For office use only
T1 _____
T2 _____
T3 _____
T4 _____

For office use only
F1 _____
F2 _____
F3 _____
F4 _____

2016

19th Annual High School Mathematical Contest in Modeling (HiMCM)

Summary Sheet

Team Control Number: 6879

Problem Chosen: A

Summary

A triathlon is a three-stage competition involving continuous swimming, cycling and running in its most popular form. Athletes are required to complete these 3 stages without succession in the shortest amount of time. The sport requires tremendous tenacity and endurance. Transition time between swimming and cycling is counted towards the total time and is known as T1, similarly transition time between cycling and running is called T2. Athletes start in waves, and each wave should not affect any other wave. Our objective was to organize an open triathlon competition that suits both the requirements of the government and our sponsor which includes: 1.the duration of the road closed time should not exceed 5.5 hours; 2.each wave of athletes should prevent hindering other waves. There would be approximately 2000 people with different genders, ages and status (professional, premier or open).

To seek for a viable arrangement of 2000 athletes, it is necessary to estimate the size of each category of athletes (discriminated by gender, age and status) as well as their swimming time, biking time, running time, and transition time. 8 different categories and an appropriate age were randomly assigned to the 2000 athletes based on the data set of the previous competition. This was done by a random process based on the cumulative probability. We constructed a linear regression equation to predict the time each athlete took by using the assigned category and age as external variables to complete different stages.

According to our definition, congestion happened when a participant caught up with another participant starting before him. Total number of congestions happened in the competition was used as a criterion to evaluate schedules of ordering. We assumed that the best way to prevent congestion was to arrange the faster triathletes before the slower ones; this was done by ranking the average total time of each category. A model was built to compute the duration of the triathlon and the number of congestions, which were directly influenced by the size of each wave and the time gap between waves. We tested different sizes of waves and the time gasp between each wave to select the optimal solution from a variety of solutions.

Adjusting the race distances of the competition might bring advantages in terms of congestion and road closure time. We also used our model to determine which kind of changes could contribute to reduce the congestion and road closure time most effectively.

Dear Mayor:

Here is our summary of the Triathlon which will be held in the city.

This time, we expect about 2,000 people attending the triathlon. There will be some professional athletes attending the triathlon. We also hope to attract some famous triathlon athletes in the world either professional or premier or open athletes. So there will be a lot of people arriving at the city.

Based on our calculations, about 67% of people will be male open athletes and about 28% of people will be female open athletes. Male professional athletes use the less time to finish the triathlon, so they will start the triathlon first in order to avoid congestion. The athletes will begin the triathlon one by one and this schedule of start time can reduce the congestion and road closure time.

Next, due to our calculation, we suggest that the local road should be closed before the triathlon begins because the athletes won't spend too much time (no more than 20min) in the water before they begin cycling.

We also find that adjusting distances can improve the congestion and reduce the road closure time. You can consider reducing the distance of a section to meet your requirements.

Yours,

Team 6879

Here is the schedule of the triathlon.

11 a.m.	All the athletes need to arrive to be identified for the triathlon. Athletes will take a urinalysis. Any athletes who fail to pass the test will be disqualified.
12p.m.	Athletes who pass the urinalysis will receive their number by and start time according to their category.
2 p.m.	The race begins. The athletes will begin the triathlon in order of their number. The athletes should wear their swimming suit before the event starts.
8 p.m.	Prize-awarding ceremony We will award the 1st, 2nd and 3rd prize. Every athlete who attend the race will be given a trophy.
8:30p.m	The ceremony ends.

Restatement

We were asked to organize a traditional open triathlon at our local town with the mayor. There would be about 2000 participants and the triathlon competition consists of 1.5 km of swimming, 40 km bike riding and 10 km of running. The objective was to attract professional athletes and premier athletes. The race begins with the swimming event and participants start the race in a sequence of waves of groups of swimmer at intervals of some number of minutes apart. The sponsor wants us to minimize congestion on the course. In other words, participants should be able to proceed without hinndrance during each phase of the triathlon. At the mean time, mayor of our town wants to minimize the length of time the local roads in the town are closed for the cycling and running portions of the triathlon to no more than 5.5 hours. Thereby, our objective is to minmizes the congestion of the course while allow each and every preticipates be able to finish their race withinin 5.5 hours.

Assumptions and Justifications

Our assumptions and justifications are listed below. They are all based on rational reasons and will be mentioned in relevant sections afterwards.

Assumption 1: Roads are closed as soon as the competition starts, and ends as the last participant reaches the finish line. Participants will be asked to leave the competition after 5.5 hours (in the case of extremely slow competitors).

Justification: The competition starts with swimming section. Usually, athletes spend about 10 to 20 minutes swimming, which is a very short time. This time is not enough for the local road to be closed. So the road needs to be closed before the event begins.Also, the event only takes 5.5 hours. After this time, the road will be open again. There may be some athletes who are very

slow. So to prevent the delay of the traffic, those very slow competitors may be cleaned out.

Assumption 2: There would be exactly 2000 participants

Justification: This is due to our predictions based on the recent events.

Assumption 3: A congestion is defined by when a participant catch up with another participant starts before him

Justification: The athletes begin the event in different groups which consist of several people. Each group begins the event in order. However, different athlete has different velocity. There may be some athletes catching up the previous group. This is called a congestion.

Assumption 4: A wave consists one or more and less than two hundred participants.

Assumption 5: Wave sizes should be identical (except the last wave might be different).

Justification: The number of people each wave begins the trathlon should be the same and it's easy to control the whole situation.

Assumption 6: The time gap between the waves should be the same and timed in intergral seconds.

Justification: It's reasonable and fair for the athletes.

1. Data description

We analyzed the provided data in various ways. We assumed that the total race time was affected by gender, age and status, which would be justified later in the paper. Therefore the objective of our analysis was to visually demonstrate the data and to calculate the percentage of each gender, age group, and status, which was crucial to the construction of our model. We have utilized pie graphs to demonstrate the percentage of each gender (male/female) in Figure 1.1.

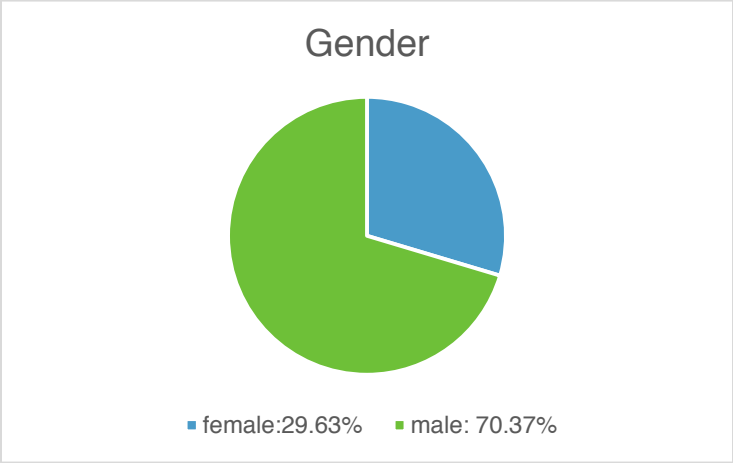

Figure 1.1

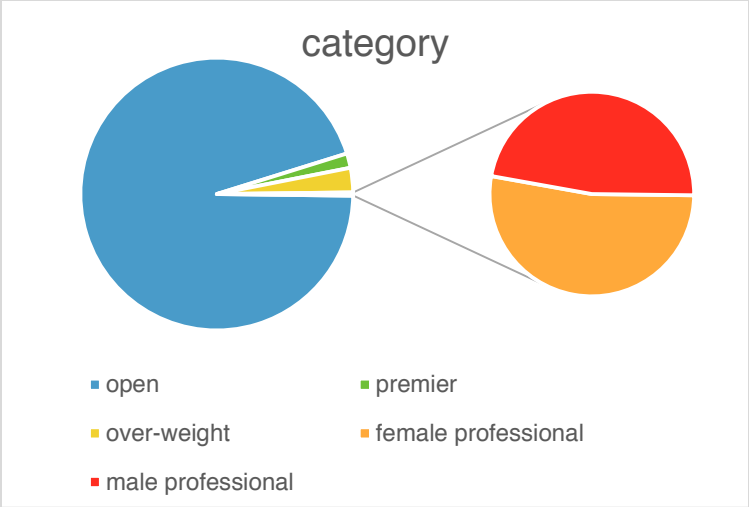

Figure 1.2

Table 1.1

Status	Percentage
Professional male	0.2%
Premier male	1.0%
Open male	67%
CLY	1.9%
Professional female	0.18%
Premier female	0.56%
Open female	28%
ATH	1.0%

Figure 1.1 demonstrates that there was a majority of male participants in the previous competition; therefore we can deduce that the majority of the competitors would be male as well in this triathlon competition. Figure 1.2 and Table 1.1 shows that there were a lot of male and female open athletes taking part in the recent event. The second largest group was the premier group and the smallest was the professional group. Through this we can deduce that not many professional triathletes would participate in our open triathlon.

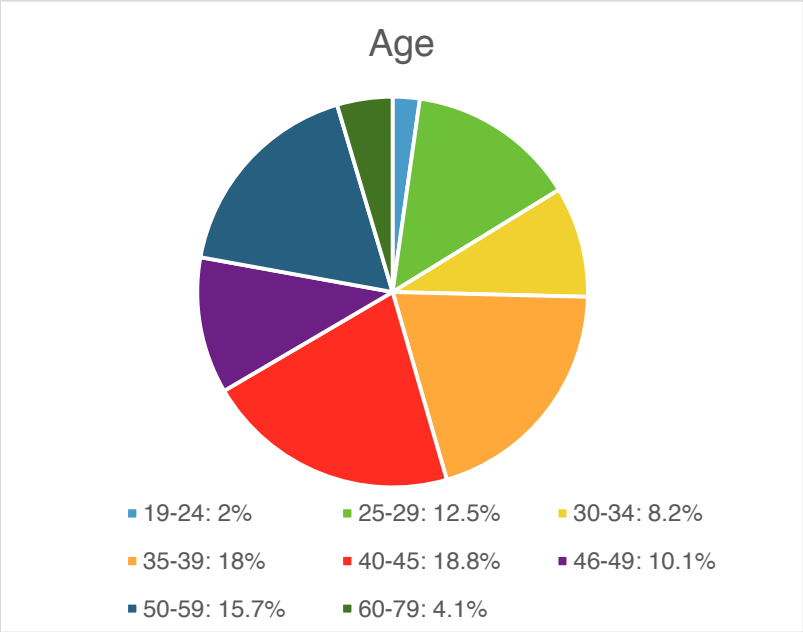

Figure 1.3

Figure 1.3 shows that most people who joined the recent triathlon were young and middle-aged people from 25 years old to 59 years old, because 95.9% of people were between these ranges of age. So we concluded that there would be about the same percent of middle age people attending this event.

Figure 1.4 shows the range of completion time and the frequency of participants of each range of completion time. As we can see from the graph, most players complete the triathlon between 9000 and 12000 seconds.

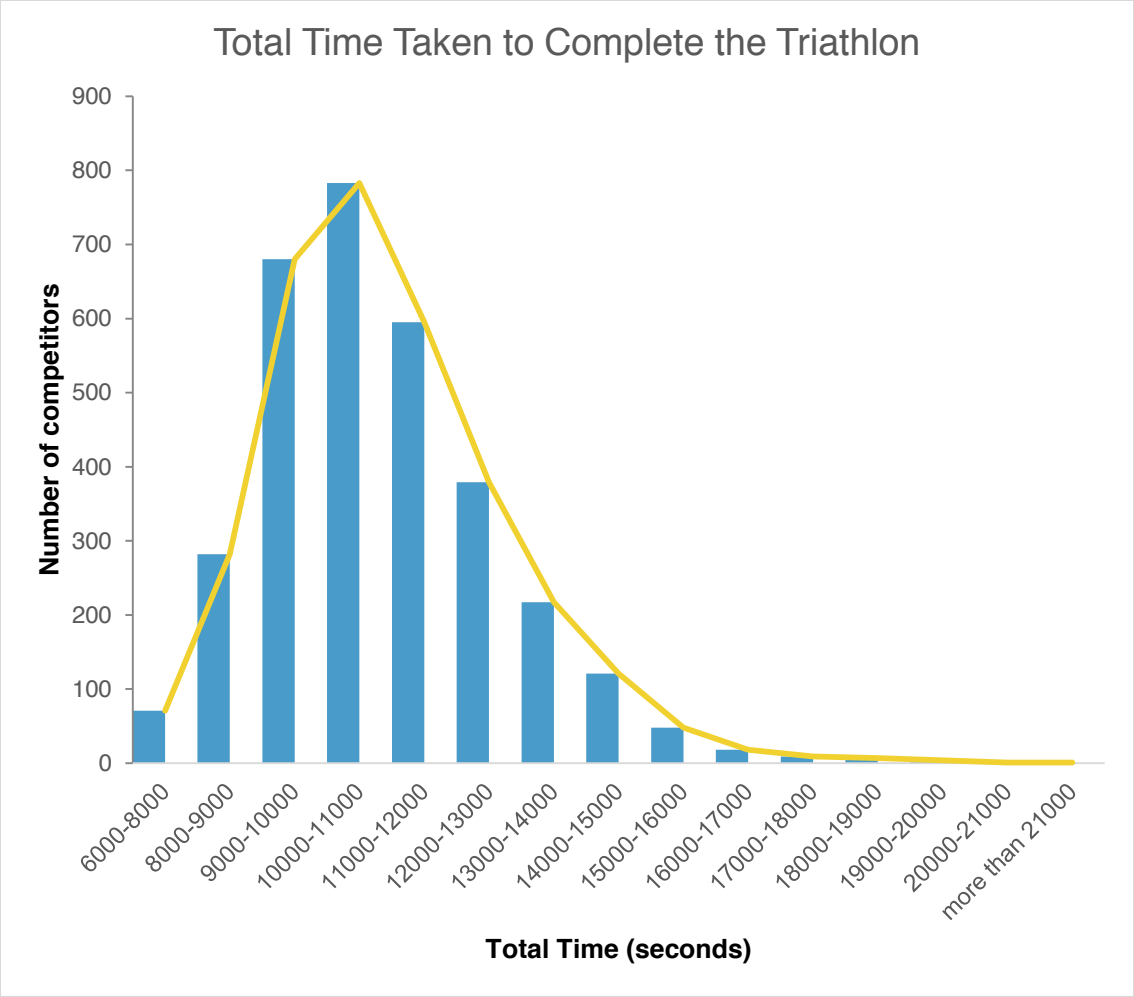
Total Time Taken to Complete the Triathlon
Number of competitors
900
800
700
600
500
400
300
200
100
0
6000-8000
8000-9000
9000-10000
10000-11000
11000-12000
12000-13000
13000-14000
14000-15000
15000-16000
16000-17000
17000-18000
18000-19000
19000-20000
20000-21000
more than 21000
Total Time (seconds)

Figure 1.4

2. Part I

Categorization

The categorization of participants is necessary in order to arrange them in different waves accordingly. We have decided to categorize participants by their status and gender. Our categories include professional male (category 1), professional female (category 2), premier male (category 3), premier female (category 4), open male (category 5), CLY (category 6), open female (category 7), ATH (category 8).

Distribution of Age and Category

We have estimated the size of each category and age group by calculating the percentage of competitors based on the given data set, and generated a cumulative probability table (Table 2.1 and Table 2.2) in order to assign category and age to each of the 2000 people. The cumulative probability was obtained by calculating the percentage of the age groups and catagories in the previous data set.

The process was as follows: we randomly generated 2000 numbers, each followed a 0 to 1 uniform distribution; we asssigned the corresponding age and category to each participant accoding to the probability interval in which the random numbers fell into.

We have given 2000 participants an appropriate age and category based on rational probability calculations. We therefore needed to obtain the time each athlete would take to complete the triathlon. This was done by linear regression.

Table 2.1

Age	cumulative probability	Age	cumulative probability	Age	cumulative probability	Age	cumulative probability
19	0.0009325	36	0.412807	53	0.8877836	70	0.9956481
20	0.0031085	37	0.443581	54	0.8989742	71	0.9962698
21	0.0046627	38	0.4793286	55	0.911719	72	0.9965807
22	0.0074604	39	0.5132111	56	0.9250855	73	0.9965807
23	0.0118122	40	0.5470936	57	0.9341001	74	0.9972024
24	0.0198943	41	0.5766242	58	0.9468449	75	0.9975132
25	0.0379235	42	0.6089524	59	0.9571029	76	0.9975132
26	0.0578178	43	0.6400373	60	0.9645633	77	0.9981349
27	0.0848617	44	0.6726764	61	0.9723345	78	0.9987566
28	0.1106621	45	0.7009636	62	0.9804165	79	0.9990675
29	0.1451663	46	0.7280075	63	0.9841467	80	0.9993783
30	0.1784271	47	0.7525645	64	0.9869444	81	0.9996892
31	0.217594	48	0.7733914	65	0.9894311	82	0.9996892
32	0.2589369	49	0.8016786	66	0.9909854	83	1
33	0.2959279	50	0.8302767	67	0.9919179		
34	0.3332297	51	0.8507927	68	0.9931613		
35	0.3742617	52	0.8713087	69	0.993783		

Table 2.2

Category	Cumulative probability
male professional	0.0021759
male premier	0.0174075
male open	0.6847995
CLY	0.7034504
female professional	0.7053155
female premier	0.7109108
female open	0.9900528
ATH	1

Linear Regression

We used linear regression to predict the time each athlete would take to complete different stages of the competition. In this step, we predict follwing time variables:

- Swim time: t_1
- Transition time from swim to bike: t_2
- Bike time: t_3
- Transition time from bike to run: t_4
- Run time: t_5

Based on follwing covariates:

- Age of the participants: A
- Male professional participants: B
- Male premier participants: C

- Male open participants: D
- CLY Participants: E
- Female professional participants: F
- Female premier participants: G
- Female open participants: H
- ATH participants: I

The relationship between time taken in each section and the number of each kind of participants are shown in the linear regression model below:

$$t_i = \beta_0^i + \beta_1^i * A + \beta_2^i * B + \beta_3^i * C + \beta_4^i * D + \beta_5^i * E + \beta_6^i * F + \beta_7^i * G + \beta_8^i * H + \beta_9^i * I$$

(i=1, 2, 3, 4, 5)

In this model β_j^irepresents the coefficient related to each covariate in our linear regression model (e.g.for each time stage i, β_0 represents the intercept, β_1 is the coefficient of "age", $\beta_{2,3,4\ldots9}$ are the coefficients of the remaining 8 variables).

The results from those 2000 random generations are demonstrated below in Figure 2.1 and Figure 2.2. The structure of athletes is similar to the data sets. Figure 2.1displayed the relationship between age and total time in our estimation. Regardless of some special cases athletes between 25 and 55 are the fastest, justifying our assumption in the data analysis section. The total time required to complete the triathlon increases as the age of the triathlet gets smaller than 25 or exceed 55. There are two exceptional cases represented by the 2 conspicuous dots that do not follow the seuquence; the dots represent a very fast 21 years old triathlet and a very fast 68 years old triathlet. Figure 2.2 demonstrates the relationship between categories and average total time, which indicates the order of completion time of the categories.

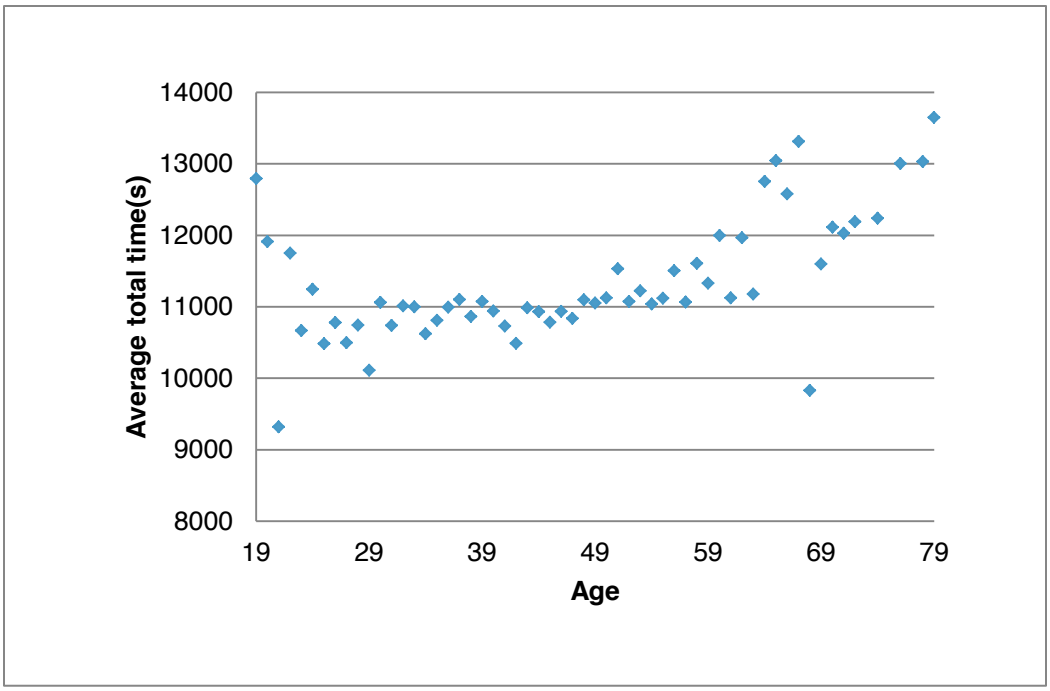

Figure 2.1

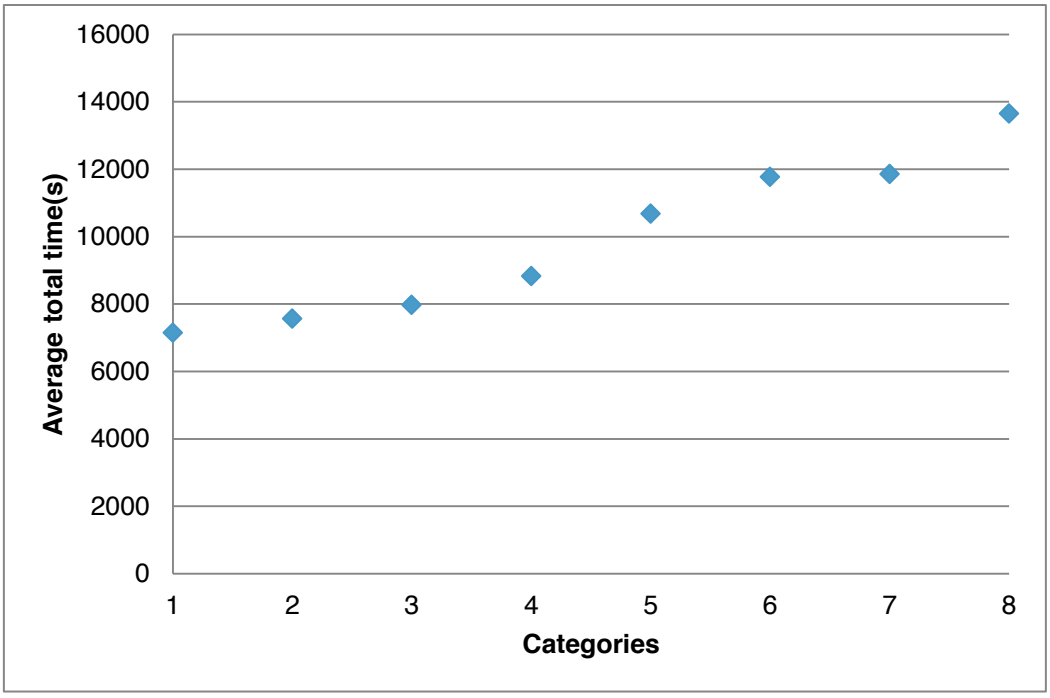

Figure 2.2

Arrangement of the waves and start times

In order to establish a well organized activity, we decided to mainly focus on minimizing congestions while keeping the length of road closed time less than 5.5 hours. Time gap between each wave is very limited, therefore each time we increased the time gap we decreased the number of congestions tremendously.

Our main concept was to arrange the faster participants before the slower participants to prevent possible influences. Therefore we have determined the order based on category and age, which are the two most influential factors of triathletes' speed. We have utilized the average total time of each category calculated by the linear regression model listed below:

Table 2.3

Categories	Average Total times(s)
1	7148.5
3	7971.6
5	10676.6
6	11771.2
2	7557.6
4	8826.8
7	11852.1
8	13649.8

A primary arrangement based on category was determined based on the table above. A secondary arrangement was determined by the age of the participants. Average total times of

each age group are shown in Table 2.4. After all the athletes were assigned according to their category, we then further assigned athletes in each category to groups according to their age. The age groups and their rankings are shown in Table 2.5.

Table 2.4

Age	Time/s	Age	Time/s	Age	Time/s
19	12794.3	39	11075.4	59	11331.2
20	11910	40	10940.0	60	11997.3
21	9320	41	10727.9	61	11125.2
22	11750.5	42	10487.9	62	11966.8
23	10668.04	43	10985.6	63	11178.9
24	11244.04	44	10929.1	64	12752.8
25	10481.04	45	10783.2	65	11042.3
26	10777	46	10934.3	66	12579.4
27	10496.2	47	10838.6	67	13314
28	10740.9	48	11095.7	68	9830.3
29	10110.6	49	11052.8	69	11597.5
30	11060.8	50	11124.4	70	12114.2
31	10738.02	51	11528.4	71	12028
32	11012.2	52	11074.0	72	12190
33	10996.7	53	11220.2	74	12237
34	10622.3	54	11036.2	76	13305
35	10811.1	55	11117.8	78	13032
36	10993.7	56	11504.6	79	13649
37	11100	57	11063.2		
38	10864.9	58	11605.9		

Table 2.5

Age group	19-28	29-38	39-48	49-58	59-68	69-79
Rank/Evaluation	3	1	2	4	5	6

Models and computation process

As we have decided the order of start time of all the athletes, we could determine the start time of each athletes under each arrangement. We then used the estimated time produced by linear

regression to calculate the following time points: (1) finish time of swimming; (2) start time of cycling; (3) finish time of cycling; (4) start time of running; (5) finish time of running. We then compared these time points between each pair of athletes. In each pair, if any of these time points of the athlete starts earlier was later than the corresponding time points of the athlete starts later, it meant at least one encounter happened, and the total number of congestions would be added by one. We then compared different kinds of arrangement in terms of average congestions per person.

Conclusion

We have tested three kinds of arrangements: (1) the wave size was one, which meant athletes would start one by one; (2) the wave size was 5; (3) the wave size was 10. We tried different integer time gap with the constraint that total road closure time was less than 5.5 hours. We then found 3 best arrangements from a myriad of different solutions as shown in Table 2.6. From Table 2.6, we could conclude that the arrangement that athletes starts one by one with a time gap of 2 seconds was the best arrangement in terms of congestion per person.

Table 2.6

Arrangement no.	**1**	**2**	**3**
Wave size	1	5	10
Time gap (seconds)	2	14	25
Congestion per person	1.575	1.8045	1.745
Total Time (hours)	5.458	5.478	5.309

3. Part 2

From Part 1, we have obtained the best arrangement, which was to let athletes start one by one with the time gap of 2 seconds. In this part, we tested the effect of change of distances under this arrangement. Same modelling and computation process as in Part I was used in Part 2.

We tried to change the distances of swimming, biking, and running by the same percentage (10%, 20%, and 30%). To make the problem simpler, we only changed the distance of one section at each time, which meant that to change two or three sections were not allowed in this part. Table 3.1 and Figure 3.1 display the average congestion per person under different changes. From the table and figure, we could see that decreasing the distances of swimming, biking, or running reduced the average congestion while increasing the distances increase the average congestion. As to reduce the congestion was the motivation of this part, we could conclude that to decrease the distances would bring advantages and the more to decrease the more significant the effect

Table 3.1

Change	Swim	Bike	Run
-30%	1.342	1.56	2.38
-20%	1.41	1.58	2.49
-10%	1.4645	1.61	2.64
0	1.613	1.66	2.78
10%	1.76	1.71	2.92
20%	2.1	1.78	3.06
30%	2.2	1.87	3.21

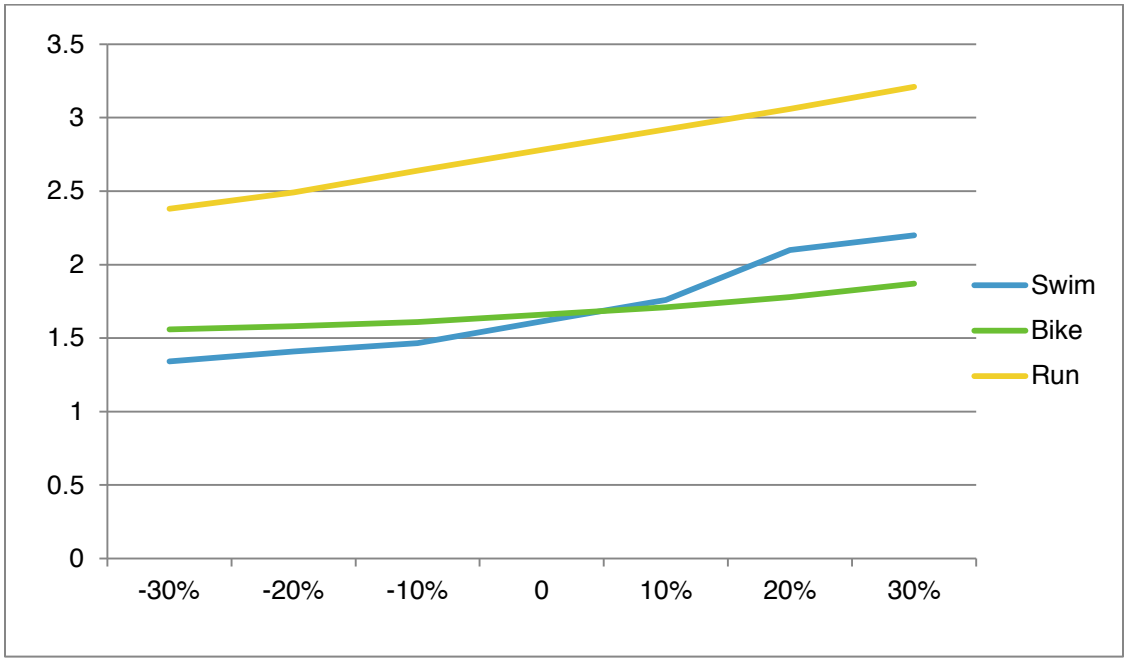

Figure 3.1

4. Reference

http://results.active.com/events/transamerica-chicago-triathlon/international-elite-amateur/expanded

http://www.chicagotriathlon.com/results-photos/

http://www.triathlon.org/about/documents

http://3oxwmd40w09z1u8ihn1492m3.wpengine.netdna-cdn.com/wp-content/uploads/sites/26/2016/08/2016-Chicago-Tri-Wave-Assignments-8-18-16.pdf

5. Appendix

Part 1: Linear Regression Matlab Fomula

```
x=HIMCM(1:3217,2:3);
y1= HIMCM (1:3217,5);
y2= HIMCM (1:3217,7);
y3= HIMCM (1:3217,9);
y4= HIMCM (1:3217,11);
y5= HIMCM (1:3217,13);
x2=[ones(3217,1),x(:,1),dummyvar(x(:,2))];
[b5,bint5,r5,rint5,stats5]=regress(y5,x2)
t1=8.59-2.40A-0.71B+2.24C+6.39D+7.79E+0F+2.44G+5.82H+6.39I
t2=1.04+3.37A-6.63B+43.60C+2.15D+2.37E+0F+78.20G+3.68H+4.63I
t3=3.93+3.71A-3.11B+57.54C+1.01D+1.21E+0F+5.17G+1.82H+2.46I
t4=4.14+1.49A+3.41B+19.61C+1.15D+1.41E+0F+31.17G+1.60H+2.08I
t5=1.85+13.28A-1.18B+78.35C+1.17D+1.87E+0F+3.46G+1.38H+2.28I

x=x(1:2000,1)
x5=dummyvar([Group])
x6=[ones(2000,1),AGE,x5]
A=x6(1:2000,2)
B=x6(1:2000,3)
C=x6(1:2000,4)
D=x6(1:2000,5)
E=x6(1:2000,6)
F=x6(1:2000,7)
G=x6(1:2000,8)
H=x6(1:2000,9)
I=x6(1:2000,10)
SEM = std(r5)/sqrt(length(r5));
ts = tinv([0.025  0.975],length(r5)-5);
CI5 = mean(r5) + ts*SEM;
for i=1:2000
ei=rand*(Cl5(1,2)-Cl5(1,1))+Cl5(1,1)
e5(i)=ei;
end
tp1=t1-e1'
tp2=t2-e2'
tp3=t3-e3'
tp4=t4-e4'
tp5=t5-e5'
```

Part 2:

```
chase=zeros(2000,2000);
for x=1:1900
   for y=x+1:x+100
      if t1(x,1)>t1(y,1)|t2(x,1)>t2(y,1)|t3(x,1)>t3(y,1)|t4(x,1)>t4(y,1)|t5(x,1)>t5(y,1)
         chase(y,x)=1;
```

```
        end
    end
end
chasee=zeros(2000,1);
for x=1:2000
    chasee(x)=sum(chase(x,:));
end
mean=(sum(chasee))/2000;


chase=zeros(2000,2000);
for x=1:1950
    for y=x+5:x+50
        if t20(x,1)>t20(y,1)|t21(x,1)>t21(y,1)|t22(x,1)>t22(y,1)|t23(x,1)>t23(y,1)|t24(x,1)>t24(y,1)
            chase(y,x)=1;
        end
    end
end
chasee=zeros(2000,1);
for x=1:2000
    chasee(x)=sum(chase(x,:));
end
mean=(sum(chasee))/2000;
```