交通数据分析期末大作业

1852127 赵冠华
2018 级交信

一、 数据采集。

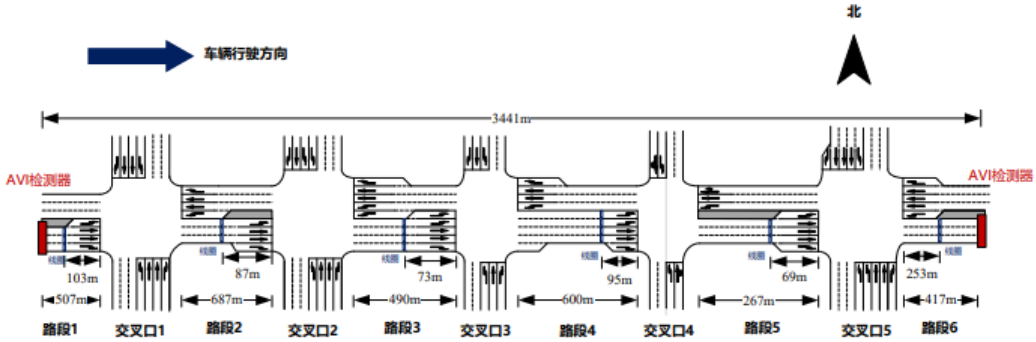


图 1 干道基本示意图

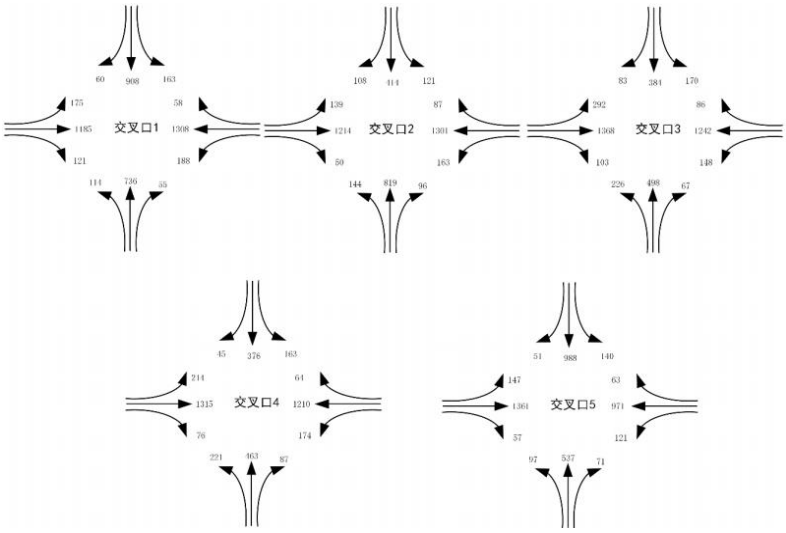


图 2 交叉口机动车交通量

本次作业数据来源于某城市干道（如图 1 所示），该干道包含 5 个信号控制交叉口和 6 个路段，限速为 60km/h，各交叉口流量与信号配时方案如图 2 与表 1 所示。

每个路段上安装有线圈检测器，干道两端安装有 AVI 设备，干道车流中有 10%的出租车浮动车。因此，我们可以得到线圈、浮动车、AVI 检测数据估计得到的各路段 5 分钟平均行程速度与真实的行程速度数据。

	相位 1：南北直行	相位 2：南北左转	相位 3：东西直行	相位 4：东西左转	周期
交叉口 1	39s	12s	50s	14s	131s
交叉口 2	36s	11s	56s	12s	131s
交叉口 3	34s	15s	50s	12s	127s
交叉口 4	36s	14s	49s	12s	127s
交叉口 5	32s	18s	45s	19s	130s

注：所有交叉口每个相位的黄灯均设为 3 秒、全红设为 1 秒，损失时间 4 秒/相位、16 秒/周期；所有交叉口的右转车辆都不受信号灯控制（即红灯期间可右转）。

表格 1 各交叉口信号配时方案（单位：秒）

二、 干线平均行程速度与统计分析。

在本章中，我将以真实行程速度数据为基础，借助 Matlab 工具先以路段长度加权的方式计算 168 个 5 分钟时间间隔的整条干线的平均行程速度，再对其进行基本统计分析与可视化。

1. 整条干线的平均行程速度。

首先计算各路段的权重，计算依据为路段长度，计算结果如下表 2 所示。

路段编号	路段1	路段2	路段3	路段4	路段5	路段6
长度/m	507	687	490	600	267	417
总长/m	2968					
权重	0.1708	0.2315	0.1651	0.2022	0.0900	0.1405

表格 2 路段权重计算

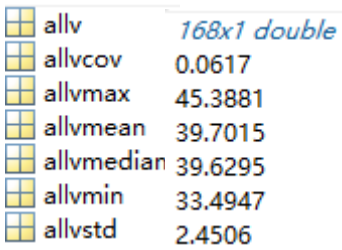
接着计算每个时间间隔整条干线的平均行程速度(单位 km/h)，公式如下：

$$v_{干线} = \sum_{i=1}^{6} v_i * w_i$$

将计算结果储存为新的表格“allv.xls”，以便后续处理。

2. 基本统计分析。

我将对计算得到的全线平均速度进行基本的统计分析，得到算数平均值、中列数、中位数、标准差、变异系数、最大值、最小值与样本数。其中，样本数为 168，中列数指样本中极大值与极小值的平均，结果为 39.4414，变异系数指标准差与平均值之比，结果如下图所示。



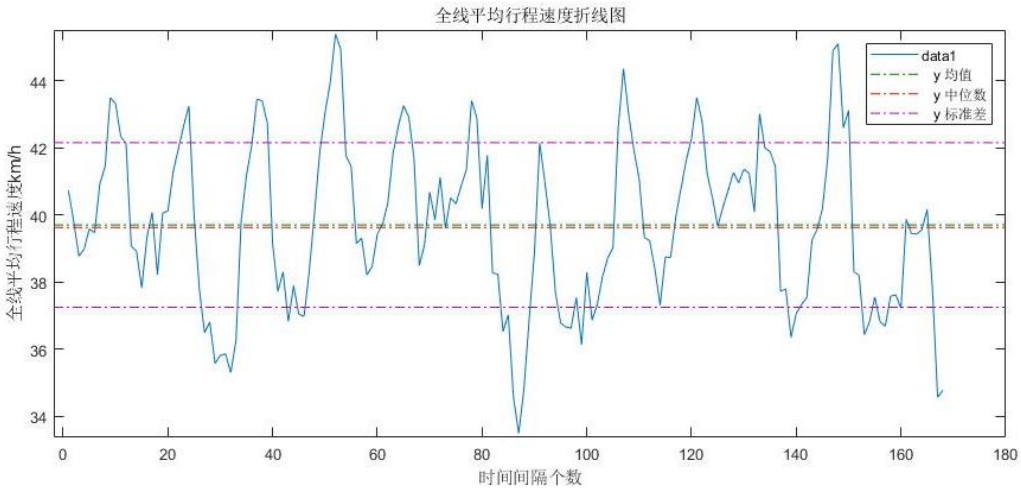


图 3 全线平均行程速度折线图

3. 箱图。

使用五数概括法描述数据，求五分位数，并绘制箱图如下所示。

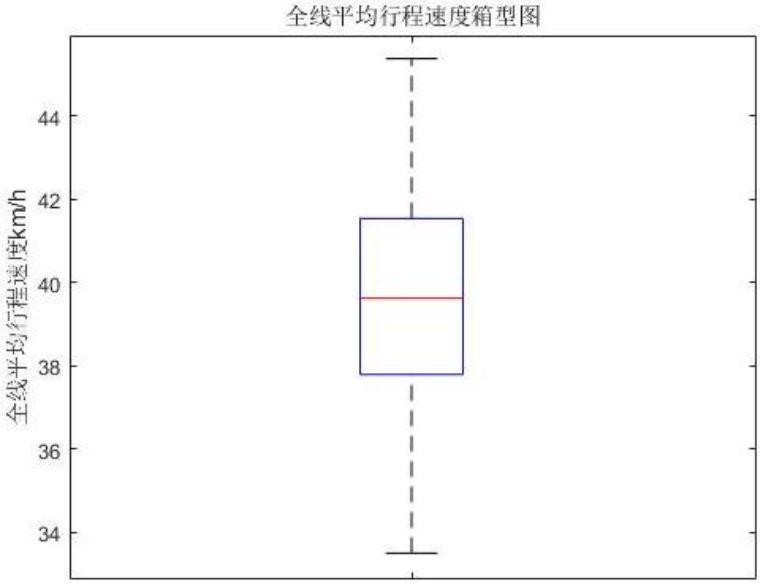


图 4 全线平均行程速度箱图

Q1	37.5368
Q2	39.0553
Q3	40.2019
Q4	41.9129

Min	33.4947
Q1	37.7923
Median	39.6295
Q3	41.5424
Max	45.3881

表格 3 五数概括法

4. 区间频数与累计分布频率。

以 10km/h 为区间长度，计算整条干线平均行程速度的区间频数与累计分布频率。由上分析已知速度最小为 33.4947km/h，最大为 45.3881km/h，故只分两组即满足要求。编辑函数“pinshutongji.m”如下，并得到计算结果。利用 histc 与 cumsum 函数得到更平滑的累计分布频率曲线，绘制结果如下。

编辑器 - D:\matlabhomework\dalast\pinshutongji.m

byes.m | quehsi.m | zhifangtu.m | untitled.m | pinshutongji.m

```
function  [d,zb1,ps] = pinshutongji(a,b)
% a: 输入的矩阵
% b: 要分成的组别，默认为10
% 输出为: d(组别zb 频数ps 累积lj(%)) zb1(组别数值矩阵double) ps(频数数值矩阵double)
    if nargin==1
        b = 10;
    end
    ps = zeros(b,1);lj = zeros(b,1);
    zb = (min(a):(max(a)-min(a))/b:max(a));zb = zb'; % 建立组别
    for i = 1:1:length(a) % 计算频数
        temp = a(i);
        for j = 2:1:b+1
            if temp<=zb(j,1)
                ps(j-1) = ps(j-1)+1;
            end
        end
    end
    asum = length(a);
    for i = 1:1:b % 计算累积
         lj(i) = (ps(i,1)/asum)*100;
    end
    c = ["          组别          "," 频数","累积(%)"];
    zb1 = [zb(1:end-1,1),zb(2:end,1)];
    zb2 = num2str(zb1); zb2 = cellstr(zb2);
    ps = [ps(1,1);ps(2:end,1)-ps(1:end-1,1)];
    ps1 = num2cell(ps); lj = num2cell(lj);
    d = [c;zb2,ps1,lj];
end
```

	1	2	3
1	组别	频数	累积(%)
2	29.4414 39.4414	78	46.4286
3	39.4414 49.4414	90	100

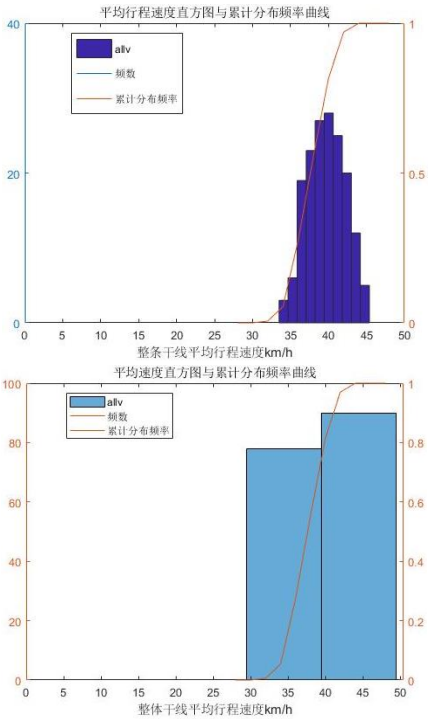


图 5 直方图与累计分布频率曲线

三、 线圈数据预处理。

在本章，我将对 Sheet1 中线圈检测数据进行预处理，包括剔除异常数据以及对缺失数据的补充。

1. 剔除异常数据。

由于从线圈得到的检测数据包括平均速度、流量、以及占有率，故剔除异常数据时可以采用的方法有许多种，我选择采用独立判断法，之后再对速度与流量使用三倍标准差法。

独立判断法的依据在于：①地点平均车速的合理范围为$0 \leq v \leq f * V$，其中 V 为道路的限制速度，这里限制速度为 60km/h；f 为修正系数，一般取 1.3~1.5,这里取 1.5，故速度最大值取 90km/h。②时间占有率的合理范围为 0~100%，但在实际应用中多车道集成为检测站的时间占有率在 5min 间隔下的合理值最大为 80。③流量的合理范围为$0 \leq q \leq f * C * T/60$，其中 C 为道路通行能力（veh/h），T 为采集周期（min），f 为修正系数（1.3~1.5），一般标准为 3060pcu/h，对于 5min 的时间间隔而言，最大值取 255pcu。经过程序判断后的不合理数据记录为 9999，并生成标记矩阵，运行此法后并未发现异常数据。

num | b1 | threemean

1x12 double

	1	2	3	4	5	6	7	8	9	10	11	12
1	43.8000	50.5000	44.5000	38.4000	40.6000	56.9000	118.7000	113.4000	115.4000	110.3000	109.5000	113.8000

threestd

1x12 double

	1	2	3	4	5	6	7	8	9	10	11	12
1	6.2500	3.6500	7.9100	4.4600	8.0100	1.0600	23.6900	25.3500	25.4900	25.3600	25.1500	26.8900

表格 4 速度与流量数据的均值、标准差

若数据服从正态分布，在 3σ 原则下，异常值如超过 3 倍标准差，那么则将其视为异常值，且数据落入正负 3σ 的概率是 99.7%。由于各路段状况不同，因此这里分路段进行求解。求得速度与流量原始数据的均值与标准差如下表所示。接着先后分别对速度与流量的数据进行 3σ 检验，并完成异常值的标注。经过程序判断的不合理数据记录为 9999，并生成标记矩阵，共有 25 个异常结果，部分结果如表 5 所示。

1	时间段(min)		平均速度(km/h)						流量(Veh)						占有率					
92	445	450	38.5	51.9	44.8	34.3	38	56.9	130	128	161	113	143	154	0.196	0.046	0.062	0.073	0.143	0.049
93	450	455	36	48.8	51.2	9999	52.7	56.8	141	124	130	205	116	117	0.186	0.23	0.038	0.111	0.043	0.038
94	455	460	29.2	47.5	53.3	37.8	47.5	56.6	149	151	126	128	158	149	0.311	0.062	0.033	0.07	0.07	0.048
95	460	465	26.8	53.7	51.2	36.8	48.5	57	140	130	144	134	118	146	0.408	0.048	0.043	0.083	0.053	0.048
96	465	470	34.6	46.5	45.2	36.4	43.9	56.5	134	127	144	137	125	113	0.228	0.054	0.099	0.088	0.058	0.035
97	470	475	39.6	47	43.1	39.1	35.8	9999	168	143	117	113	133	195	0.359	0.059	0.231	0.068	0.092	0.041
98	475	480	39.9	46.7	41.4	38.7	37.4	56.6	144	170	152	130	98	131	0.225	0.29	0.115	0.08	0.082	0.043
99	480	485	59.7	44.7	37	35.8	36.1	56.8	104	100	133	128	136	111	0.061	0.12	0.239	0.1	0.093	0.034
100	485	490	46.2	51.2	36.4	38.5		57.1	113	104	102	87	109	99	0.053	0.038	0.09	0.059	0.118	0.031
101	490	495	42.2	44.8	40.5	36.9	9999	56.7	114	128	120	114	188	141	0.064	0.055	0.121	0.077	0.113	0.045
102	495	500	47.7		39.6	39.2	9999		115	117	139	142	195	121	0.052	0.044	0.145	0.112	0.211	0.039
103	500	505	45	50.9	38.3	36.1	39.1		99	90	111	96	106	121	0.077	0.032	0.07	0.073	0.163	0.039
104	505	510	39.8	45.9	43.9	34	37.7	56.8	134	127	109	119	107	140	0.115	0.18	0.079	0.083	0.217	0.047
105	510	515	38.8	49.9	42	38.3	38.2	56.4	143	157	132	133	151	35	0.162	0.06	0.072	0.103	0.271	0.035
106	515	520	37.1	47.5	42.8	36.7	36.4	56.6	117	114	163	130	116	117	0.147	0.045	0.066	0.088	0.323	0.065
107	520	525	38.2	45.2	49.9	34.9	37.2	56.6	133	116	127	145	141	156	0.11	0.052	0.038	0.092	0.283	0.05
108	525	530	9999	53.2	54	42.8	40	57.1	198	136	105	104	128	127	0.080	0.048	0.027	0.044	0.198	0.042
109	530	535	25.8	50.7	55.8	42.7	48.8	57	125	114	145	113	97	99	0.08	0.041		0.047	0.039	0.031
110	535	540	39.6	51.5	57.8	9999	51.8	9999	120	110	125	205	119	130	0.091	0.04	0.042	0.047	0.043	0.042
111	540	545	44	51.1	42.3	40.8	51.8	9999	114	118	103	91	109	138	0.067	0.043	0.048	0.043	0.041	0.043
112	545	550	42.9	48.2	40.1	42.7	54.1	57.1	108	105	113	112	84	83	0.053	0.043	0.137	0.068	0.03	0.027
113	550	555	45.4	52.5	36.5	32.6	50.6	56.7	92	92	108	103	116	86	0.061	0.032	0.259	0.092	0.043	0.027
114	555	560	48.9	51.4	36.1	9999	44	56.4	96	80	78	70	75	110	0.038	0.032	0.134	0.035	0.035	0.035
115	560	565	44.7	46.6	41.5	34.7	40.8	55.9	108	113	89	93	77	76	0.066	0.046	0.093	0.057	0.046	0.029
116	565	570	43.3	60.8	36.6	36.2	36	56.9	108	95	116	98	112	88	0.093		0.177	0.072	0.085	0.027
117	570	575		51.8	37.6	28	37.9	56.6	93	96	105		110	125	0.05	0.034	0.082	0.05	0.09	0.041
118	575	580		49.2	36.7	32.9	38.2	56.9	91	104	77	116	101	111	0.036	0.039	0.056	0.118	0.128	0.034
119	580	585	50.6	50.4	39.6	36.8	33.2	57	96	89	105	78	101	78	0.043	0.033	0.053	0.049	0.188	0.025
120	585	590	49.1	49.8	9999	42.7	9999	56.5	98	73	100	88	78	111	0.039	0.028	0.036	0.041		0.067
121	590	595	44.7	45.4	9999	34.4	9999	56.5	103	91	86	97	96	91	0.05	0.04	0.022	0.053	0.165	0.029
122	595	600	52.3	48.5	55.9	40.8	31.7	57.4		50	98	78	87	70	0.03	0.04	0.028	0.035	0.211	0.023

表格 5 3σ 检验出的部分异常结果

2. 补全缺失数据。

经过统计，原始数据表格中共有 90 个缺失数据，与检测出的故障数据一起，共有 115 个缺失数据。这里采用基于加权移动平均法对其进行补缺。

加权移动平均法认为各时期的历史数据对缺失数据的影响不相同，离预测期越近的采样时刻数据的影响越大，应在移动平均中赋予更大的权重。其公式表示如下：

$$S_t = \frac{\sum_{i=t-n}^{t-1} W_i X_i}{\sum_{i=1}^{n} W_i}$$

式中：S_t为t时刻丢失样本中的待修补数据；X_i为(t-n)至(t-1)时刻采集的有效数据；W_i是权重值，**一般取n=5**，W_{t-1}=5，W_{t-2}=4，W_{t-3}=3，W_{t-2}=2，W_{t-5}=1。

由于在数据的前 5 组存在 2 条空缺数据，这两个数据采用相邻两检测器的平均值，其余部分对应的 Matlab 代码与部分结果如下所示：

```
%%%%补缺
for i=6:len
    for j=3:20
    if isnan(num(i,j))==1
        num(i,j)=1/15*(5*num(i-1,j)+4*num(i-2,j)+3*num(i-3,j)+2*num(i-4,j)+num(i-5,j));
    end
    if num(i,j)==9999
        num(i,j)=1/15*(5*num(i-1,j)+4*num(i-2,j)+3*num(i-3,j)+2*num(i-4,j)+num(i-5,j));
    end
    end
end
xlswrite('buque',num);
```

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	时间段(min)		平均速度(km/h)						流量(Veh)						占有率					
23	100	105	42.6	45.3	46.2	38.1	43.5	56.8	120	110	155	101	106	145	0.112	0.047	0.053	0.055	0.06	0.045
24	105	110	41.7	49.5	52.9	36.2	40	56.5	124	103	130.8	131	115.86667	113	0.08	0.039	0.027	0.085	0.067	0.035
25	110	115	43.6	52.4	55.9	40.9	33.3	56.3	130	132	102	100	111	103	0.093	0.27	0.026	0.044	0.096	0.032
26	115	120	41.9	51.8	53.3	46.6	35.9	56.6	125	121	141	103	104	139	0.067	0.042	0.036	0.032	0.117	0.044
27	120	125	38.6	49	43.4	40.6	36.2	56.3	135	117	133	124	109	99	0.114	0.046	0.148	0.058	0.09	0.031
28	125	130	39.4	49.9	39.1	39	33.9	56.4	166	137	112	112	133	121	0.191	0.051	0.141	0.064	0.219	0.038
29	130	135	37.3	50.4	41.3	41.013333	35.24	56.6	137	159	151	134	124	137	0.172	0.061	0.16	0.09	0.172	0.045
30	135	140	38.4	51.3	38.1	36	35.033333	56.8	136	121	138	149	139	136	0.1462	0.045	0.240	0.112	0.236	0.043
31	140	145	39.9	50.9	40.7	37.5	36.6	56.8	143	133	134	115	125	130	0.129	0.048	0.183	0.07	0.141	0.041
32	145	150	39.7	49.5	44.8	37.5	39.3	56.6	141	164	161	128.06667	142	146	0.146	0.061	0.092	0.088	0.256	0.046
33	150	155	28.1	50.5	40.4	38.1	33.7	56.4	135	120	139	142	132	145	0.56	0.045	0.089	0.101	0.2076667	0.045
34	155	160	35.2	50.4	41.2	38.3	37.6	56.9	131	125	143	127	129	132	0.23	0.047	0.082	0.086	0.337	0.042
35	160	165	37.8	50.373333	43.9	35	37.3	56.5	159	192	139	145	137	140	0.304	0.054	0.062	0.1	0.285	0.044
36	165	170	37.4	49.6	43.5	37.1	44.4	56.5	136	174	148	149	153	162	0.18	0.065	0.24	0.09	0.17	0.053
37	170	175	50	51.2	46.1	35.7	48.2	56.7	144	122	169	130	114	125	0.109	0.045	0.058	0.08	0.048	0.039
38	175	180	38.4	51.2	49.5	35.7	48.7	56.9	152	121	123	149	132	133	0.167	0.044	0.038	0.089	0.056	0.042
39	180	185	44.2	48.2	51.4	41.7	52.8	57.2	136	161	126	107	130	140	0.106	0.065	0.035	0.049	0.049	0.046
40	185	190	41.9	51	53.1	39.9	48.4	56.4	113	119	147	125	101	112	0.105	0.045	0.038	0.051	0.04	0.036
41	190	195	42.593333	48.9	48.3	43.1	42.1	57.1	112	94	103	105	122	94	0.091	0.037	0.081	0.064	0.063	0.031
42	195	200	41.8	49.78	50.213333	38.5	38.7	57.2	141	130	104	96	101	109	0.48	0.058	0.116	0.054	0.063	0.034
43	200	205	39.9	47.2	50.391111	36.4	36.7	57.1	110	142	140	110	95	111	0.124	0.06	0.168	0.072	0.087	0.036
44	205	210	41.2	52	38	37.4	36.7	57.1	118	102	124	107	128	106	0.119	0.035	0.218	0.085	0.156	0.034
45	210	215	45.9	51	37.3	38.28	36.2	56.6	113	96	110	98	102	111	0.054	0.035	0.29	0.062	0.125	0.034

表格 6 数据补缺部分结果

四、 聚类分析与评价。

在本章，我将对补缺后的路段 3 的线圈数据进行聚类，同时分析得到的各类簇数据的统计特征并评价聚类质量。

1. 聚类。

要求对速度、流量和占有率三个变量进行聚类，由于每个变量只有 168 个数据，且考虑划分为 3-5 类，异常数据已被剔除，故选择 K-means 算法。

首先使用 Calinski-Harabaz 指数确定 K-means 参数。Calinski-Harabaz 通过评估类之间的方差与类内方差来计算得分。将参数分别设置为 2-9，结果如下所示，因此选择 2/7 作为分类簇数。分别对其进行聚类，得到类间间距/类内间距得分：2-2.32；7-3.78，故选择 7 作为最终参数。

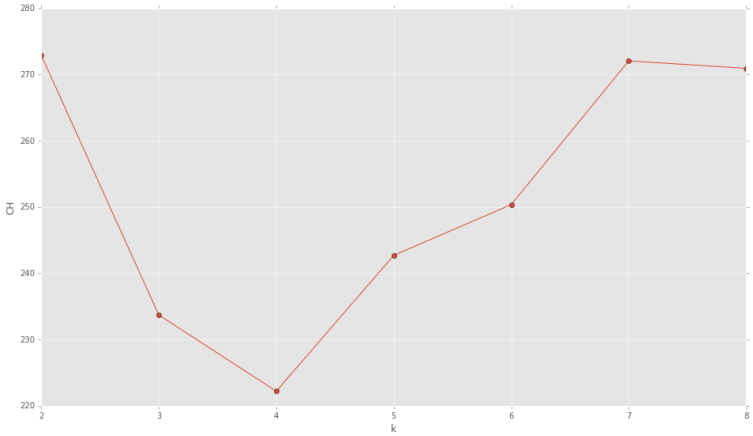


图 6 不同簇数对应 CH 得分

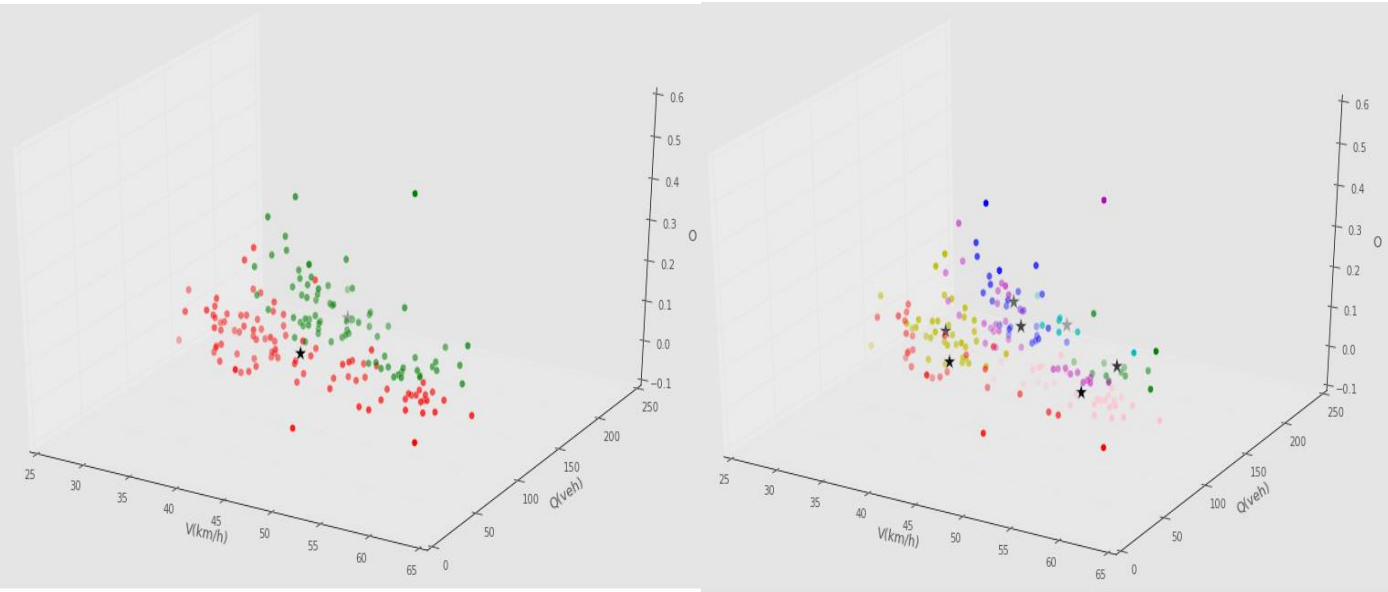


图 7 聚类可视化结果

2. 类簇数据统计特征。

使用数据透视表计算各类簇数据平均速度、流量和占有率三个变量的均值、方差、最大值、最小值、样本数。结果如下表 7 所示。从中可以看出，这 7 个类簇的平均速度基本上 2~3km/h 划分一个类别，流量 10~20veh 划分一个类别，占有率则差距较大，范围从 0.005~0.04。

类簇	0	1	2	3	4	5	6
样本数	26	29	35	13	13	26	26
平均速度 km/h							
类簇	0	1	2	3	4	5	6
最大值	61.30	44.70	43.90	57.60	52.70	59.90	57.80
最小值	43.90	37.00	29.50	43.50	39.15	32.80	36.00
均值	52.34	40.54	37.75	50.22	43.98	40.66	47.85
方差	20.11	3.49	8.34	23.25	10.93	45.23	40.52
流量 veh							
类簇	0	1	2	3	4	5	6
最大值	110.00	145.00	116.00	148.00	197.00	91.00	130.80
最小值	87.00	128.00	93.00	132.00	151.00	32.00	112.00
均值	99.87	136.50	105.06	141.38	161.69	78.92	121.84
方差	38.30	20.14	40.57	28.09	156.06	230.71	26.31
占有率							
类簇	0	1	2	3	4	5	6
最大值	0.116	0.38	0.29	0.24	0.163	0.186	0.48
最小值	0.021	0.052	0.029	0.035	0.049	0.021	0.027
均值	0.042	0.156	0.109	0.080	0.089	0.086	0.093
方差	0.001	0.007	0.003	0.004	0.001	0.003	0.010

表格 7 各类簇统计特征

3. 评价聚类质量

采用轮廓系数评价聚类质量。轮廓系数指向量与簇内部各点距离求均值，衡量簇内部的紧凑程度；再与簇外部所有点的距离求均值，衡量簇外部的分散程度；后者减掉前者，再除以两者最大值。结果在[-1,1]之间，越趋于 1 说明分类质量越好。这里采用 sklearn 包内置的 silhouette_score 函数计算所有点的平均轮廓系数，结果为 0.414，聚类质量尚可。

```
sil_score=silhouette_score(X,estimator.labels_,metric='euclidean')
print (sil_score)
```

```
In [115]: print (sil_score)
0.4143491628647458
```

图 8 轮廓系数评价

五、 相异性与相关性分析。

在本章，我分析路段 3 与路段 6 线圈和浮动车数据估计得到的行程速度与真实行程速度数据的相异性和相关性。

1. 相异性分析

相异性分析包括欧几里得距离与 DTW 距离的计算。

欧几里得距离公式为$\sqrt{(X_{I1}-X_{J1})^2+\cdots+(X_{Ip}-X_{Jp})^2}$，经过计算，路段 3 与路段 6 线圈和浮动车数据估计得到的行程速度与真实行程速度数据的欧几里得距离表如附表中‘oxr3’‘ofr3’‘oxr6’‘ofr6’所示，每个表均为 168*168 大小。最后，分别求得其欧几里得距离均值如下表 8 所示：

路段	路段3		路段6	
数据来源	线圈数据	浮动车数据	线圈数据	浮动车数据
欧几里得距离均值	13.781	15.316	0.464	0.790

表格 8 欧几里得距离均值

从中可以看出：

① 对于路段 3 与路段 6 而言，线圈数据均较浮动车数据更为准确；

② 横向对比，无论是通过线圈还是浮动车估计得到的行程速度，路段 6 都比路段 3 更接近于真实数据。

由于数据是按照时间序列排序的，因此还可以计算 DTW 距离。DTW 算法基于动态规划的思想，通过构建邻接矩阵，寻找最短路径和。DTW 距离公式为$DTW(Q,C)=min\sqrt{\sum_{k=1}^{K}w_k}/K$。经过计算，得到 DTW 距离如下表 9 所示：

```
In [182]: dtw=[]
     ...: dtw.append(dtw_distance(x3,r3,d=lambda x,y: abs(x-y), mww=10000))
     ...: dtw.append(dtw_distance(f3,r3,d=lambda x,y: abs(x-y), mww=10000))
     ...: dtw.append(dtw_distance(x6,r6,d=lambda x,y: abs(x-y), mww=10000))
     ...: dtw.append(dtw_distance(f6,r6,d=lambda x,y: abs(x-y), mww=10000))
     ...: print(dtw)
[1406.8608888888882, 564.2999999999998, 29.767111111111184, 59.70000000000002]
```

路段	路段3		路段6	
数据来源	线圈数据	浮动车数据	线圈数据	浮动车数据
DTW距离	1406.861	564.300	29.767	59.700

表格 9 DTW 距离

观察发现：

① 对于路段 3 而言，浮动车数据比线圈数据的预测更为准确；

② 对于路段 6 而言，线圈数据比浮动车数据的预测更为准确；

③ 横向对比，无论是通过线圈还是浮动车估计得到的行程速度，路段 6 都比路段 3 更接近于真实数据。

2. 相关性分析

我将通过相关系数与协方差评价路段 3/6 中线圈、浮动车预测数据与真实数据的相关性。

协方差公式为$COV(X,Y) = E(X,Y) - E(X) * E(Y)$,这里使用 numpy 自带的 cov 函数计算协方差，结果如下表 10 所示。

路段	路段3		路段6	
数据来源	线圈数据	浮动车数据	线圈数据	浮动车数据
协方差	49.307	70.247	0.045	0.050

表格 10 协方差

协方差越大说明两组数据的同向程度越高，因此：

① 路段 3 与路段 6 均为同相变化，且浮动车数据的同相度>线圈数据；

② 横向比较，路段 3 的同相度>路段 6 的同相度。

接着，计算相关系数。相关系数的公式为：$\rho(X,Y) = \frac{COV(X,Y)}{\sqrt{D(X)*D(Y)}}$,这里使用 numpy 的 corrcoef 函数计算相关系数，结果如下表 11 所示。

观察上表可知：

路段	路段3		路段6	
数据来源	线圈数据	浮动车数据	线圈数据	浮动车数据
相关系数	0.755	0.830	0.623	0.330

表格 11 相关系数

③ 对于路段 3，与实际数据相比浮动车数据比线圈数据的相关性更高；

④ 对于路段 6，与实际数据相比线圈数据比浮动车数据的相关性更高；

⑤ 横向对比，无论是通过线圈还是浮动车数据，路段 3 都比路段 6 与实际数据的相关性更高。

六、 行程速度估计误差。

在本章，我将分别计算线圈、浮动车和 AVI 三种检测方式的行程速度估计误差，结果由 MAPE 与 RMSE 表示。

MAPE 指平均绝对百分误差，公式为：$MAPE = \frac{\sum_{i=1}^{N} \frac{|x_i - \bar{x_i}|}{x_i}}{N}$；RMSE 指均方根

误差，公式为$RMSE = \sqrt{\frac{\sum_{i=1}^{N}|x_i - \bar{x_i}|^2}{N}}$。

线圈行程速度的估计误差按照每个路段分别计算，然后进行平均。浮动车行程速度的估计误差按照每个路段分别计算，然后进行平均。AVI 行程速度估计误差直接按照整条干线平均行程速度计算。

首先计算 RMSE，计算结果如下：

路段	1	2	3	4	5	6	终值
线圈RMSE	12.63132	2.189114	12.31784	5.419747	14.73845	0.298597	5.296691
浮动车RMSE	7.418578	5.575195	6.028864	6.771061	29.88648	25.21685	9.37922
avi	4.995559277						

表格 12 线圈、浮动车与 AVI 数据 RMSE

接着计算 MAPE，计算结果如下：

路段	1	2	3	4	5	6	终值
线圈MAPE	0.40112	0.032468	0.392909	0.138377	0.646046	0.004422	0.210801
浮动车MAPE	0.188561	0.083258	0.13942	0.153929	0.577433	0.842758	0.252802
aviMAPE	0.109838423						

表格 13 线圈、浮动车与 AVI 数据 MAPE

观察表格发现，行程速度的估计误差排序（误差越小代表预测越好）为：

AVI 数据 < 线圈数据 < 浮动车数据

七、 数据融合。

在本章，我将对线圈、浮动车和 AVI 检测数据进行融合，估计每个路段的行程速度，最终目的是使估计结果最接近于给定的真实行程速度(Sheet4)，也就是说 MAPE 越小越好。

数据融合的方案为：AVI 提供较可靠的干道平均速度信息，线圈和浮动车提供每个路段的行程速度信息，同时计算交叉口信号控制延误及其对行程速度的影响。整个数据集 14h，取前 10 小时为训练集，后 4 小时为测试集，检验模型精度。

1. 交叉口信号控制延误

由干道基本示意图可知，共有 5 个交叉口,每个交叉口均为定时控制。由于数据采集的行车方向为从西向东，故这里的延误只计算西进口道车道的延误。该指标是 15min 分析期间的平均每辆车的信号控制延误，用以下公式进行计算（在本次作业中，只需计算d_1、d_2）：

$$d = d_1 + d_2 + d_3$$

$$d_1 = d_s * \frac{t_u}{T} + f_s * d_u * \frac{T - t_u}{T}$$

$$d_2 = 900 * T[(x-1) + \sqrt{\left((x-1)^2 + \frac{8 * e * x}{CAP * T}\right)}]$$

$$d_s = 0.5 * C * (1 - \lambda)$$

$$d_u = 0.5 * C * \frac{(1-\lambda)^2}{1 - \min[1, x]\lambda}$$

$$t_u = \min[T, \frac{Q_b}{CAP[1 - \min[1, x]]}]$$

$$f_s = \frac{1 - P}{1 - \lambda}$$

其中，根据以下公式计算得到各交叉口绿信比(计算来源于交叉口机动车交通量，本次作业不考虑行人和非机动车的流量及其影响)。

$$\lambda = \frac{g}{C} = \frac{G + Y - L}{C}$$

以交叉口 1 西进道口直行车道的信控延误为例，计算过程如下：

交叉口 1 西进道口共有 4 个车道，1 个左转车道、1 个右转车道、2 个直行车道。由于该城市干道限速为 60km/h，根据《城市道路设计规范》中单车道理论通行能力表可知：每个车道的通行能力 CAP=1800pcu/h。

指标		城市道路规范建议值		
V	km/h	40	50	60
基本通行能力 N0	pcu/h	1650	1700	1800

表格 14 单车道理论通行能力

接着计算饱和度 x。x 定义为实际车流量与最大车流量之比，则：

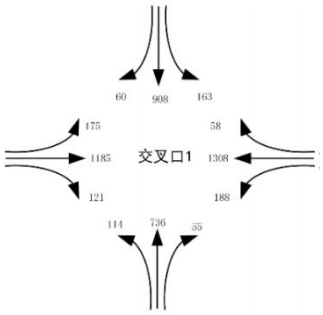


图 8 交叉口 1 基础流量

$$x = \frac{1185}{1800 * 2} = 0.3292$$

由于从西向东方向的车辆只在相位 3：东西直行，中出现，故其绿信比：

$$\lambda = \frac{50 + 3 - 4}{131} = 0.3740$$

交叉口 1 信号延误的其他计算指标还包括：C=131s、T=0.25、e=0.5、Qb=10、P=0.387。

计算得到$d_s = 0.5 * 131 * (1 - 0.3740) = 41.003 s/pcu$

$$d_u = 0.5 * 131 * \frac{(1 - 0.3740)^2}{1 - \min(1, 0.3292) * 0.3740} = 29.272 s/pcu$$

$$t_u = \min\left[0.25, \frac{Q_b}{(1-\min[1,0.3292])*1800}\right] = \min\left[0.25, \frac{10}{1207.44}\right] = 0.0083$$

$$f_s = \frac{1-P}{1-\lambda} = \frac{1-0.387}{1-0.3740} = 0.9792$$

则计算得到 d1、d2 与 d 如下：

$$d_1 = 41.003 * \frac{0.0083}{0.25} + 0.9792 * 29.272 * \frac{0.25-0.0083}{0.25} = 29.07s/pcu$$

$$d_2 = 900 * 0.25 * \left[(0.3292-1) + \sqrt{\left((0.3292-1)^2 + \frac{8*0.5*0.3292}{1800*0.25}\right)}\right]$$

$$= 0.49s/pcu$$

$$d = d_1 + d_2 = 29.56s/pcu$$

同理，得到所有交叉口西进口道直行车道与左转车道的延误如下所示：

交叉口	1	2	3	4	5
直行车流量(pcu/h	1185	1214	1368	1315	1361
直行车道数	2	2	2	2	3
x	0.3292	0.3372	0.3800	0.3653	0.2520
C(s)	131	131	127	127	130
相位3绿灯时间(s)	50	56	50	49	45
绿信比	0.374	0.420	0.386	0.378	0.338
T(h)	0.25				
e	0.5				
QB	10				
P	0.387				
ds(s/pcu)	41.003	37.990	38.989	39.497	43.030
du(s/pcu)	29.272	25.670	28.054	28.503	31.138
tu(h)	0.0083	0.0084	0.0090	0.0088	0.0074
fs	0.97923	1.0569	0.99837	0.98553	0.92598
d1(s/pcu)	29.07	27.49	28.40	28.49	29.26
d2(s/pcu)	0.49	0.51	0.61	0.57	0.34
d(s/pcu)	29.56	28.00	29.01	29.06	29.59

交叉口	1	2	3	4	5
左转车流量(pcu/h	175	139	292	214	147
左转车道数	1	1	2	2	1
x	0.0972	0.0772	0.0811	0.0594	0.0817
C(s)	131	131	127	127	130
相位4绿灯时间(s)	14	12	12	12	19
绿信比	0.0992	0.0840	0.0866	0.0866	0.1385
T(h)	0.25				
e	0.5				
QB	5				
P	0.107				
ds(s/pcu)	59.000	60.000	58.000	58.000	56.000
du(s/pcu)	58.0905	59.0780	57.0802	57.0819	55.1286
tu(h)	0.0030769	0.00301	0.00302	0.00295	0.00302
fs	0.9913814	0.97486	0.97768	0.97768	1.03652
d1(s/pcu)	57.61	57.62	55.83	55.83	57.13
d2(s/pcu)	0.49	0.51	0.61	0.57	0.34
d(s/pcu)	58.10	58.13	56.44	56.41	57.46

图 6 各交叉口西进口道直行车道与左转车道的信控延误

由于所有交叉口的右转车辆都不受信号灯的控制，则其信控延误为 0，取各方向车道交通量与延误之积除于总交通量，得到各交叉口西进道口的平均控制延误。

交叉口	1	2	3	4	5
直行车辆流量	1185	1214	1368	1315	1361
直行车辆延误	29.56	28.00	29.01	29.06	29.59
左转车辆流量	175	139	292	215	147
左转车辆延误	58.10	58.13	56.44	56.41	57.46
右转车辆流量	121	50	103	76	57
右转车辆延误	0	0	0	0	0
西进道口总流量	1481	1403	1763	1606	1565
西进道口总延误	45195.60	42074.61	56172.44	50347.25	48721.49
西进道口平均延误	30.52	29.99	31.86	31.35	31.13
流量单位：pcu/h；延误单位：s/pcu					

图 7 各交叉口西进道口平均信控延误

2. 信控延误的影响

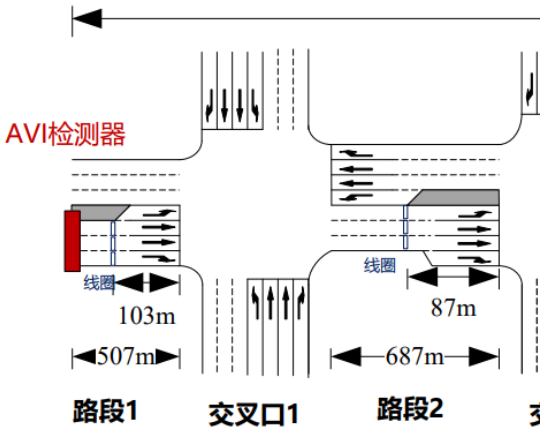


图 8 路段一

因为使用环形线圈测速时，线圈埋设地点距离交叉口还有一定的距离，而对于交叉口前的路段而言，其检测数据估计得到的路段平均行程速度较真实值往往偏大，因此，这里应结合各交叉口的信控延误对其进行优化处理。

由于延误并不全部作用在交叉口前的路段，因此这里对其进行处理时，只采用部分信控延误；采用比例通过对比获得（这里采用路段一的数据）。

$$新线圈速度=\frac{该路段全长}{\frac{对应交叉口信控延误}{某个大于1的值}+\frac{线圈前路段长度}{原线圈速度}}$$

在这里，比例的选择范围是 1-4 之间的数，精度为 0.1。选择路段一真实行程速度数据与新线圈速度之差的标准差作为评价指标，得到比例-标准差曲线如下：

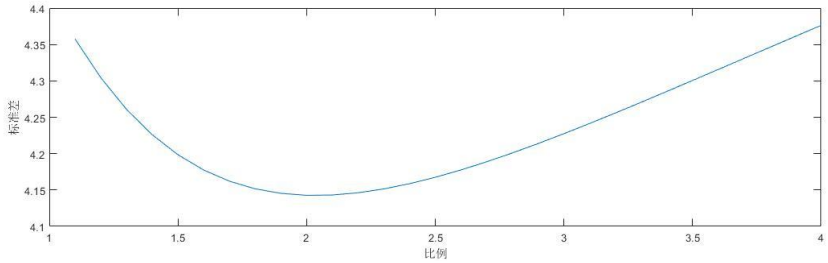


图 9 比例-标准差曲线

当比例=2 时，新线圈速度最接近真实行程速度。处理后路段一示例如下：

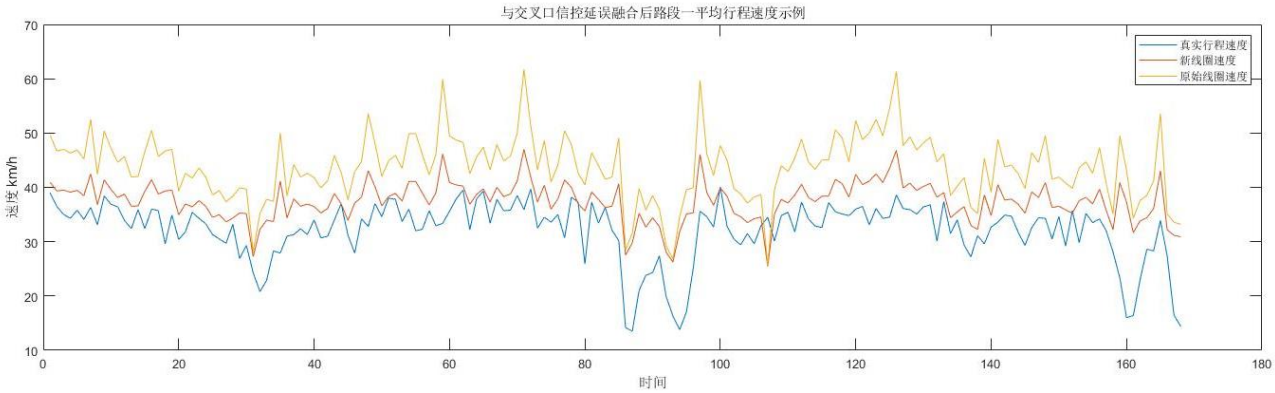


图 10 路段一处理结果

将处理后的数据输出到 excel 表格中备用。

3. 数据融合

选择神经网络 BP 反向传播算法对每个路段的平均行程速度进行预测。对于每个路段而言，输入层有三个，分别为：新线圈估计得到的该路段行程速度、浮动车估计得到的该路段行程速度与 avi 检测器计算的干道行程速度。输出层有一个，即数据融合后得到的该路段平均行程速度。

首先，将数据集分为训练集与测试集。根据题目要求，前 10 小时，即 1~120 的数据为训练集，后 4 小时，即 121~168 的数据为测试集。

接着，为了使网络快速的收敛，对样本数据进行归一化处理。常用的归一化操作有 Z-score 与 Max-min。由于 Z-score 方法是基于数据的均值和方差来将数据标准化，适用于数据的分布类似高斯分布的情况；Max-min 是对数据进行一次线性变换，将数据映射到[0,1]，适用于数据较为零散或者是线性关系，并且没有很多离群值的时候。因此这里采用 Mapminmax 函数归一化。

```
%样本数据归一化 min=0 max=1 [0,1]
[inputn,inputps]=mapminmax(input_train1,0,1);
[outputn,outputps]=mapminmax(output_train1,0,1);
```

图 11 归一化处理

归一化后创建 BP 网络，这里需要确定隐含层神经元的个数。一般而言，增加隐含层数可以降低网络误差，提高精度，但也会使网络复杂化，增加了网络的训练时间和出现“过拟合”的倾向。隐含层神经元的个数满足公式：

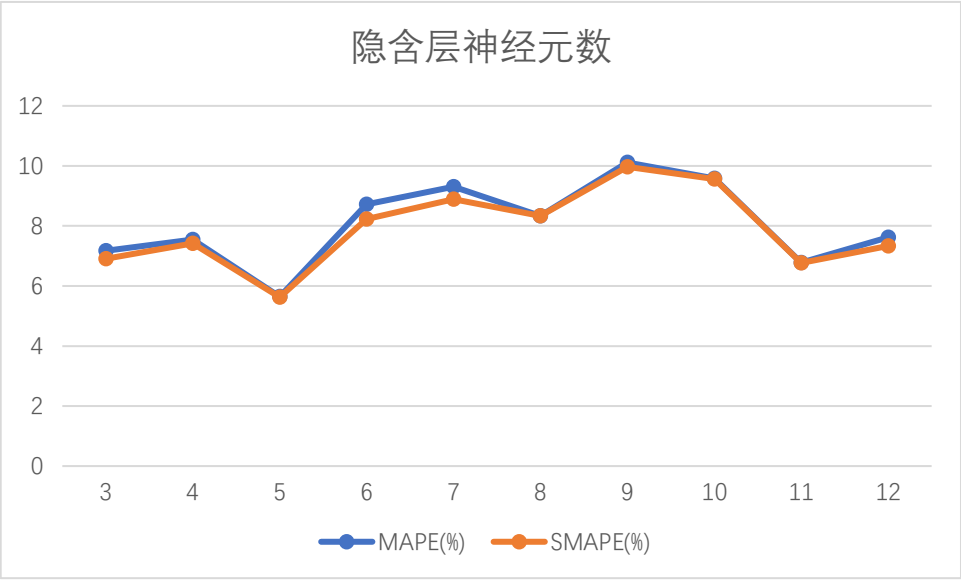


图 12 隐含层神经元数与 MAPE

$$n_1 = \sqrt{(n+m)} + a$$

其中，n 为输入单元数，m 为输出单元数，n1 位隐含层单元数，a 为[1, 10]之间。得到最佳隐藏层数的边界[$\sqrt{(n+m)}+1$，$\sqrt{(n+m)}+10$]，即[3, 12]。在收敛后比较收敛速度，根据得到的训练误差与收敛速度综合选择最佳隐藏层数。在本次实验中，由于收敛速度都很快，因此这里采用 MAPE 作为主要的评价指标。结果如图 12 所示，观察可知，当隐含层神经元数为 5 时，达到最好效果。

然后，设置 BP 网络的训练参数，包括迭代次数 epochs、学习率 lr 与目标值 goal。其中，学习率越大，输出误差对参数的影响就越大，参数更新的就越快，同时受到异常数据的影响就越大，很容易发散。经过测试后，参数分别设置为：$epochs = 1000, goal = 0.001, lr = 0.1$。网络训练后，将预测数据归一化，进行预测；将预测结果输出，注意要反归一化。

最终结果如下所示：

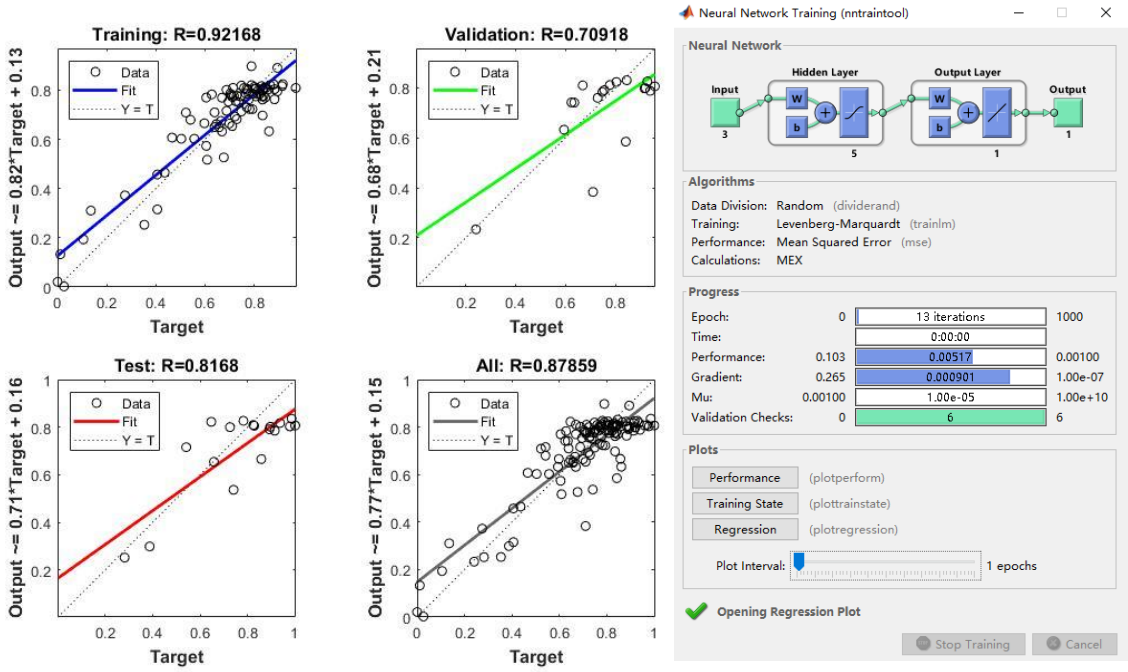


图 13 BP 模型图

同理，分别预测路段一至路段六的平均行程速度（km/h），绘制真实值-预测值曲线，并计算误差。题目要求以 MAPE（平均绝对百分误差）为测量值，然而，MAPE 是不对称的，它对负误差（预测值>真实值）比正误差（预测值<

真实值）的惩罚更大。因此，这里同时给出 SMAPE（对称平均绝对百分比误差）克服上述的不对称性。

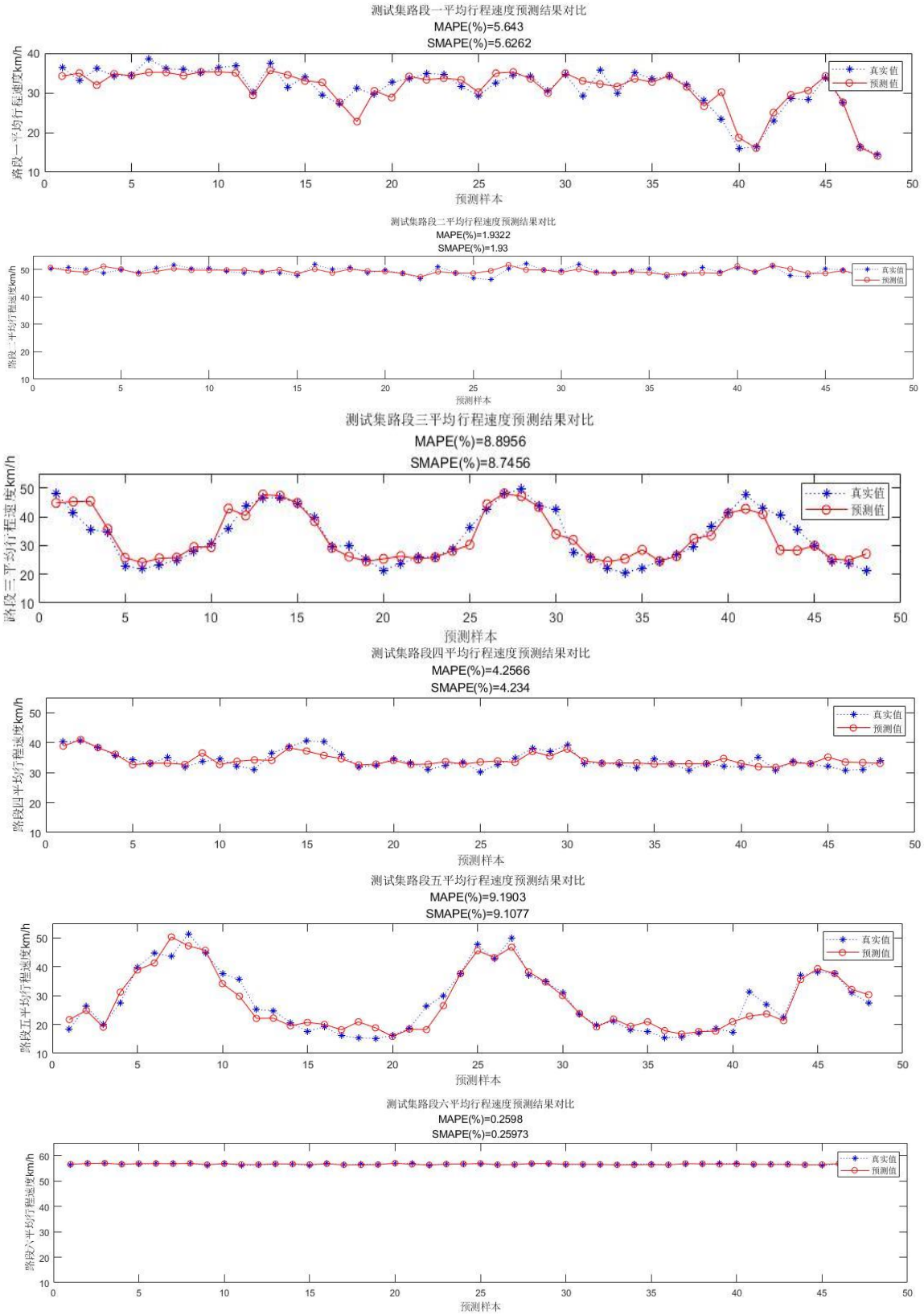


图 14 路段 1-6 的 BP 预测平均行程速度与真实值对比

表格 15 路段平均 MAPE 与 SMAPE

路段	1	2	3	4	5	6	平均
MAPE(%)	5.6430	1.9322	8.8956	4.2566	9.1903	0.2598	5.0296
SMAPE(%)	5.6262	1.9300	8.7456	4.2340	9.1077	0.2597	4.9839

计算得出路段一到路段六的平均 MAPE 与平均 SMAPE 分别为 5.0296%与 4.9839%。同时，观察表格与图可以发现，路段一、二、四、六的机动车平均行程速度较为平稳，该 BP 模型的拟合程度较好，MAPE 误差均在 6%以内；而对于路段三、五，机动车平均行程速度随时间变化较大，MAPE 误差大于 6%，但仍然小于 10%，模型效果较为理想。