# 原始数据探查

## 1.1数据总体情况

共计提供1w辆新能源车辆数据，数据时间范围为20200101到20201116，总计数据文件大小为870G。车辆从3家车企抽取的数据，分别对应的车辆为[1-5000, 5001-7000,7001-10000]。

提供共计39个字段，具体字段见下表

表一

|  |  |  |
| --- | --- | --- |
| 字段 | 英文 | 字段含义 |
| vin | vin | 车辆唯一识别码 |
| time | time | 格式：20200914233201 |
| 2201 | spd | 车速,0.1km/h |
| 2615 | soc | 剩余电量，1% |
| 2202 | mileage | 里程，0.1km |
| 2502 | lon | 经度，10^-6 |
| 2503 | lat | 纬度, 10^-6 |
| 2910 | w2910 | 动力蓄电池包不匹配报警 |
| 2911 | w2911 | 动力蓄电池一致性差报警 |
| 2203 | gear | 挡位 |
| 2501 | locationStatus | 定位状态 |
| 2901 | w2901 | 温度差异报警 |
| 2902 | w2902 | 电池极柱高温报警 |
| 2903 | w2903 | 动力蓄电池包过压报警 |
| 2904 | w2904 | 动力蓄电池包欠压报警 |
| 2905 | w2905 | SOC低报警 |
| 2906 | w2906 | 单体蓄电池过压报警 |
| 2907 | w2907 | 单体蓄电池欠压报警 |
| 2808 | n2808 | 正极对地电阻 |
| 2804 | c2804 | 电机故障总数N |
| 2909 | w2909 | SOC过高报警 |
| 2301 | chargeStatus | 充放电状态 |
| 2204 | brakeStatus | 制动踏板状态 |
| 2205 | powerStatus | 驱动电机状态 |
| 3201 | carStatus | 车辆状态 |
| 2213 | status | 运行模式 |
| 2214 | dcStatus | DC-DC状态 |
| 2307 | c2307 | 驱动电机个数 |
| 2913 | w2913 | DC-DC温度报警 |
| 2914 | w2914 | 制动系统报警 |
| 2915 | w2915 | DC-DC状态报警 |
| 2916 | w2916 | 驱动电机控制器温度报警 |
| 2917 | w2917 | 高压互锁状态报警 |
| 2918 | w2918 | 驱动电机温度报警 |
| 2919 | w2919 | 车载储能装置类型过充 |
| 2930 | w2930 | SOC跳变报警 |
| 2920 | n2920 | 最高报警等级 |
| 2921 | c2921 | 可充电储能装置故障总数N1 |
| 2923 | c2923 | 发动机故障总数N3 |

数据分为卫星定位类数据，报警类数据，车辆状态类数据。

## 1.2卫星定位数据

|  |  |  |
| --- | --- | --- |
| 字段 | 是否提供 | 影响 |
| 卫星定位时间 | 否 | 卫星定位时间与设备终端时间都可能存在异常，可以互相校验，未提供卫星定位时间只能依赖设备终端时间的可靠性 |
| 设备终端时间 | 是 |  |
| 经度 | 是 |  |
| 纬度 | 是 |  |
| 卫星定位方向 | 否 | 急转弯判断依据，道路匹配需要字段 |
| 卫星定位精度 | 否 | 表示卫星定位数据的水平精度，可用于判断数据的准确度 |
| 可见卫星数 | 否 | 用于判断卫星定位数据的准确度 |
| 速度 | 否 | 可用车辆数据的速度代替 |
| 卫星定位高度 | 否 | 无影响 |

实际系统中，应存在三个时间：设备终端时间、卫星定位时间和服务器接收时间。服务器接收时间通常是准确的。设备终端时间与卫星定位时间都可能存在异常。卫星定位时间异常通常发生在冷启动、山地高楼遮挡导致的可见卫星数少等情况，通过精度、卫星数等数据可以过滤掉异常值。设备终端时间可能由于用户手工操作或网络授时延迟存在异常值，可由卫星定位时间校验保证时间的准确性，无卫星定位时间可能导致：实时数据的顺序混乱；数据欺诈等道德风险。

卫星定位精度与可见卫星数直接影响着数据的可靠性，缺失该数据导致数据过滤和修正需要的事实基础缺失，除卫星定位经纬度异常外的数据异常难以检测，对里程结果的准确性有较大影响。本次工作只能假设卫星定位数据的精度处于可用范围内。

尽管数据中提供了定位状态字段（2501），但经过检查该数据取值包含{0, 4}。根据字段含义解析均表示“有效定位”，无法对定位数据的有效性提供帮助。而取值4的数据表示经纬度的经度为西经，而中国范围内均为东经，经过检查取值为4的数据的经纬度均为0，确认为异常数据。

pdata['locationStatus'].value\_counts()

0.0 1027083

4.0 32892

#按照解析4代表 有效定位;北纬;西经，可能有异常点

#状态为4时经纬度均为0

location = data[data['locationStatus']==4]

location.describe()

## 1.3报警类数据

报警类数据包含各单个数据报警和最高报警等级字段（2920），由于报警发生频率极低，大部分情况下数据全为零值，且单个数据报警发生时均会上报最高报警等级，因此只保留最高报警等级字段，单个报警不做具体分析。

#最高报警等级

data['n2920'].value\_counts()

0.0 998397

2.0 5580

3.0 107

1.0 15

sortdata['warn1'] = sortdata['n2920'].shift(1)

sortdata['warndiff'] = sortdata['n2920']-sortdata['warn1']

sortdata['warndiff'].value\_counts()

#共计发生13次报警

0.0 1003770

-2.0 10

2.0 10

1.0 2

-1.0 2

-3.0 1

3.0 1

#最高报警与单个报警关系

sortdata['warnsum'] = sortdata['w2901']+sortdata['w2902']+sortdata['w2903']+sortdata['w2904']+sortdata['w2905']+sortdata['w2906']+sortdata['w2907']+sortdata['w2909']+sortdata['w2913']+sortdata['w2914']+sortdata['w2915']+sortdata['w2916']+sortdata['w2917']+sortdata['w2918']+sortdata['w2919']+sortdata['w2930']

sortdata['warnsum'].value\_counts()

0.0 1003909

1.0 190

warn = sortdata[['n2920','warnsum']]

warn.value\_counts()

#最高等级报警覆盖所有单个报警的结果

n2920 warnsum

0.0 0.0 998397

2.0 0.0 5497

3.0 1.0 107

2.0 1.0 83

1.0 0.0 15

## 1.4车辆状态类数据

|  |  |  |  |
| --- | --- | --- | --- |
| 字段编码 | 字段 | 值域 | 是否保留 |
| 2205 | 驱动电机状态 | 0和1,0值不应存在 | 否 |
| 2307 | 驱动电机个数 | 均为1 | 否 |
| 2213 | 运行模式 | 均为1 | 否 |
| 2808 | 正极对地电阻 | 均为0 | 否 |
| 2804 | 电机故障总数 | 均为0 | 否 |
| 2921 | 可充电储能装置故障总数 | 均为0 | 否 |
| 2923 | 发动机故障总数 | 均为0 | 否 |
| 2204 | 制动踏板状态 | 取值0和1，按标准应为0-100 | 否 |
| 2214 | DC-DC状态 | 取值1和2，工作与断开，车辆状态覆盖 | 否 |
| 2301 | 充放电状态 | 取值{1,3,4}代表{停车充电,未充电,充电完成},车辆状态覆盖 | 否 |
| 2203 | 档位 | {0,13,14}代表{空档,倒挡,d档} | 是 |

pdata['gear'].value\_counts()

14.0 812909

0.0 237864

13.0 9202

#13代表 有驱动力：档位编码“1101”， 倒挡

#14代表 有驱动力：档位编码“1110”， d档

pdata['chargeStatus'].value\_counts()

3.0 970576

1.0 83004

4.0 6115

#1 停车充电

#3 未充电状态

#4 充电完成

pdata['brakeStatus'].value\_counts()

1.0 533784

0.0 526191

#制动踏板状态

#0 代表制动关的状态，需要解释，待定

pdata['powerStatus'].value\_counts()

0.0 596438

1.0 463537

#0 未知，标准中不存在

#1 耗电

#驱动电机状态

pdata['carStatus'].value\_counts()

1.0 913583

2.0 146112

#1 车辆启动状态

#2 熄火

pdata['status'].value\_counts()

1.0 1059695

#1 纯电

pdata['dcStatus'].value\_counts()

1.0 1039441

2.0 20534

#1 工作

#2 断开

sta = sortdata[['chargeStatus','carStatus']]

sta.value\_counts()

chargeStatus carStatus

3.0 1.0 880359

1.0 2.0 85449

3.0 2.0 28236

4.0 2.0 10055

#熄火状态不会处于充电，熄火状态准确

sta = sortdata[['chargeStatus','carStatus','dcStatus']]

sta.value\_counts()

#车辆状态的熄火状态准确,dcStatus字段不使用

chargeStatus carStatus dcStatus

3.0 1.0 1.0 880359

1.0 2.0 1.0 85449

3.0 2.0 1.0 28177

4.0 2.0 1.0 10049

3.0 2.0 2.0 59

4.0 2.0 2.0 6

gear = sortdata[['gear','carStatus']]

gear.value\_counts()

#熄火状态时档位都为0，空挡

gear carStatus

14.0 1.0 707968

0.0 1.0 165021

0.0 2.0 123740

13.0 1.0 7370

## 1.5 保留字段

|  |  |  |
| --- | --- | --- |
| 字段 | 含义 | 说明 |
| vin | vin |  |
| time | Time |  |
| 2201 | 车速 | 0.1km/h |
| 2615 | Soc | 1% |
| 2202 | 里程 | 0.1km |
| 2502 | 经度 | 0.000001度 |
| 2503 | 纬度 | 0.000001度 |
| 2203 | 档位 | {0,13,14};{P,R,D} |
| 2501 | 定位状态 | {0,4};{正常,异常} |
| 3201 | 车辆状态 | {1,2};{启动,熄火} |
| 2920 | 最高报警等级 | {0,1,2,3}报警等级 |

其中里程字段存在异常，使用前需进行过滤

mileage = sortdata[['time','mileage']]

mileage['mileage1'] = mileage['mileage'].shift(1)

mileage['diff']=mileage['mileage']-mileage['mileage1']

mileage['diff'].value\_counts()

#mileagediff文件

mileage['diff'].value\_counts()

mileagediff,count

0.0,692125

1.0,147730

2.0,136873

3.0,22879

4.0,1715

5.0,747

7.0,163

8.0,145

10.0,143

6.0,132

9.0,87

12.0,71

15.0,58

11.0,43

13.0,42

20.0,42

17.0,35

14.0,30

25.0,29

18.0,29

-1.0,26

19.0,24

30.0,22

16.0,22

22.0,20

21.0,19

28.0,18

23.0,17

24.0,14

35.0,14

-5.0,14

38.0,13

-4.0,13

40.0,12

-2.0,12

-3.0,12

-7.0,12

-6.0,12

29.0,10

37.0,10

27.0,10

48.0,9

-9.0,9

-11.0,8

32.0,8

39.0,8

34.0,8

-10.0,7

45.0,7

43.0,7

50.0,7

36.0,7

60.0,6

33.0,6

68.0,6

65.0,6

26.0,6

-13.0,5

31.0,5

-8.0,5

55.0,5

47.0,5

54.0,5

42.0,5

49.0,5

44.0,4

97.0,4

62.0,4

70.0,4

-12.0,3

41.0,3

120.0,3

-16.0,3

83.0,3

-15.0,3

72.0,3

105.0,2

73.0,2

69.0,2

100.0,2

98.0,2

66.0,2

80.0,2

-19.0,2

77.0,2

-14.0,2

103.0,2

57.0,2

90.0,2

78.0,2

-18.0,2

87.0,2

56.0,2

53.0,2

64.0,2

110.0,2

92.0,2

84.0,2

52.0,2

3125.0,1

175.0,1

2012.0,1

177.0,1

2845.0,1

112.0,1

-25.0,1

2065.0,1

182.0,1

180.0,1

122.0,1

46.0,1

185.0,1

-42.0,1

2784.0,1

133.0,1

2067.0,1

150.0,1

-20.0,1

2190.0,1

2192.0,1

2196.0,1

138.0,1

2209.0,1

2070.0,1

141.0,1

125.0,1

2329.0,1

147.0,1

2360.0,1

96.0,1

76.0,1

610.0,1

618.0,1

622.0,1

623.0,1

1927.0,1

-21.0,1

1956.0,1

2594.0,1

658.0,1

167.0,1

136.0,1

684.0,1

-37.0,1

1709.0,1

202.0,1

332.0,1

1396.0,1

1748.0,1

346.0,1

344.0,1

85.0,1

1759.0,1

81.0,1

3244.0,1

444.0,1

319.0,1

79.0,1

-38.0,1

1233.0,1

1223.0,1

1747.0,1

1418.0,1

348.0,1

363.0,1

365.0,1

1723.0,1

1490.0,1

95.0,1

1719.0,1

1560.0,1

1588.0,1

102.0,1

6579.0,1

1685.0,1

106.0,1

107.0,1

1701.0,1

1200.0,1

1179.0,1

113.0,1

115.0,1

13245.0,1

1924.0,1

839.0,1

857.0,1

128.0,1

225.0,1

1916.0,1

924.0,1

58.0,1

427.0,1

929.0,1

931.0,1

59.0,1

220.0,1

108.0,1

1122.0,1

244.0,1

75.0,1

63.0,1

114.0,1

1032.0,1

1050.0,1

160.0,1

2112.0,1

67.0,1

2014.0,1

-47.0,1

274.0,1

111.0,1

279.0,1

1075.0,1