482492 1221046 423607 1
```

P-F or N

n [67]:
```
# get time_dict with filter of 'F' &'P'
time_dict_f_p = get_time_dict('P')

# get numbers of count in each bin, for F & P
# and a dict for each bin for counting num of diff stations in each bin afterwards
# update count and dict_{} for F & P
count1,count2,count3,count4,count5,count6,count7,count8,count9,count10,count11,count12,count13,c

# get percentage of num of diff stations in each bin as a list
prop_lst_f_p = get_prop()
all_lst = get_all_prop(dict_70, prop_lst_f_p)
```

```
100%|| 5597872/5597872 [00:24<00:00, 231641.52it/s]
```

n [68]:
```
print(all_lst)
print(count1,count2,count3,count4,count5,count6,count7,count8,count9,count10,count11,count12,cou
```

```
[0.726530612244898, 0.7797029702970297, 0.7311015118790497, 0.7511870845204178, 0.736030828516376,
6, 0.7396226415094339, 0.6827253957329663, 0.4731993299832496, 0.4138948161164805, 0.452218507136
6756, 0.5969387755102041, 0.6801787916152897]
735 808 926 1053 1038 994 1060 1453 4776 15934 44264 19532 6860 6488
```

F or N-P

n [64]:
```
# get time_dict with filter of 'F' &'P'
time_dict_f_p = get_time_dict('P')

# get numbers of count in each bin, for F & P
# and a dict for each bin for counting num of diff stations in each bin afterwards
# update count and dict_{} for F & P
count1,count2,count3,count4,count5,count6,count7,count8,count9,count10,count11,count1

# get percentage of num of diff stations in each bin as a list
prop_lst_f_p = get_prop()
all_lst = get_all_prop(dict_70, prop_lst_f_p)
```

```
100%|| 5597872/5597872 [00:33<00:00, 167335.30it/s]
```

n [65]:
```
print(all_lst)
print(count1,count2,count3,count4,count5,count6,count7,count8,count9,count10,count11,
```

```
[0.1786631585541565, 0.40950552088334136, 0.5446947674418605, 0.6213086276780544, 0.6
72, 0.7228637413394919, 0.6693306693306693, 0.4605077574047955, 0.4122257053291536, 0
173913, 0.5747463961558996, 0.030049945651426133]
47889 6249 2752 1727 1235 939 866 1001 2836 8932 23961 11500 3746 72679
```

F or N -F or N

n [70]:
```
# get time_dict with filter of 'F' &'P'
time_dict_f_p = get_time_dict('P')

# get numbers of count in each bin, for F & P
# and a dict for each bin for counting num of diff stations in each bin afterwards
# update count and dict_{} for F & P
count1,count2,count3,count4,count5,count6,count7,count8,count9,count10,count11,count12,count13,co

# get percentage of num of diff stations in each bin as a list
prop_lst_f_p = get_prop()
all_lst = get_all_prop(dict_70, prop_lst_f_p)
```

```
100%|| 5597872/5597872 [00:17<00:00, 325527.43it/s]
```

n [71]:
```
print(all_lst)
print(count1,count2,count3,count4,count5,count6,count7,count8,count9,count10,count11,count12,coun
```

```
[0.3175411475857052, 0.5074780814853017, 0.5794285714285714, 0.6620370370370371, 0.679518072289152,
2, 0.6594827586206896, 0.6156716417910447, 0.31405895691609975, 0.26571936056838363, 0.28797242695
556962, 0.476078431372549, 0.043741043055642095]
10997 1939 875 648 415 298 232 268 882 2815 8414 3950 1275 16049
```

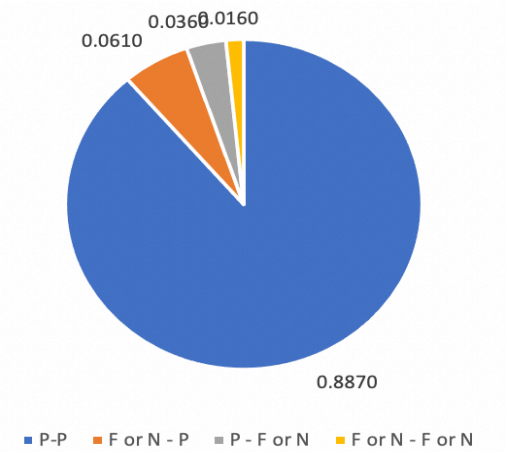
2 Time since last pass calculation for year 2015 and 2016 for PA

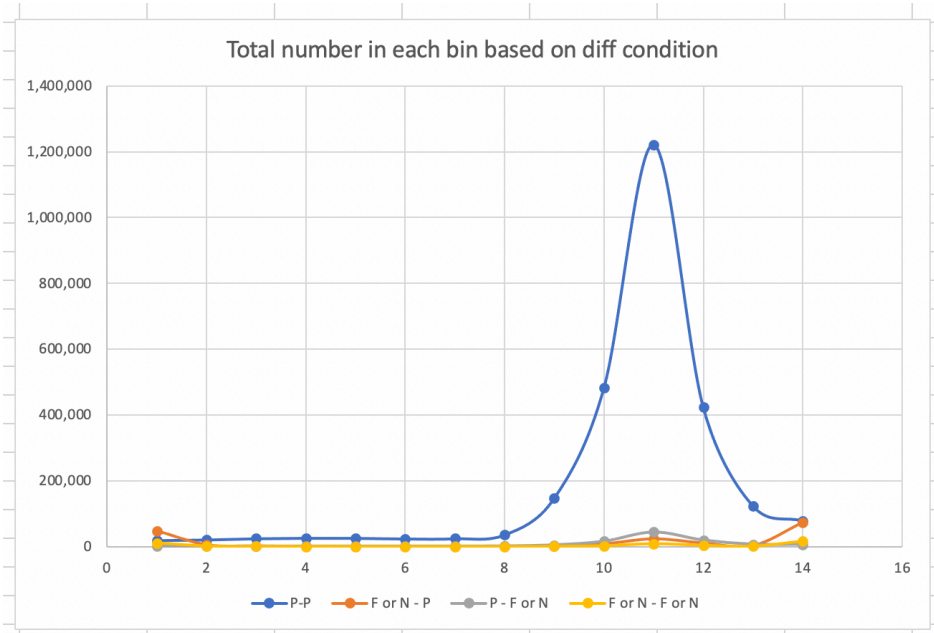
Bins' length 5 weeks

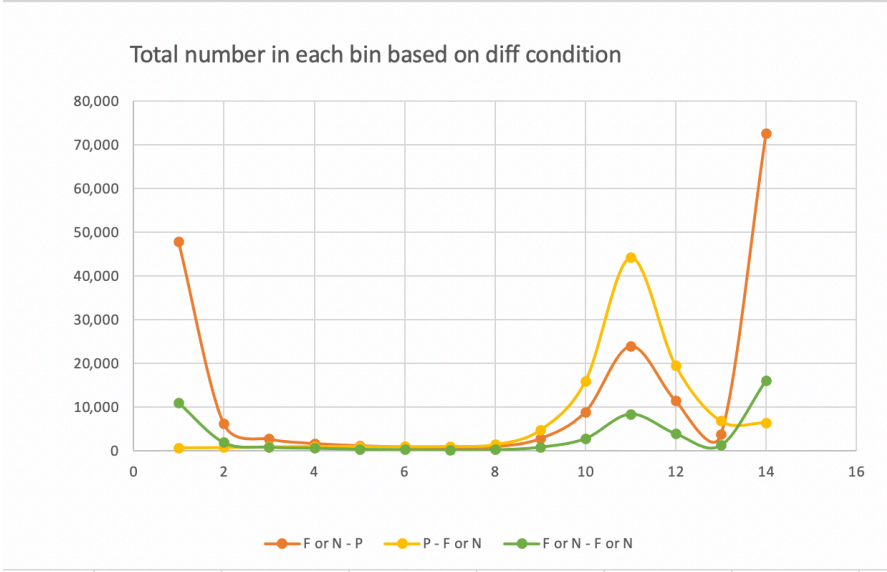
Penn	year 2015 2016												
	Total	P-P			F or N-P			P-F or N			F/N-F/N		
Bin weeks (time since last pass)	original records in each bin	numbers in each bins, all stations time since last pass(P-P)	# P-P / total in bin(5 weeks)	P-P num of diff stations/total in each bin	numbers in each bins, all stations time since last pass(F or N-P)	# F or N-P / total in bin(5 weeks)	F or N-P num of diff stations/total in each bin	numbers in each bins, all stations time since last pass(P-F or N)	# P-F or N/ total in bin(5 weeks)	P-F or N num of diff stations/total in each bin	numbers in each bins, all stations time since last pass(F/N-F/N)	(F/N - F/N) / total in bin(5 weeks)	F/N-F/N num of diff stations/total in each bin
0-5	78,861	19,240	0.2440	0.7275	47,889	0.6073	0.1787	735	0.0093	0.7265	10,997	0.1394	0.3175
5-10	30,276	21,280	0.7029	0.7410	6,249	0.2064	0.4095	808	0.0267	0.7797	1,939	0.0640	0.5075
10-15	29,388	24,835	0.8451	0.7224	2,752	0.0936	0.5447	926	0.0315	0.7311	875	0.0298	0.5794
15-20	29,809	26,381	0.8850	0.6963	1,727	0.0579	0.6213	1,053	0.0353	0.7512	648	0.0217	0.6620
20-25	28,818	26,130	0.9067	0.6776	1,235	0.0429	0.6672	1,038	0.0360	0.7360	415	0.0144	0.6795
25-30	26,423	24,192	0.9156	0.6929	939	0.0355	0.6912	994	0.0376	0.7384	298	0.0113	0.6745
30-35	26,892	24,734	0.9198	0.6984	866	0.0322	0.7229	1,060	0.0394	0.7396	232	0.0086	0.6595
35-40	38,898	36,176	0.9300	0.5843	1,001	0.0257	0.6693	1,453	0.0374	0.6827	268	0.0069	0.6157
40-45	156,263	147,769	0.9456	0.3732	2,836	0.0181	0.4605	4,776	0.0306	0.4732	882	0.0056	0.3141
45-50	510,173	482,492	0.9457	0.3271	8,932	0.0175	0.4122	15,934	0.0312	0.4139	2,815	0.0055	0.2657
50-55	1,297,685	1,221,046	0.9409	0.3586	23,961	0.0185	0.4356	44,264	0.0341	0.4522	8,414	0.0065	0.2880
55-60	458,589	423,607	0.9237	0.4351	11,500	0.0251	0.5269	19,532	0.0426	0.5369	3,950	0.0086	0.4003
60-65	136,159	124,278	0.9127	0.4838	3,746	0.0275	0.5747	6,860	0.0504	0.5969	1,275	0.0094	0.4761
>65	174,235	79,019	0.4535	0.5769	72,679	0.4171	0.0300	6,488	0.0372	0.6802	16,049	0.0921	0.0437
sum	3,022,469	2,681,179	88.7%.	/	186,312	6.1%.	/	105,921	3.6%.	/	49,057	1.60%	/

2 Time since last pass calculation for year 2015 and 2016 for PA

Bins' length 5 weeks



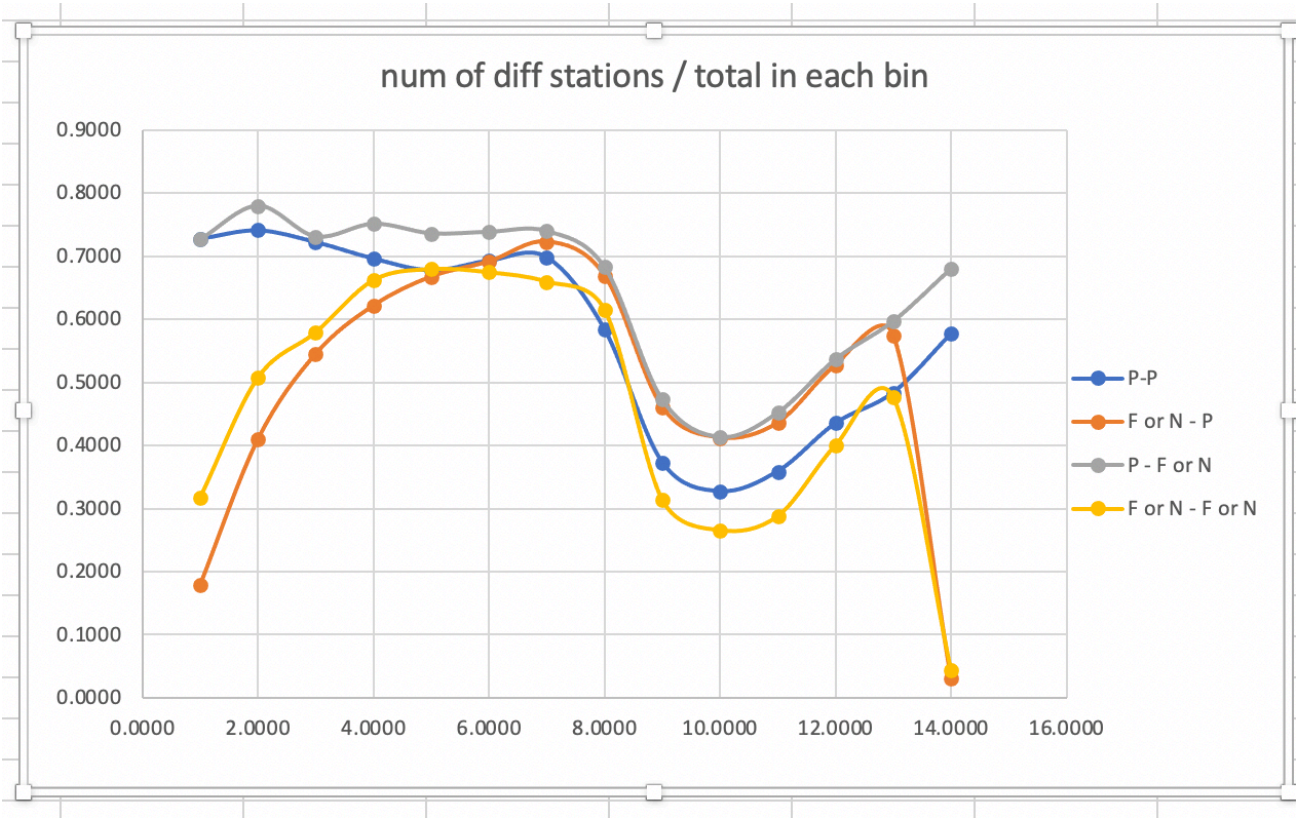




- ***Most test happens in 50-55 weeks, a year***

2 Time since last pass calculation for year 2015 and 2016 for PA

Bins' length 5 weeks



- ***0-5, 5-10, 10-15 weeks, if first result is F, number of switching stations is less than that is first result is P, which means drivers would be more likely going to the same station after the first one fails.***
- ***Same situation happens in bin of > 65 weeks.***

Thanks!

Lin Lyu