

一、数据前处理（以 LPHC7PE0K1A07972 为例）

1、空值及异常数据帧删除

根据文件《数字汽车大赛数据字段介绍》中的“sing_volt_num”和“sing_temp_num”字段的众数判断电池组的电池单体数量为 96，温度探针数量为 48，将“cell_volt_list”和“cell_temp_list”中的数据分割为单独命名的列，便于后续取值。

数据处理中不可避免的会出现空值，若通过均值填补或众数填补的方式处理，可能破坏数据原有的特点，因此选择直接删除有空值的数据帧。

由于后续需要划分充电桩充电片段和驾驶过程片段，充电状态是片段划分的重要依据，因此需要处理充电状态的记录异常数据帧。根据文件《数字汽车大赛数据字段介绍》中关于充电状态的描述，当字段值为“0xFE”为异常数据，“0xFF”时为无效数据，删除这两类数据帧。

总电流、单体电压以及温度为构建模型时的重要输入特征，因此需要处理充电状态的记录异常数据帧，同样删除表示异常和无效数据的“0xFE”和“0xFF”数据帧。

查看处理完成后的总电压、总电流的分布（图 1.1），符合常理认知，可以进行下一步数据清洗。

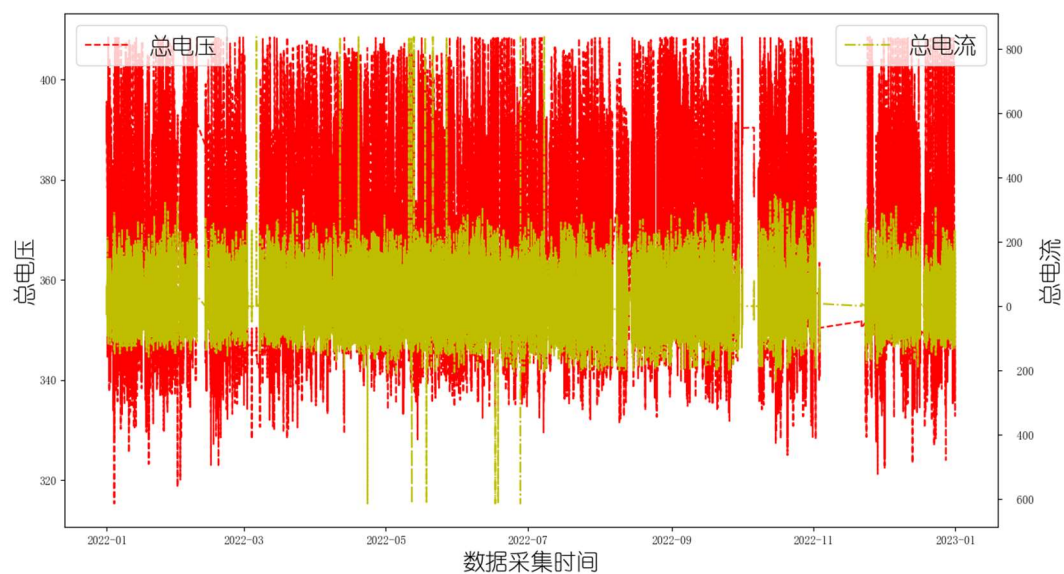


图 1.1 全过程总电压、总电流变化

2、里程异常数据帧删除

查看累计里程随数据采集时间的变化曲线（图 1.2），可以看到有部分数据帧记录的累计里程存在异常跳变，显然这是记录错误的数据帧，通过设置累计里程的范围进行筛选，清理后的累计里程变化如图 1.3 所示。

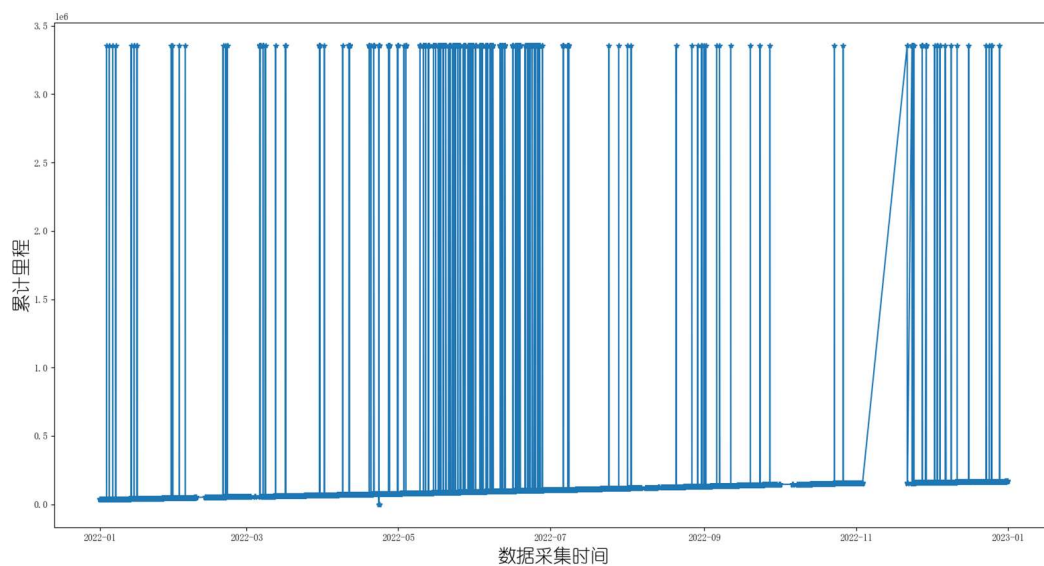


图 1.2 清洗前累计里程变化

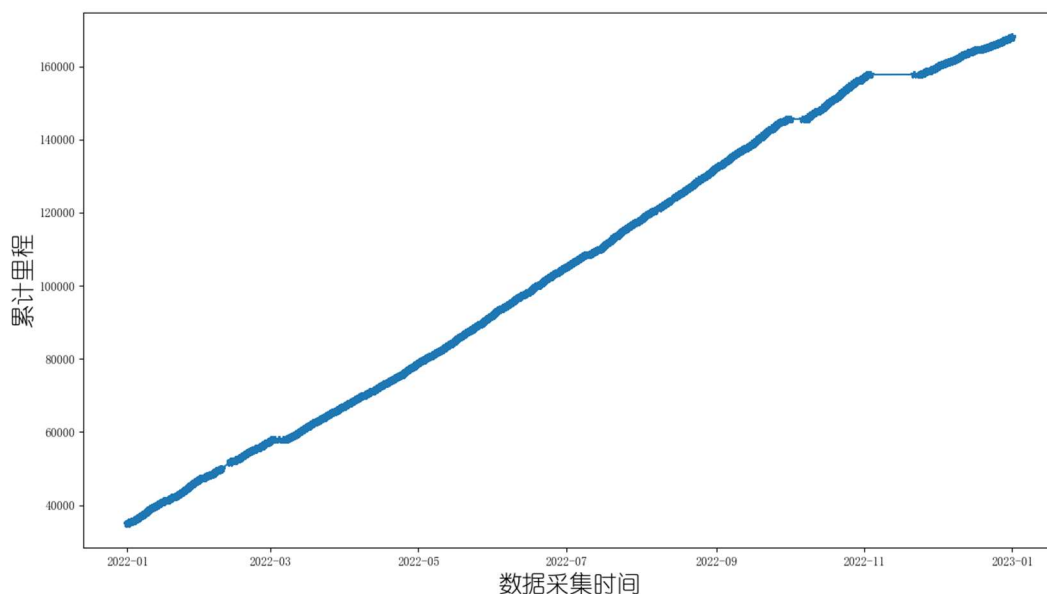


图 1.3 清洗后累计里程变化

在全部十三辆车的数据集中，vin 码为 LFPHC7PE0K1B16707 的车辆里程错乱数据过多（图 1.4），删除后影响整体数据集质量，删除该数据。vin 码为 LFPHC7PE5K1A24931 的车辆数据过少，深度充电片段过少（图 1.5），数据质量较低，删除该数据。vin 码为 LFPHC7PEXK1B02846 的车辆 SOC 算法发生过改变，导致容量显著改变（图 1.6），无法统一利用，因此删除该数据。

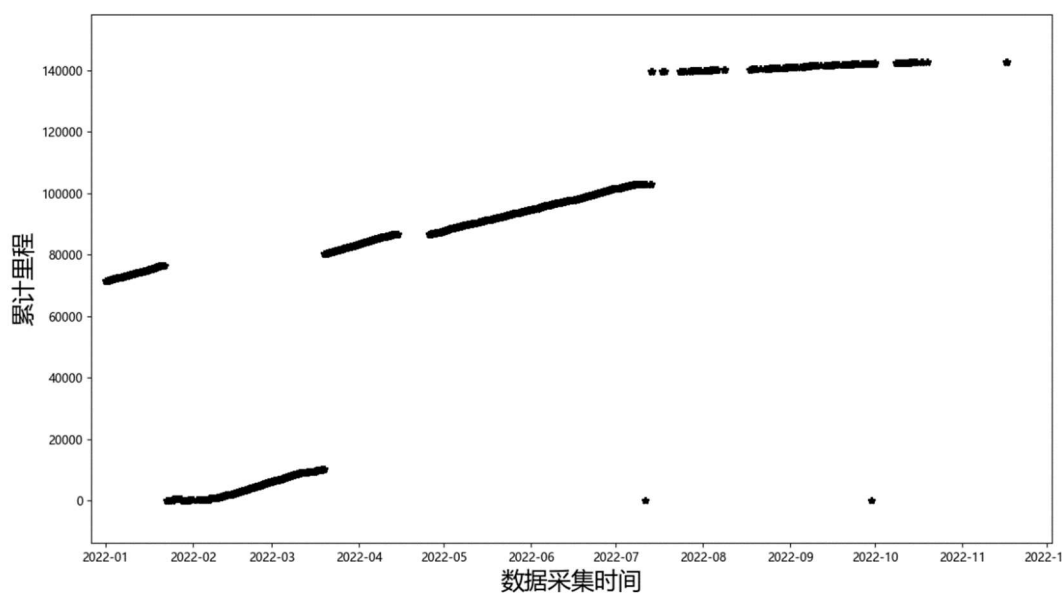


图 1.4 LFPHC7PE0K1B16707 车辆里程随时间的变化

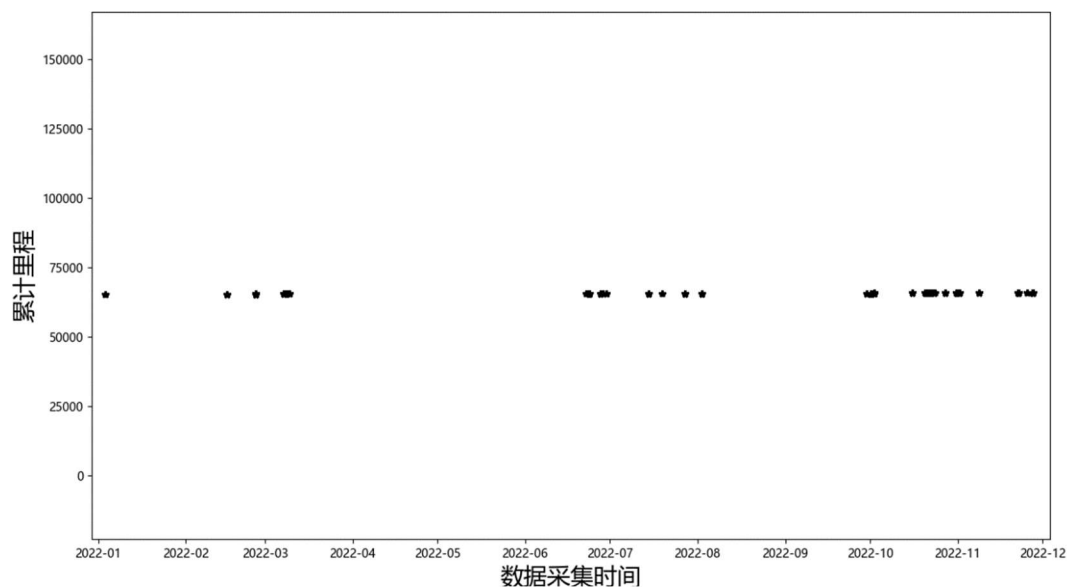


图 1.5 LFPHC7PE5K1A24931 车辆数据过少

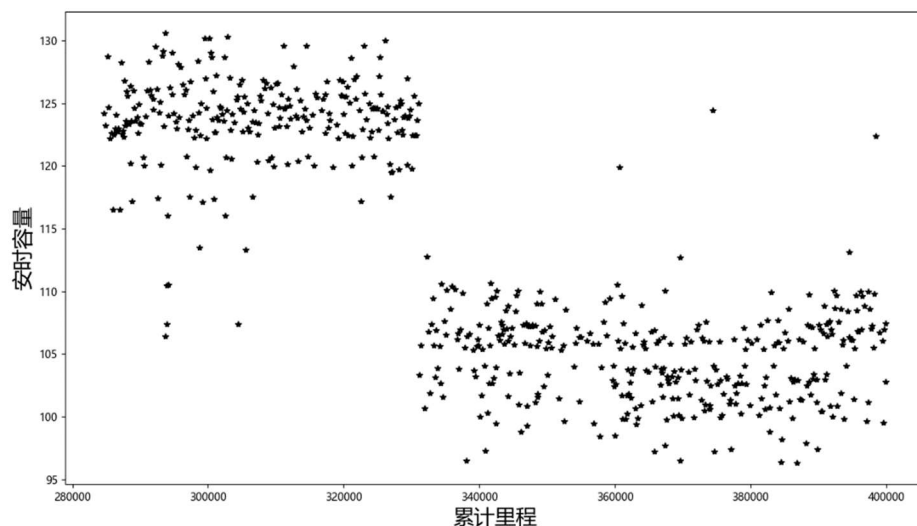


图 1.6 LFPHC7PEXK1B02846 车辆容量算法发生切换

3、数据片段划分

在构建电池预测模型之前，首先需要获取电池的容量值作为模型训练过程中的标签，这里选择停车充电片段计算电池组实际容量。

理由如下：

1、电池组容量计算方法无法避免的需要安时积分计算充入容量或放出容量，即：

$$C^* = \int_{t_1}^{t_2} I dt$$

使用这种方法的前提假设为 d_t 时间内总电流保持不变，但在驾驶过程电流变化迅速且波动剧烈，在 10s 采样间隔内电流可能经历多次变化，导致与假设的理想情况偏差较大，因此若通过驾驶数据片段按时积分计算电池组容量会有较大误差；

2、通过清洗后的电流数据可以看出本车的充电方式为多段恒流充电（图 1.7），在每个恒流阶段，电流接近恒定，接近假设情况，且按照常用的 SOC 估计算法，当电流稳定时 SOC 估计精度更高；

3、不同于 SOC 估计或热失控预测等需要实时更新状态的算法，SOH 的估计通常以循环为单位，由于锂离子电池的高库伦效率特点，每次循环中充电过程计算得到的电池容量可以视为等于放电过程计算得到的容量。

综上所述，选择停车充电片段的数据计算电池容量标签能够极大地降低准确预测的难度。因此，需要从原数据中分离出停车充电数据和驾驶过程数据。

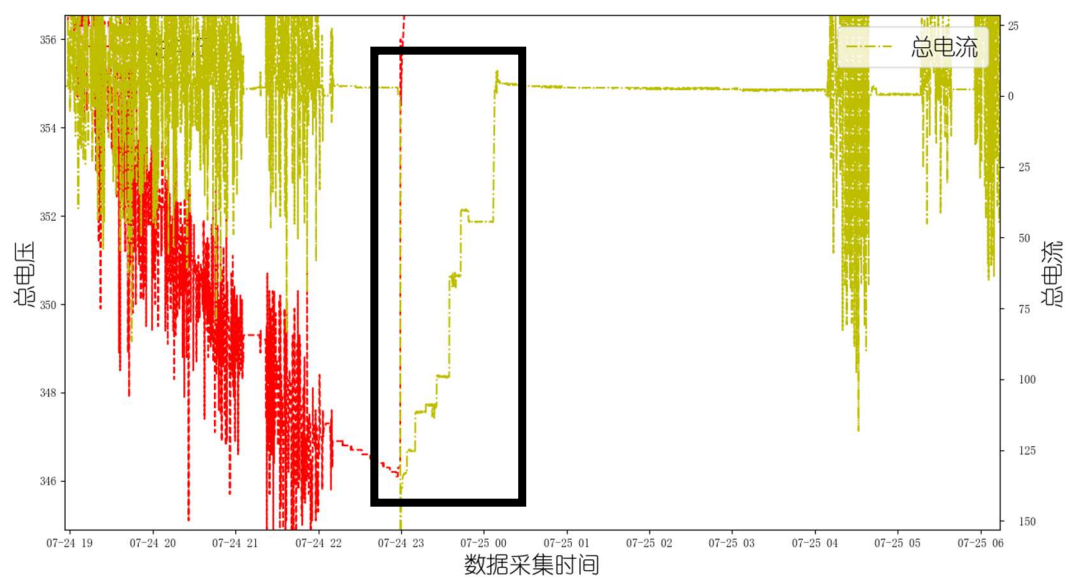
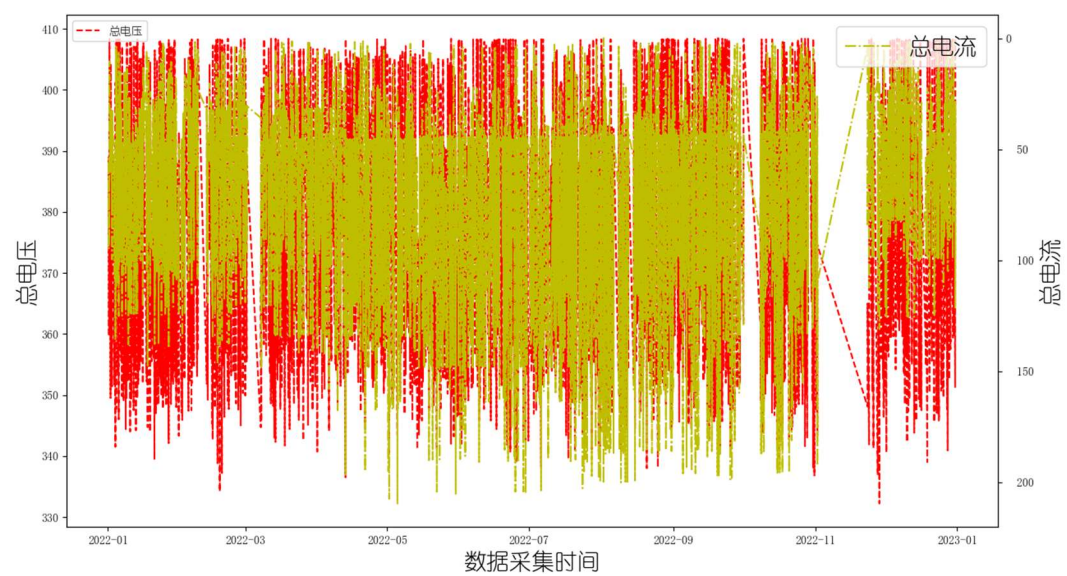
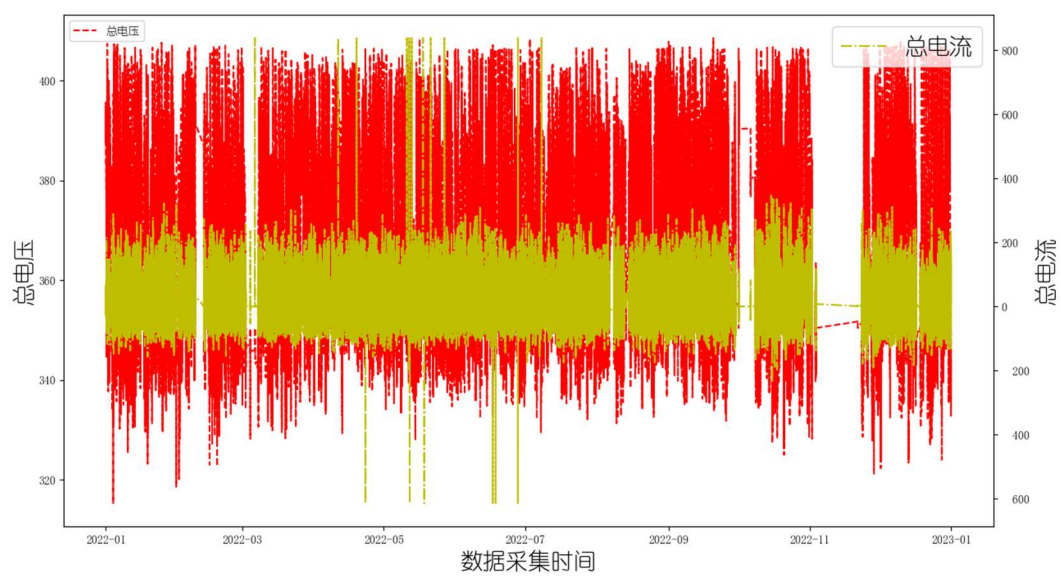


图 1.7 充电方式说明

两类数据片段划分完成后，各自的总电压、总电流变化如图 1.8 所示。



(a)



(b)

图 1.8 (a) 停车充电数据片段；(b) 驾驶过程数据片段

二、数据分析

1、充电数据片段的容量标签计算

在充电片段的数据帧较少时，有两种情况：1、大电流充电， ΔSOC 较大， C^* 计算误差偏大；2、小电流充电， ΔSOC 过小，会导致 C^* 的计算误差被倍数放大。因此在选取充电片段时仅保留数据长度大于 100 条的片段。

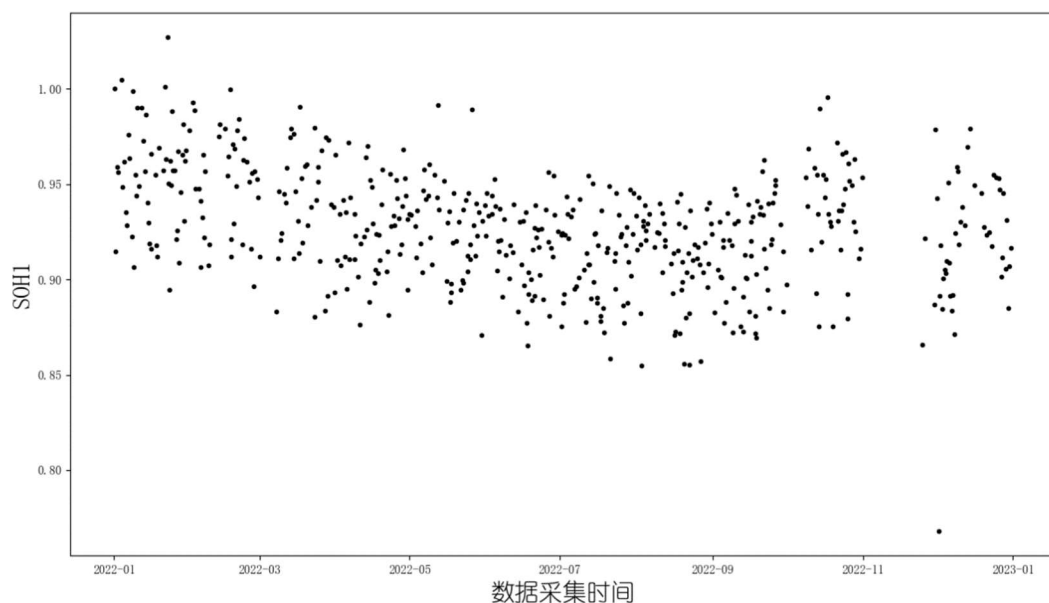
对于未知电池组额定容量的 SOH 计算，通常有两种方法：

- 1、 计算充入容量后，通过该过程变化的 SOC 计算电池组最大容量，并以第一个循环的最大容量为标准计算 SOH；
- 2、 选择特定电压范围内的充电数据，计算该范围内的电池充入容量，称为区域容量，以该指标代表电池组最大容量，并以第一个满足电压条件的循环的区域容量为标准计算 SOH。

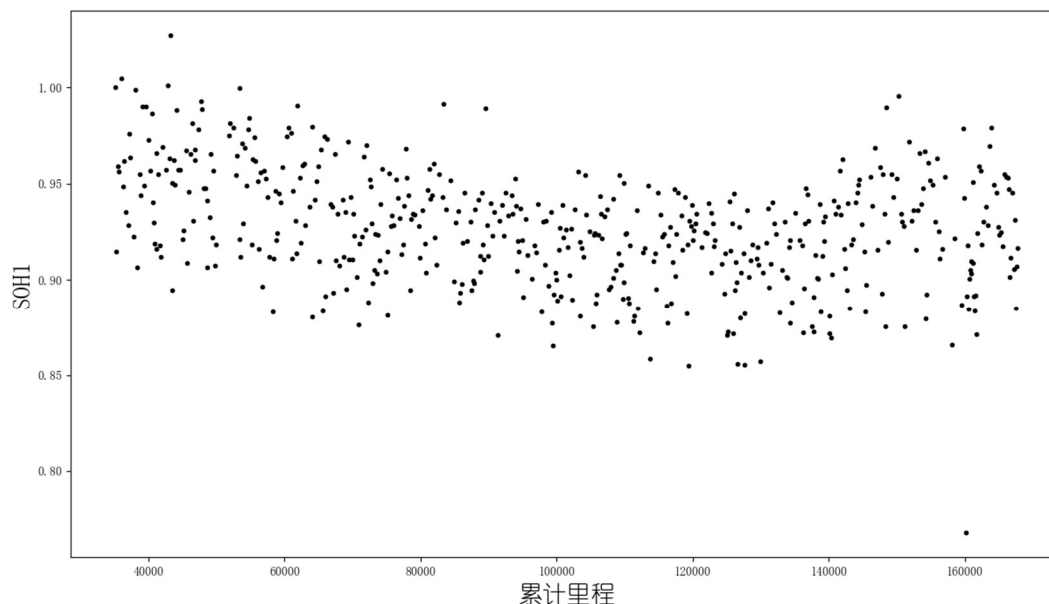
首先，按照第一种方法的计算方式，即：

$$SOH_1 = \frac{C^*}{\Delta SOC}$$

由于实验室中的电池容量变化总是有明确的循环次数，而实车数据中去计算该参数准确性较差。从图 2.1 中可以看出，采用累计里程作为横坐标时 SOH 变化的连续性更强，因此在后续的分析中应采用累计里程代替循环次数参数。



(a)



(b)

图 2.1 (a) SOH 随时间的变化; (b) SOH 随累计里程的变化

按照第二种方法计算 SOH 时, 需要选择触发计算的电压范围, 有研究表明, 当该电压范围包含 IC 峰时能够更准确的表示整体的 SOH 变化, 因此计算每个充电片段的 IC 曲线(图 2.2)。从图中可以看到, 在 3.8V 以下和 4.0V-4.12V 之间各存在一个明显的 IC 峰。统计各停车充电片段的充电开始时单体电压和结束时单体电压的分布(图 2.3), 可以看到包含 3.8V 以下电压范围的充电片段较少, 而包含 4.0V-

4. 12V 电压范围的充电片段样本丰富。

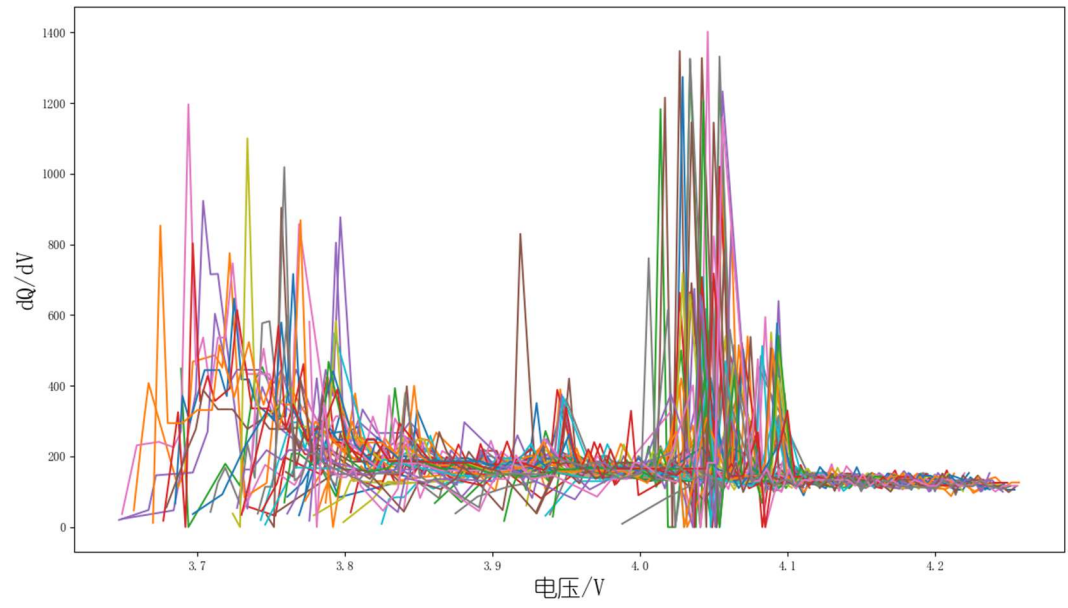
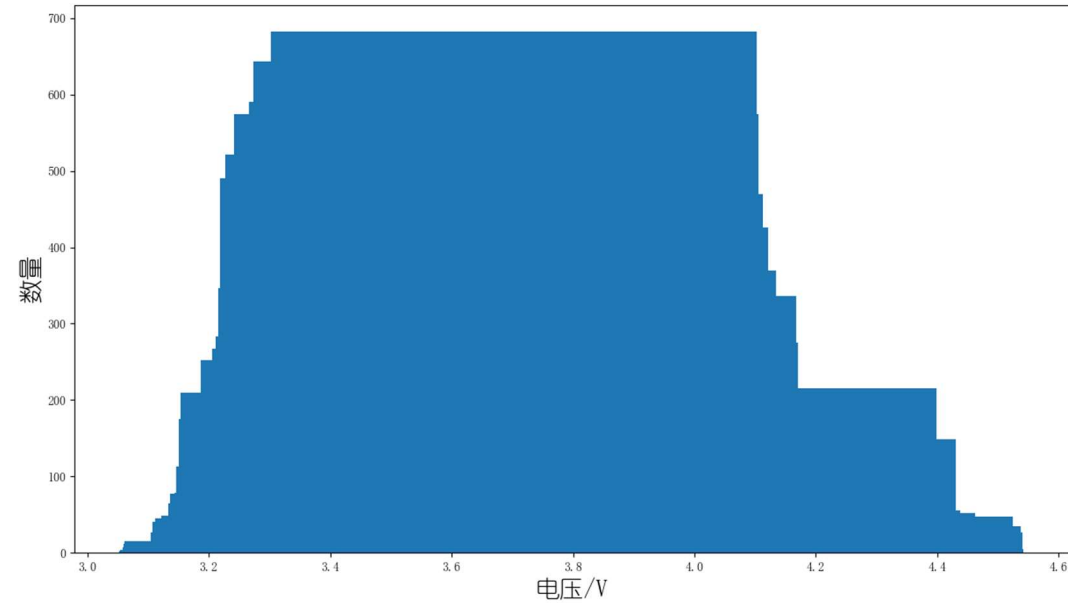
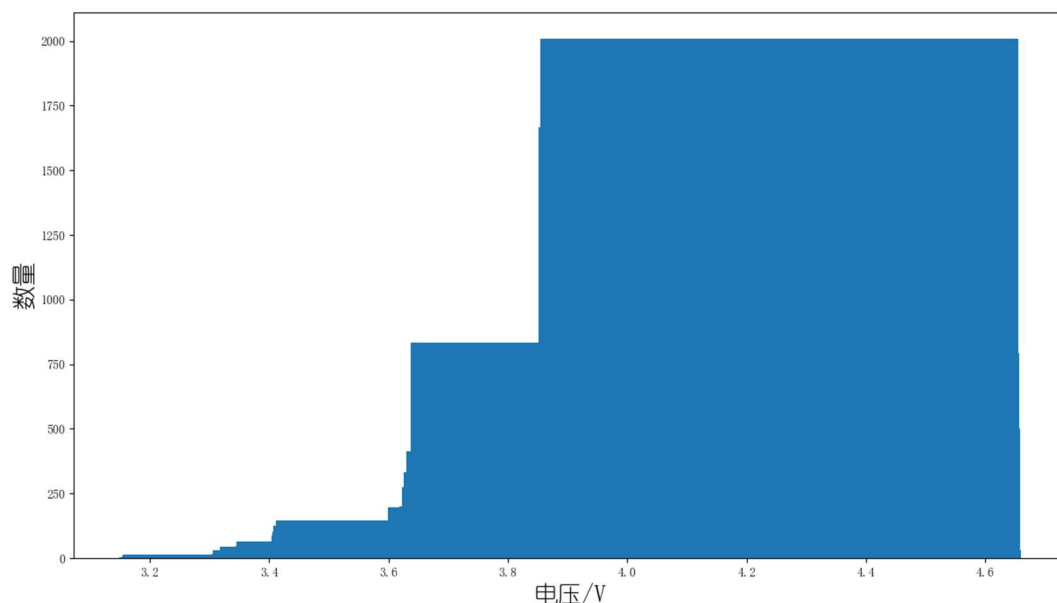


图 2.2 不同停车充电片段的 IC 曲线
(由于片段过多, 仅展示 1、11、21……等片段)



(a)



(b)

图 2.3 (a) 充电开始电压分布；(b) 充电结束电压分布

选择充电起始电压的单体最大值小于 4.000V 且充电结束电压的单体最小值大于 4.120V 的充电片段计算由区域容量定义的 SOH (图 2.4)。

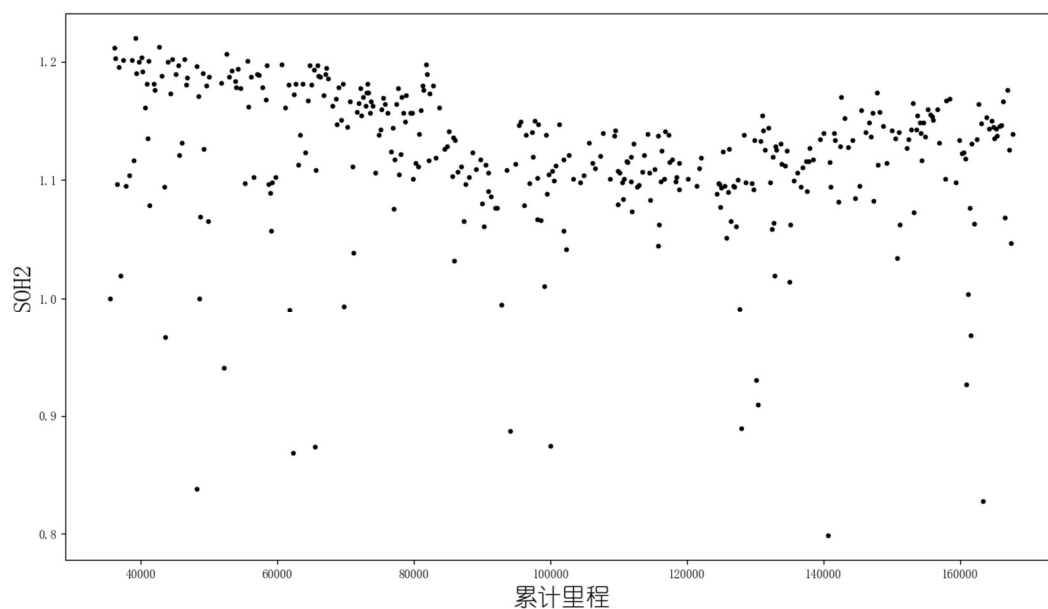


图 2.4 区域容量计算的 SOH 变化

由于电池组的容量变化应该是缓慢的，而这两种方法计算出的结果都带有较大的波动，而这两种方法计算结果波动较大时的里程数不同 (图 2.5)，因此可以通过分别给予权重结合两种方法的计算结果，

使 SOH 的波动减小，即：

$$SOH_3 = w_1 * SOH_1 + w_2 * SOH_2$$

$$w_1 + w_2 = 1$$

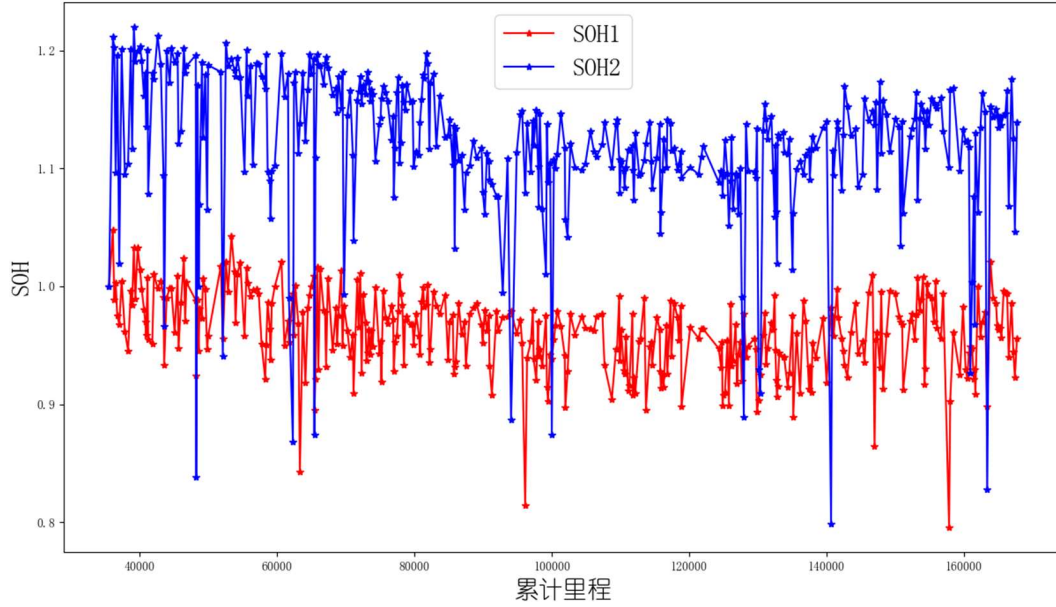


图 2.5 不同方法计算的 SOH 变化

根据目前市场销售纯电车辆的续驶里程及电池的寿命，可以假设行驶里程在 10000km 内时电池组容量衰减很小，将数据记录中的里程按 10000km 一段划分，最小化各段里程范围的电池组 SOH 方差即可得到权重值，即：
$$w = \min_{voltage_range} \sum_{i=0}^{n-1} var(SOH_3|_{mileage_i}^{mileage_{i+1}})$$

通过粒子群算法寻优，得到权重值为[0.87336437 0.12663563]，从优化后的结果（图 2.6）可以看到结合两种方法后计算的 SOH3 的变化更加稳定。

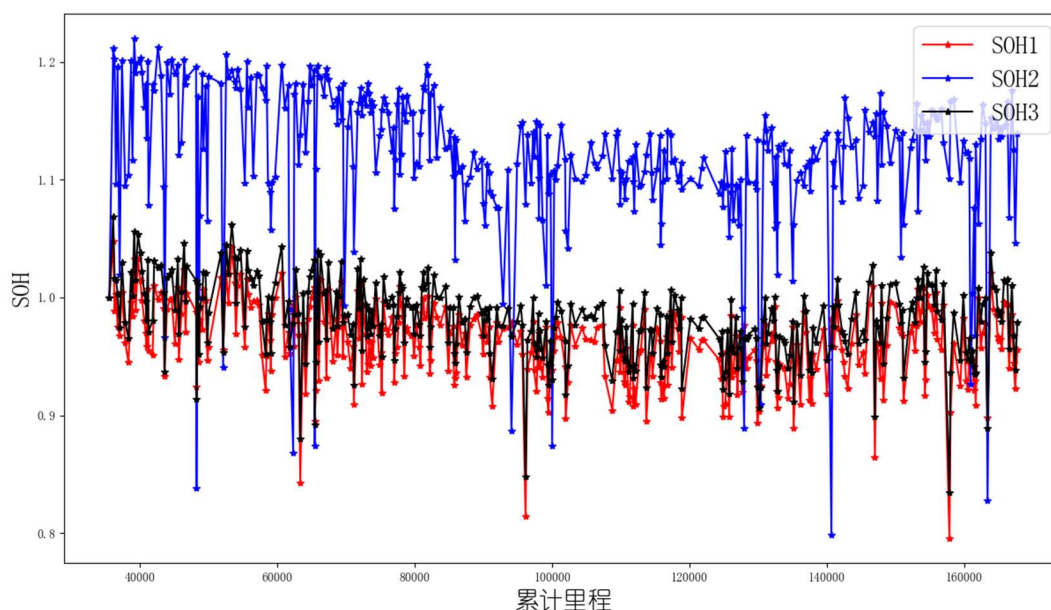


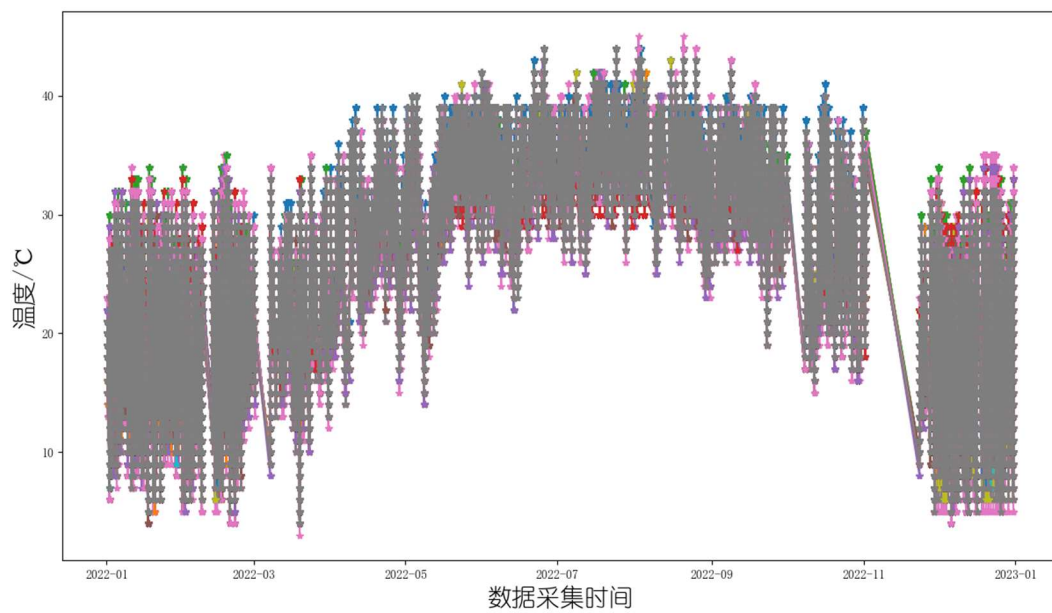
图 2.6 三种 SOH 的对比变化

2、温度对电池组的影响

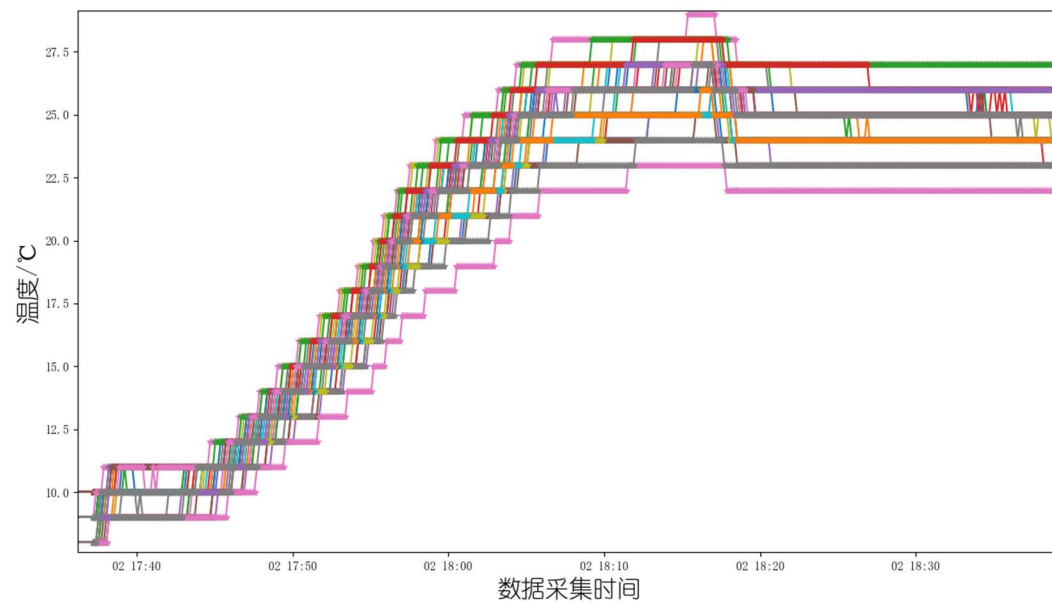
电池温度是影响电池容量的重要因素。

由于温度探针在电池组内的空间布置未知、电池组内电池单体的排布未知，且探针数量与单体数量不一致，因此对于某一个电池单体，无法得知哪个温度探针能够最准确的反应该单体附近的温度；；而如果在计算单体时考虑所有温度探针的值，则会造成大量的数据冗余，极大增加计算成本。

绘制各温度探针的温度值变化情况（图 2.7），从图中可以看出在充电开始段温度值的一致性较高，随充电时间增长，一致性逐渐变差。绘制每个数据帧温度探针的方差图来反映电池温度的一致性情况（图 2.8），方差最大数据帧位于 2022-12-18 16:24:08，此时温度范围为 18°C - 28°C ，温差为 10°C （图 2.9），该温度仍位于电池的适宜使用区间，对电池容量的差异性影响较小。



(a)



(b)

图 2.7 (a) 探针温度值变化；(b) 某片段的局部放大情况

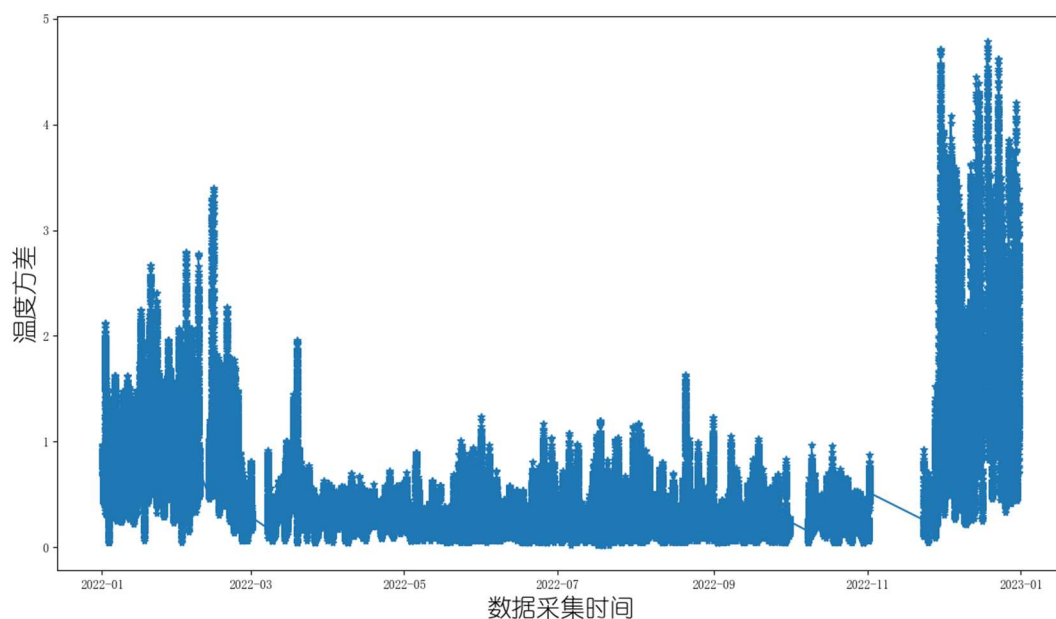


图 2.8 温度方差变化

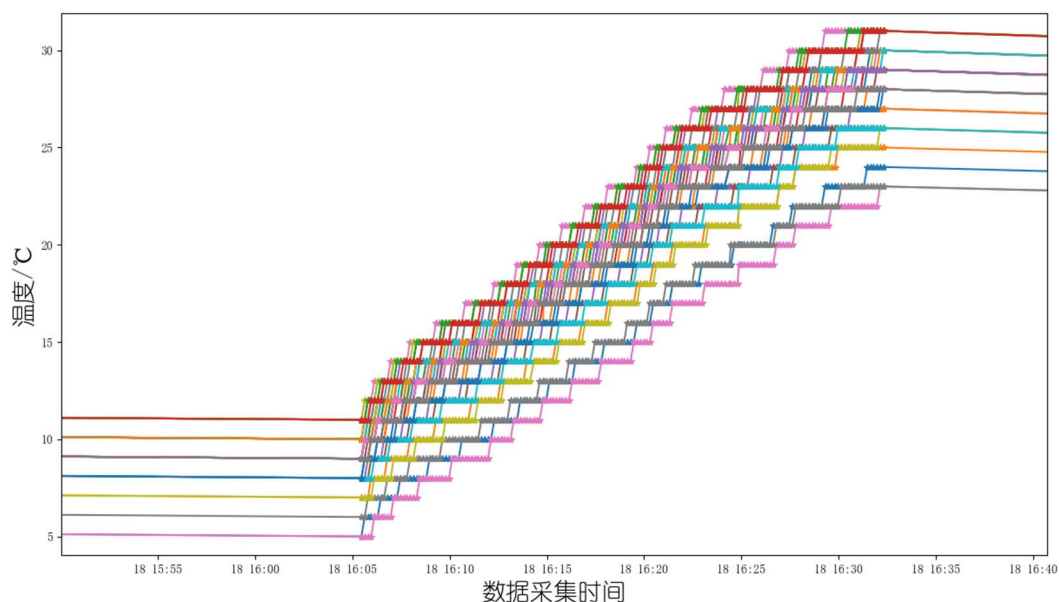


图 2.9 最大方差处温度情况

从数据集中“sing_volt_num”和“sing_temp_num”字段可以得知单体数量为 96，温度探针数量为 48，恰好为 2:1 的比例，且温度探针在电池组内通常为均匀布置，猜测每两个电池单体共用一个温度探针，这两个电池单体的一致性可以认为很好，因此可以在模型构建中将温度探针的特征矩阵从 48 直接升为 96，再与其他单体特征匹配。

三、特征选择

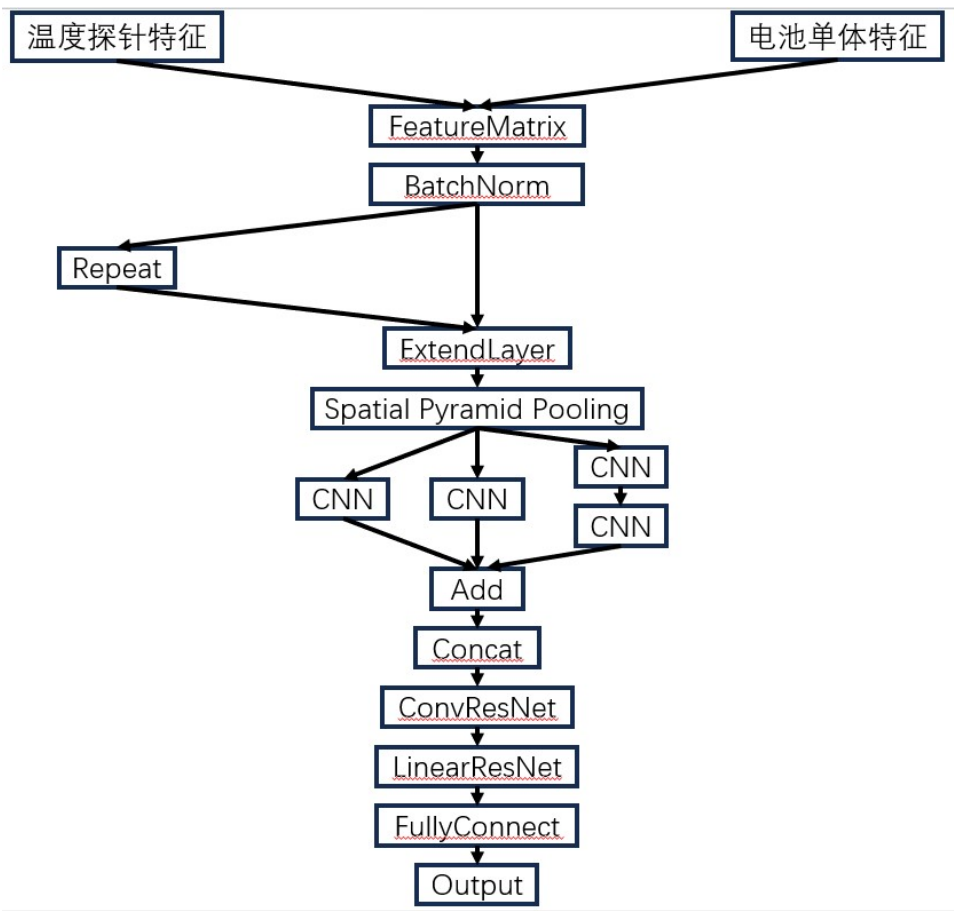
根据锂电池老化影响因素研究现状,可以得知电池循环的 SOC 区间、电池放电深度、放电倍率、电池工作温度等会造成锂电池三种老化类型不同程度的变化,从而影响容量衰减速度;根据锂电池放电容量的影响因素研究现状,电池工作温度、放电倍率等条件不同会造成锂电池的最大放电容量变化;另外,根据电池容量恢复的影响因素研究现状,在电池静置过程中,由于负极表面锂枝晶的重新激活嵌入、负极活性材料的变化等现象,锂电池的最大放电容量会出现增加现象。

基于上述影响因素,对于充电片段,选择结束电流、平均电流、充电方差、充入容量、持续时间、电池组起始 SOC、电池组结束 SOC、电池组 SOC 改变量、电池组平均 SOC 和累计行驶里程代表电池组的基本信息,由于电池组的容量受限于电池组中所有单体的最大容量和最小容量,所以还选择了充电时单体的起始电压和结束电压代表单体的信息,以上数据组合作为电池组伏安特征,选择各温度探点的最大温度、最小温度、平均温度、温升、平均温升速率、温度方差作为电池组热特征;对于放电片段,由于其数量远大于可计算容量的充电片段数量,因此将静置期间代表容量恢复的放电前后的静置期时间和温度四个特征加入放电片段的容量计算特征,在充电片段的伏安特征选择基础上再增加起始里程、结束里程、平均速度,热特征选择与充电片段一致。

在神经网络的构建上,由于锂电池的当前容量受到之前容量损失模式的影响,因此需要构建时间序列数据来估计每个片段的容量。对

于时间序列数据，通常使用循环神经网络或 Transformer 神经网络结构处理，而由于锂电池技术的进步，循环寿命通常达到上千次，在车辆的使用场景下，充放电片段数量将是循环寿命的几倍，如果使用 Transformer，则需要计算庞大的相关性矩阵，因此选择循环神经网络中成熟的 LSTM 网络作为处理时间序列数据的主体。在 LSTM 中，隐状态中存放的是历史容量的所有信息，而锂电池的当前时刻状态部分信息又需要只考虑短期因素的影响，如在短期内的温度变化和活性材料损失在静置中的变化，只有最近的几个片段对当前时刻计算有参考性，对于这种短期序列，选择使用 Encoder 模块处理可以利用其中的并行计算模式，增加计算的效率。

因此，最后的网络结构如图 3.1 所示。



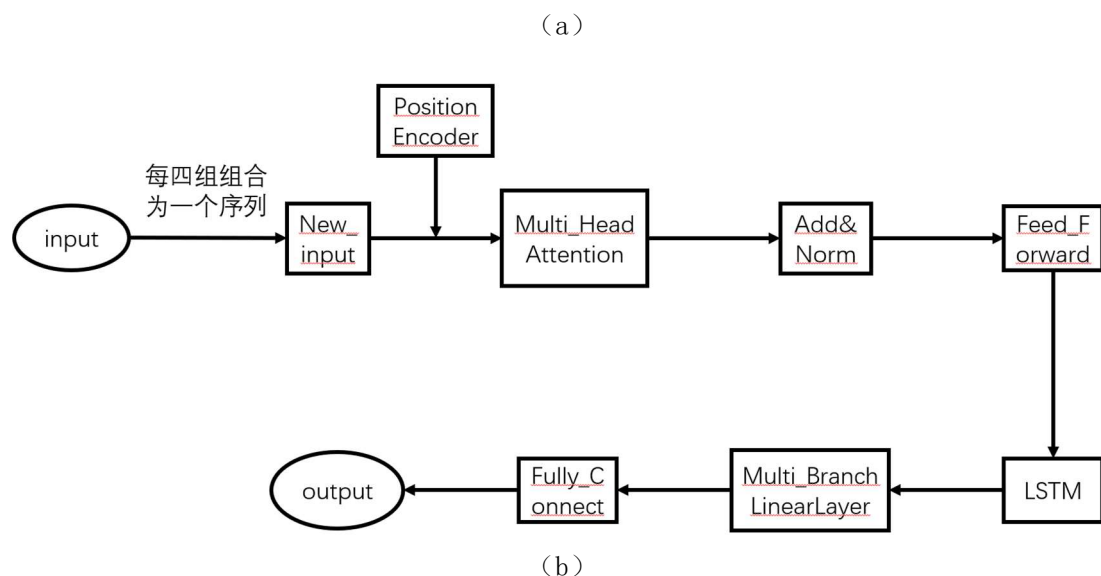


图 3.1 容量估计模型结构示意图 a)特征处理部分 b)序列处理部分

四、预测结果

十辆车充电数据共 3899 段停车充电片段,通过上文方法获得 SOH 标签的充电数据有 727 段。采用与容量相关的数据特征搭建深度学习模型,实现基于充电数据的容量估计,用于标注无 SOH 标签的 3272 段数据。

按照 4:1 的比例划分训练集与验证集,以未知 SOH 的数据集视为测试集。

由于各车辆的起始计算容量不同,因此若分别计算 SOH 作为训练目标是不合理的,选择直接计算电池组容量,以均方差损失作为损失函数。由于只有部分充电片段具有较准确的容量值,因此在数据集构造中,以-1 值填充无容量值的片段训练目标,在损失计算中,排除这些片段的损失。

训练过程中的训练集与测试集的误差变化如图 4.1 所示,可以看到两类误差基本为同步降低,可以认为未发生过拟合。

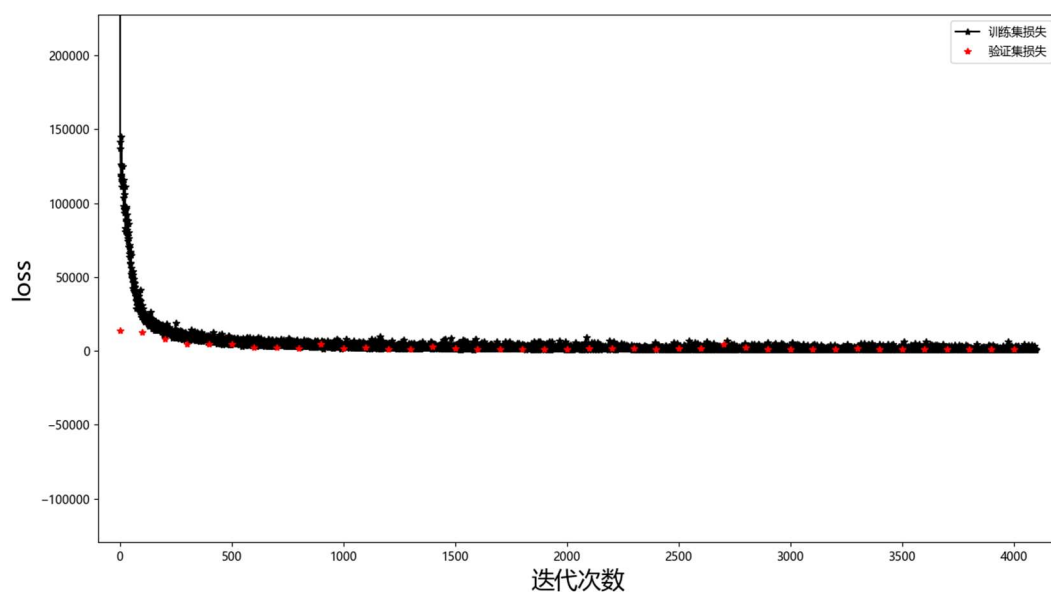


图 4.1 损失下降

十辆车的容量变化结果如图 4.2 所示，图中实际值由安时积分计算得到，可以看到经过模型估计后得到的容量值波动相较于安时积分计算减小了很多，模型在深度充电片段上的平均估计误差为 6.1%。

