基于wob和T数据的岩石属性回归预测

#### 研究数据及预处理：

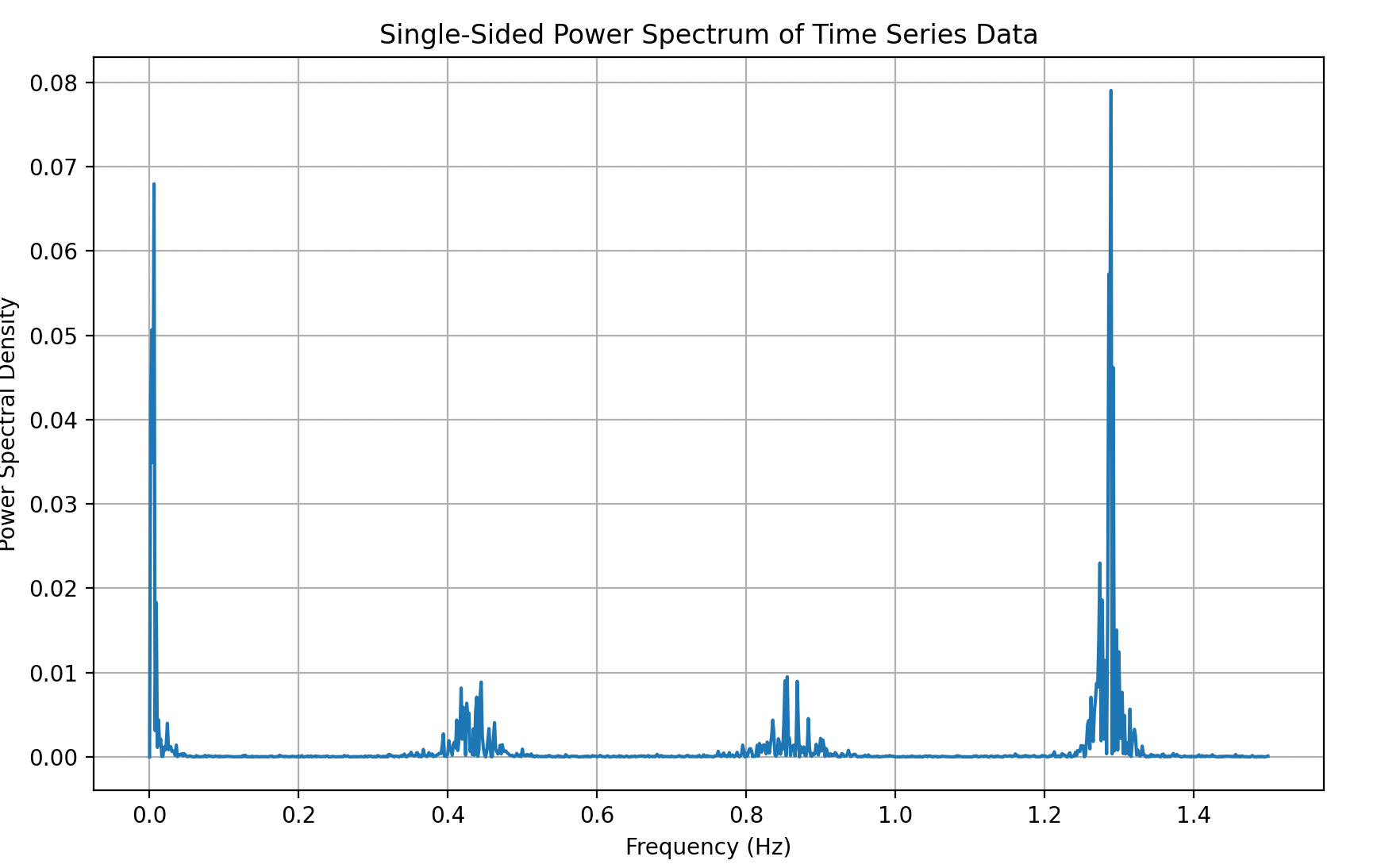
1:设备检测数据wob和T：

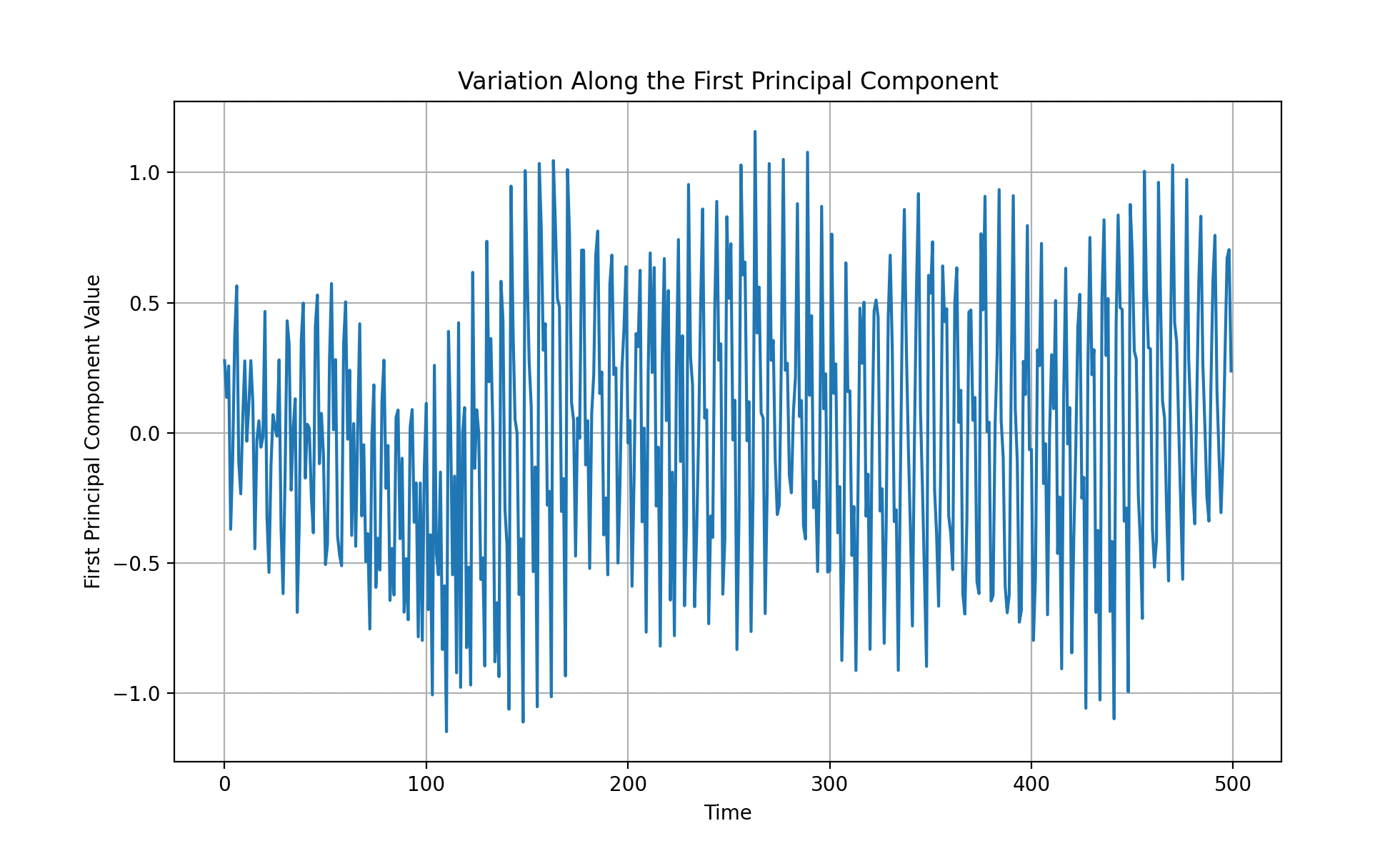
数据来源：

数据预处理过程：

基于采集的数据，采样间隔为每秒三次，数据分布如下：

对于检测的数据，我们以每三秒的时间间隔将数据划分为\*\*组，其中一组样本点数据包括三秒钟的检测数据wob和T。 对于划分好的数据，我们分别通过傅立叶变化以及K-L变化的方式提取了三秒时间内检测数据的频域特征和检测数据中的重要变化趋势，暂时结果如图：





可以看出，趋势比较稳定，适合我们直接去提取一些统计学特征。比如最大值、最小值、均值、标准差等。额外的，我们针对没列数据提取了数据的偏度系数以及峰度系数，用于拟合其属性值。

2:岩石属性数据：

数据来源：

数据预处理过程：

#### 算法设计与实现：

决策树算法是最具代表性的一种机器学习方法，它具有简单易懂、可解释性好、可根据不同的组合关系得到高精度的特点，在各类分类和回归问题中得到了广泛的应用。

集成学习是机器学习中一个非常重要的研究方向，在分类、回归和特征选择等问题上有着广泛的应用。集成学习不是单一的机器学习算法，它是一种将多种机器学习工具组合在一起的方法。该项目将为本项目的研究提供一种新的思路。

本次研究主要基于决策树的集成方案，

**决策树**：

决策树算法是一种具有代表性的机器学习方法，在处理各类分类与回归问题时得到了广泛的应用。决策树算法是一类具有代表性的树，一般由单根结点和若干叶结点构成；它既可以看作是一套 if-then规则集，又可以看作是特征空间和类别空间上条件概率分布的集合。

决策树中的节点是所有样本，每个节点中都含有一部分符合筛选条件的样本，节点之间的路径就是一个分割标准。如何寻找最佳的划分属性以及划分阈值是该算法的核心。决策树的基本构造方法是通过“分而治之”和“自顶而下”两种方法来构造决策树，该方法主要分为三个阶段：最优特征选择，决策树生成，决策树剪枝。

最佳特性的选取通常采用递归方法。首先，构造出最优的分区标准，以该分区标准对训练样本进行分割，使得各分区都能达到最适合的类别；决策树通常采用递归方法产生，它的终止条件有两个：一个是当前节点下的全部样本都是相同的；二是目前节点处的抽样数目低于预设的阈值。其中，对决策树进行剪枝是为了解决“过拟合”问题；由于去除了过分细分的叶子节点，简化树结构，提高了其推广能力。

本研究基于决策树的集成：

在GBDT模型的基础上通过构造一个目标函数来取代原先的基尼系数用以提高算法的精度。该目标函数由两个部分组成，一个是用来衡量预测值与真实值误差的函数，另一个是用来惩罚基模型复杂度的正则化项。从结构上来看，可以看做一种加法模型，通过累加的方式对结果进行预测。

算法构建决策树的基本思想是：首先，对数据集的全部特征按照特征的数值进行预排序；其次，对各特征寻找最优的分割阈值；最后根据最优的分割阈值将数据划分为左右子树。具体介绍一下该算法的流程。

假设给定训练数据集，其中是样本对应的真实值。首先给定K个回归树，其中是一个回归树，F是回归树的集合空间，T是树叶子节点数目，ω是叶子权重值，γ是叶子树的惩罚系数，λ是叶子权重的惩罚系数，为预测值，为真实值。为防止过拟合，将定义为正则化项（公式-1），则模型可以表示为公式2的形式，目标函数表示为公式3：

是第i个样本在第t次迭代的预测结果，是新增的树模型，是第t次迭代的目标函数，则其迭代过程如公式4、5所示：

其中为常量。如果模型的损失函数为均方误差时，公式5也可以写成：

当采用其他形式的损失函数时，一般在此处用其二阶泰勒展开表示：

其中，

移除掉其中的固定项，则第t次的目标函数可以表示为：

从上述过程可以看出本算法的目标函数主要依赖于损失函数的一阶偏导与二阶偏导，这是本方法相比于GBDT方法的一大优势，也是本方法和GBDT方法的主要区别。这里引入二阶导的目的主要有两方面：其一是可以提高模型拟合的精度；其二是为了能够使用二阶的泰勒展开来自定义损失函数。

不仅如此，本模型还在GBDT模型的基础上做了一些其他的改进，比如通过在目标函数中引入正则项用于控制模型的复杂度；采用了稀疏感知算法来加快节点分裂的速度；使用块结构的方式构建模型，可以很好的支持并行计算。