中国研究生创新实践系列大赛

“华为杯”第十六届中国研究生

数学建模竞赛

学　　校	华东师范大学
参赛队号	**19102690029**
队员姓名	**1. 胡亦秋**
	2. 王俊杰
	3. 向王涛

中国研究生创新实践系列大赛

“华为杯”第十六届中国研究生

数学建模竞赛

题 目 无线智能传播模型

摘 要：

本文使用赛题提供的训练数据，对数据信息进行挖掘，最终建立平均信号接收功率预测模型，在华为云平台成功部署，排名前 3%。本文主要完成了以下几方面的工作：

数据观察与清洗。本文通过特征分布信息可视化、多项地图信息可视化、计算基本统计量等方式，对数据进行了宏观观察，获取了一些先验信息。随后，从多个角度对数据集进行筛查，发现数据集无缺失值，一致性强，但存在若干异常。对异常点进行进一步分析讨论，最终确定了若干海拔异常点，由于异常点数量与数据集数量相比较小，直接对异常值进行剔除处理。最后将清洗后的数据集做了训练集与验证集的划分。

针对问题一，本文调查了 Cost 231-Hata 模型的详细资料，并从中选取出合适的基于专家经验的特征，结合数据集的几何信息、发射机工程信息、地物类型索引等信息为数据集设计了合适的特征，并阐明了构造理由。

针对问题二，本文从传统信道模型、数据集信息、算法生成三个方面构造特征，并进行筛选。构造过程中，查找了多个传统信道传播模型，并选取典型有效代表，从中构造合适的特征；在数据集上构造了几何信息特征，加入发射机工程信息、地物类型索引特征；对前两种方法构造出的单一特征进行多项式变换、对数变换、指数变换，从而丰富特征集。筛选过程中，使用方差、皮尔森系数、斯皮尔曼系数等指标评价特征，并结合随机森林算法为特征重要性评估，最终对四项排名取均值，获取了特征排名。

针对问题三，我们在比较后使用了五层神经网络作为本文主要模型，使用 leakyRelu 为激活函数，并添加 BatchNormalization 层，最终在在本地获得了 9.06 的验证集误差。

最后我们对我们的模型进行了评价总结，指出了不足。

关键字：信道建模 特征工程 机器学习

目 录

一、问题重述

1.1 问题背景

随着 5G NR 技术的发展，5G 在全球范围内的应用也在不断地扩大。运营商在部署 5G 网络的过程中，需要合理地选择覆盖区域内的基站站址，进而通过部署基站来满足用户的通信需求。在整个无线网络规划流程中，高效的网络估算对于精确的 5G 网络部署有着非常重要的意义。无线传播模型正是通过对目标通信覆盖区域内的无线电波传播特性进行预测，使得小区覆盖范围、小区间网络干扰以及通信速率等指标的估算成为可能。由于无线电波传播环境复杂，会受到传播路径上各种因素的影响，如平原、山体、建筑物、湖泊、海洋、森林、大气、地球自身曲率等，使电磁波不再以单一的方式和路径传播而产生复杂的透射、绕射、散射、反射、折射等，所以建立一个准确的模型是一项非常艰巨的任务。

现有的无线传播模型可以按照研究方法进行区分，一般分为经验模型、理论模型和改进型经验模型三类。经验模型的获得是从经验数据中获取固定的拟合公式，典型的模型有 Cost 231-Hata、Okumura 等。理论模型是根据电磁波传播理论，考虑电磁波在空间中的反射、绕射、折射等来进行损耗计算，比较有代表性的是 Volcano 模型。改进型经验模型是通过在拟合公式中引入更多的参数从而可以为更细的分类场景提供计算模型，典型的有 Standard Propagation Model（SPM）。

在实际传播模型建模中，为了获得符合目标地区实际环境的传播模型，需要收集大量额外的实测数据、工程参数以及电子地图用来对传播模型进行校正。此外无线 LTE 网络已在全球普及，全球几十亿用户，每时每刻都会产生大量数据。如何合理地运用这些数据来辅助无线网络建设就成为了一个重要的课题。

传统的无线传播模型的建立过程中，往往首先需要对传播场景进行划分，每一个场景对应一个传播经验模型。然而，经验模型在实际使用中往往不够精确，所以仍然需要通过采集大量的工程参数以及实际平均信号接收功率（Reference Signal Receiving Power，RSRP）测量值进行经验模型公式的修正。从所述过程中可以看到，传播模型建立本质上是一个函数拟合的过程，即通过调整传播模型的系数，使得利用传播模型计算得到的路径损耗值与实测路径损耗值误差最小。所以当工程参数、地理位置信息、特定地理位置测量点的 RSRP 已知的情况下，该问题可以归类为一个监督学习问题。

近年来，大数据驱动的 AI 机器学习技术获得了长足的进步，并且在语言、图像处理领域获得了非常成功的运用。伴随着并行计算架构的发展，机器学习技术也具备了在线运算的能力，其高实时性以及低复杂度使得其与无线通信的紧密结合成为了可能。因此，我们希望可以站在设备供应商以及无线运营者的角度，通过合理地运用机器学习模型来建立无线传播模型，并利用模型准确预测在新环境下无线信号覆盖强度，从而大大减少网络建设成本，提高网络建设效率。

1.2 数据集描述

数据集包含 4000 个小区，共计 120, 118, 33 条数据信息。具体包括了工程参数数据、地图数据和 RSRP 标签数据，其中为了方便数据处理，地图进行了栅格化处理，每个栅格代表了 5m×5m 的区域。

工程参数数据记录了某小区内站点的工程参数信息，共有 9 个字段，涵盖了小区编号、小区站点栅格位置、小区发射机相对地面的高度、小区发射机水平方向角、小区发射机垂直机械下倾角、小区发射机中心频率、小区发射机发射功率等信息。

地图数据记录地形地貌等信息，共有 8 个字段，涵盖了小区站点所在栅格的建筑物高度、小区站点所在栅格的海拔高度、小区站点所在栅格的地物类型索引、栅格位置、栅格上的建筑物高度、栅格上的海拔高度、栅格上的地物类型索引。其中地物类型名称的编号含义见表 1.1。

表 1.1 地物类型名称的编号含义

编号	含义	编号	含义
1	海洋	11	城区高层建筑（40m~60m）
2	内陆湖泊	12	城区中高层建筑（20m~40m）
3	湿地	13	城区<20m 高密度建筑群
4	城郊开阔区域	14	城区<20m 多层建筑
5	市区开阔区域	15	低密度工业建筑区域
6	道路开阔区域	16	高密度工业建筑区域
7	植被区	17	城郊
8	灌木植被	18	发达城郊区域
9	森林植被	19	农村
10	城区超高层建筑（>60m）	20	CBD 商务圈

数据样例见表 1.2。

表 1.2 数据集信息

工程参数数据								
Cell Index	Cell X	Cell Y	Height	Azimuth	Electrical Downtilt	Mechanical Downtilt	Frequency Band	RS Power
2	100	100	49m	45°	2°	2°	1800MHz	18.2 dBm
地图数据								
Cell Altitude	Cell Building Height	Cell Clutter Index	X	Y	Altitude	Building Height	Clutter Index	
47m	9m	11	500	500	9m	0m	1	
RSRP 标签数据								
RSRP								
-100 dBm								

由此可见，数据集较为详细地描述了站点的位置、高度、发射方向等信息，以及接收端的位置、高度、地貌信息，具有相当的数据挖掘价值。

平均信号接收功率(RSRP)标签数据作为实际测量结果，在监督学习中用于和机器学习模型预测的结果作比较，结合电子地图数据中的坐标和特征以及标签数据中的 RSRP 值，可以清晰地对信号功率分布进行可视化处理，从而明确辨识信号强弱覆盖区域。

1.3 问题描述

问题一：特征工程中的特征设计。

特征工程的本质是从原始数据中转换得到能够最好表征目标问题的参数，并使得各个参数的动态范围在一个相对稳定的范围内，从而提高机器学习模型训练的效率。

一般特征工程需要先对数据进行预处理，如去除异常值、补充缺失值等。高阶的特征工程需要充分利用与目标问题相关的专业知识。对于信道传播模型问题，可以根据已知的几何位置来挑选合理的特征。与此同时，传统经验信道模型中涉及的参数也可以纳入特征工程的考察范围。

题目要求根据 Cost 231-Hata 模型以及以表 1.2 的数据集信息设计合适的特征，并阐述原因。

问题二：特征工程中的特征选择。

完成特征设计后，通常需要选择有意义的特征输入机器学习模型进行训练。对于不同方法构造出来的特征，需要从多个层面来判断这个特征是否合适。

题目要求基于提供的各小区数据集，设计多个合适的特征，计算这些特征与目标的相关性，并将结果量化、排序，形成如下的表格，并阐明设计这些特征的原因和用于排序的量化数值的计算方法。

问题三：RSRP 预测。

根据建立的特征集以及赛题提供的训练数据集，建立基于 AI 的无线传播模型来对不同地理位置的 RSRP 进行预测。题目要求模型在预测过程中，弱覆盖识别率不能小于 20%，预测均方根误差需要尽可能地大。需要详细阐述模型的建立方法、参数设置以及训练结果等信息。

二、模型假设与符号说明

2.1 模型假设

本文模型基于以下合理假设。

假设 1：本题中所提供的数据是基本真实可靠的，本题对地图进行的栅格化是可靠的，并不会对结果带来较大影响；

假设 2：获取数据时，测量是基于相同的环境的，并且站点发射机天线内部型号、参数相同，接收设备相同；

假设 3：栅格上的信号强度仅受该栅格所在小区站点的影响，而不受其他站点信号干扰；

假设 4：所有小区栅格内的建筑为相似的建筑，接收点为栅格上建筑的顶端，并且栅格内无其他建筑；

假设 5：认为栅格内地形是均匀的，即同一海拔、被建筑物完全填充，对于发射机栅格，认为发射机被栅格内建筑物围绕。

2.2 符号说明

本文所使用的符号说明见下表 2.1。

表 2.1 符号说明

符号	含义	符号	含义
θ_{MD}	发射机机械下倾角	d	目标栅格与小区发射机栅格之间的距离
θ_{ED}	发射机垂直电下倾角	d_a	目标栅格与天线的距离
θ_A	发射机水平方向角	d_p	目标栅格与信号线在$\mathrm{x}oy$平面投影直线的距离
h_h^b	发射机相对地面的高度	d_s	目标栅格与信号线的距离
h_a^b	发射机所在栅格海拔高度	Δh_v	目标栅格与信号线的相对高度
h_b^b	发射机所在栅格的建筑物高度	φ	目标栅格与信号线夹角
h_a	目标栅格的海拔高度	C^b	发射机所在栅格的地形种类
h_b	目标栅格的建筑物高度	C	目标栅格的地形种类

三、数据观察及预处理

由于本题提供了大量的数据信息，所以本文事先分析数据集的统计特征来指导模型的建立。其中包括数据集的可视化以及各种基本信息的统计，从而能够对整个数据集有着直观全局的印象。根据可视化的结果以及相应的统计信息，我们使用数据预处理方法对数据进行清洗，包括对噪声、失真数据进行剔除，这些异常数据往往会直接影响模型的拟合效果。此外，当数据集存在缺失值时，需要对缺失值根据前后数据统计信息进行补值，帮助模型拟合生成更加通用的函数。另外，由于各个基本信息的取值范围变化幅度不一，直接放入模型中会导致效果不佳，所以还需要对数据做归一化处理，消除量纲影响，以便后期使用。

3.1 数据宏观统计与可视化观察

首先，我们对数据进行了一些宏观统计。训练集中共有 4000 个小区的文件，共计 120,118,33 条数据信息。通过对文件的遍历，我们发现每个小区仅有一个发射机站点，并且每个基站仅有一个发射机，并且同一文件内没有重复的目标栅格点。

我们将索引、栅格坐标外的数据集内原始特征进行统计，绘制出直方图，希望可以对原始特征分布有大概的印象，见下页图 3.1。从图 3.1 中可以发现，整个小区的海拔高度大致集中在 500 米左右，海拔为 0 到 400 米的小区基本不存在。同时可以发现接受点的建筑物高度为 0 居多，剩余少数高度在 100 以内，这与移动设备遍布的实际情况相符。同时，我们可以发现对于发射站所在的栅格点内建筑物高度基本为 0，可以基本认为信号塔在发射处没有受到障碍物的影响。此外，可以发现发射点的地形和接受点的地形分布大致相同，

以市区开阔区域和道路开阔区域居多。

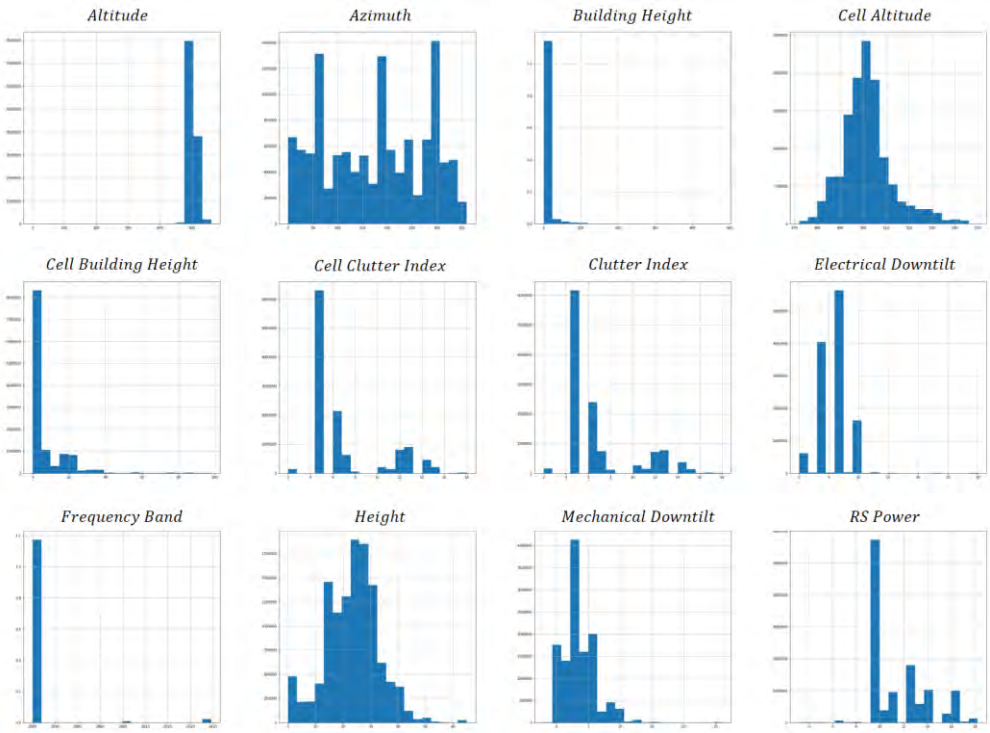

图 3.1 原始特征计数直方图

同时计算除地理坐标外其它特征的各个基本统计量，包括均值、标准差、最小值、四分位数、最大值，如下表 3.1 所示。结合可视化图片数据与统计值，进行数据的宏观观察，为下一步的数据清洗寻找依据，也为特征工程提供参考。

表 3.1 数据集基本统计量

	mean	std	min	0.25	0.5	0.75	max
Height	23.23	9.59	0	17	24	30	65
Azimuth	172.59	103.29	0	70	180	270	360
Electrical Downtilt	5.13	2.39	0	3	6	6	30
Mechanical Downtilt	3.49	2.45	-2	2	3	5	26
Frequency Band	2585.83	5.44	2585	2585	2585	2585	2624.6
RS Power	11.27	2.51	3.2	9.2	10.2	13.2	18.2
Cell Altitude	501.76	11.03	472	495	501	507	550
Cell Building Height	5.75	11.51	0	0	0	6	98
Cell Clutter Index	7.04	3.24	2	5	5	7	18
Altitude	501.57	11.00	0	495	501	507	561
Building Height	4.10	14.74	0	0	0	0	480
Clutter Index	6.85	3.04	2	5	5	7	18
RSRP	-91.80	10.71	-140	-99	-92	-84.5	-44

从统计表中可以明显看到，各个指标的量纲差距较大，后续特征工程中在此基础上构建的特征也很可能存在较大的量纲差距，因此在输入模型前，对特征进行归一化是必要的。

另外，原始特征的方差相差并不大，而特征如果不发散，则这个特征对于样本的区分并没有什么作用，说明原始特征不宜直接作为特征输入模型。

在分析基本信息的统计量之后，我们对这些数据结合位置信息以地图的形式进行可视化，以便能够对数据得到更加直观的印象。缩略图见下图 3.2，原图打包于附件中。

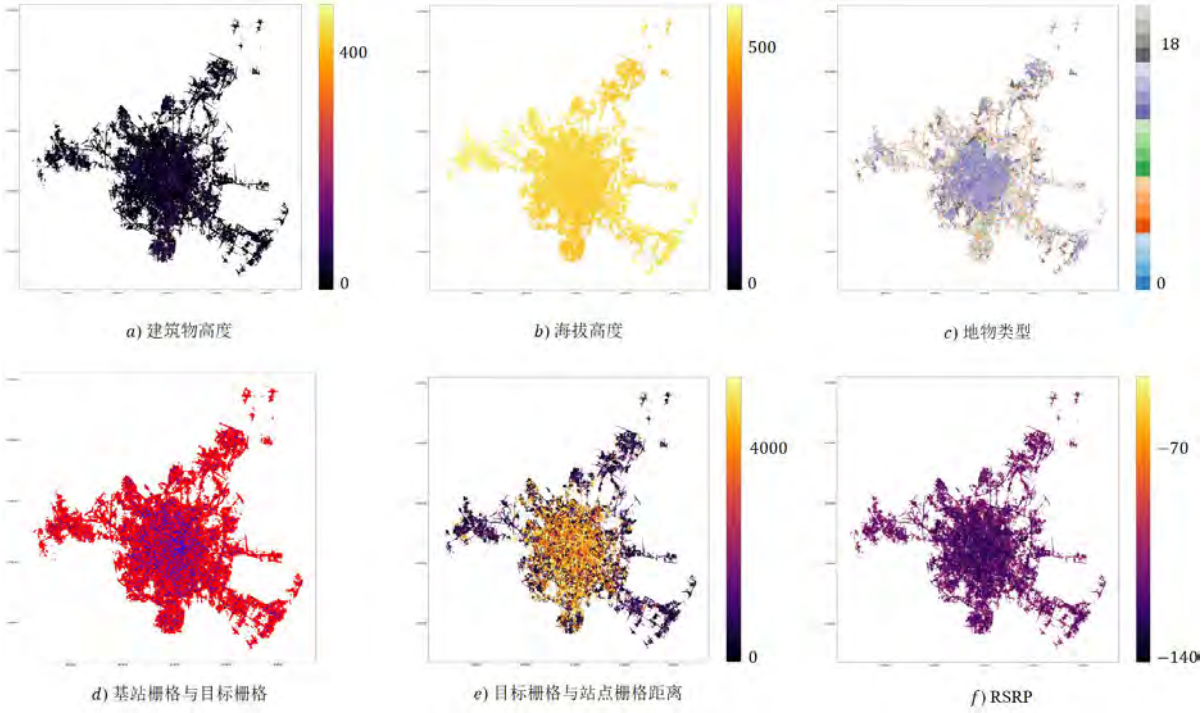

图 3.2 数据集可视化

在图 3.2.*a*中展示了建筑物高度的地图可视化，我们发现大量的建筑物高度在 0 附近，而最高值可达 400 左右，可据图猜测建筑物高度大约服从参数大于 1 的指数分布。同时我们注意到建筑物高度这一特征的第三四位数是 0，说明有建筑物的栅格在数据集中的比重小于四分之一，这也与之前大多数为道路开阔的地形信息相吻合。

在图 3.2.*b*中展示了海拔高度的地图可视化，可以发现整个图颜色十分接近，但有若干个海拔为 0 的黑点。同时，我们可以观察到海拔高度的第一四分位数是 495，平均数是 501.57，因此可以猜测海拔为相对较小值的栅格较少。根据常识，除非存在悬崖，海拔高度一般不可能存在突变，因此图中若干黑点可能是异常数据点，在训练模型时将其删除。

在图 3.2.*c*中展示了地物类型的地图可视化，可以发现存在大量的紫色区域，橙色明显为道路开阔区域，小区与小区之间联系较为密集，甚至有重叠部分，整个地物类型的分布情况符合现实中的城镇布局。

在图 3.2.*d*中，我们将发射机站点位置和小区接收点位置的地图等数据可视化，其中红色点为接收端，即目标栅格点，蓝色点为小区所在站点的栅格点。显然，栅格数据没有铺满整个地图，并且存在较大的空隙，中心区域数据更为全面，而外围区域数据则更为稀疏。

在图 3.2.*e*中展示了目标栅格与站点栅格距离的地图可视化，可以看到城市中心区域存在大量距离较大的点。

在图 3.2.*f*中展示了标签数据的可视化，可以看到靠近基站位置的 RSRP 值相比于郊区

更大，这为之后特征的设计提供了指导。

3.2 异常值筛选与预处理

对数据进行预处理能提高数据的质量，也能让数据更好的适应特定的挖掘模型，因此在对数据进行操作之前，需要先检查所得数据集是否有缺失值、异常值等情况，从而对数据进行适当清洗。

本文对数据集进行了以下筛查步骤：

1. 对数据集进行遍历，发现没有任何的缺失值；

2. 检查了每一个文件中的小区站点发射机的工程参数、发射机所在栅格的地图数据，发现其在同一个文件内均是统一的；

3. 通过在单个文件内的$[X,Y]$计数，与单个文件内的总行数比较，发现对单个文件内没有重复的目标栅格坐标点，虽然在所有数据中存在目标栅格点多次出现，但其发射机栅格不同，我们认为是正常数据；

4. 基于上一步，我们对整个数据集中$[X,Y]$相同的栅格点进行检查，发现其海拔、建筑物高度、地物类型等目标栅格均是统一的，不存在不一致；

针对可视化过程中，建筑物高度偏 0 的情况，对建筑物高度进行了进一步分析，发现高度分布较为连续，不存在明显异常点；

5. 针对可视化过程中，海拔点存在远远偏离总体分布的情况，对海拔高度分布进行了查询，发现低于 470 的值仅有五条数据，共计两个点。在图 3.3 中对该两点周围数据进行了可视化，可以看出该两点均为边缘点，且与周围临点存在明显突变，是异常值。由于数量少，因此进行了直接删除的处理。

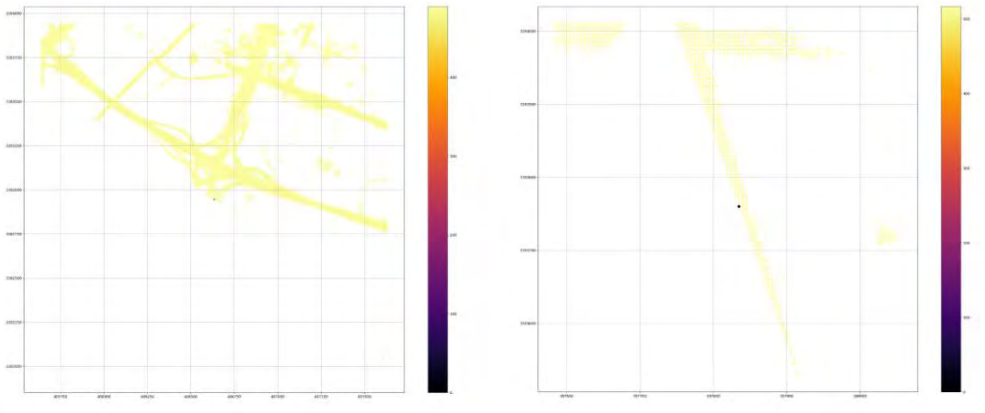

图 3.3 海拔高度异常点及附近区域可视化

删除后，重新绘制海拔高度的可视化地图，如图 3.4 所示，可以发现，海拔高度呈连续变化的趋势。

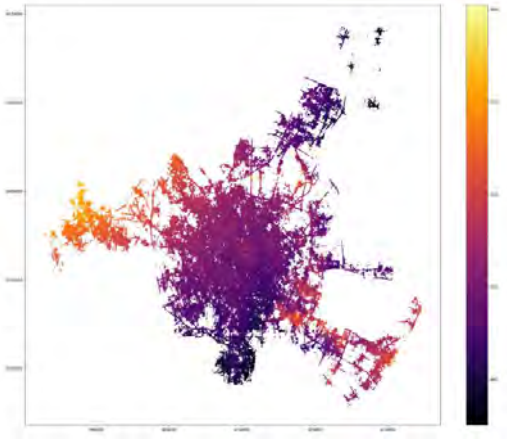

图 3.4 删除异常点后的海拔高度可视化

6. 针对可视化过程中，目标栅格与站点栅格距离分布范围较宽的情况，我们对此进行了进一步可视化分析，绘制了距离分布直方图，见图 3.5

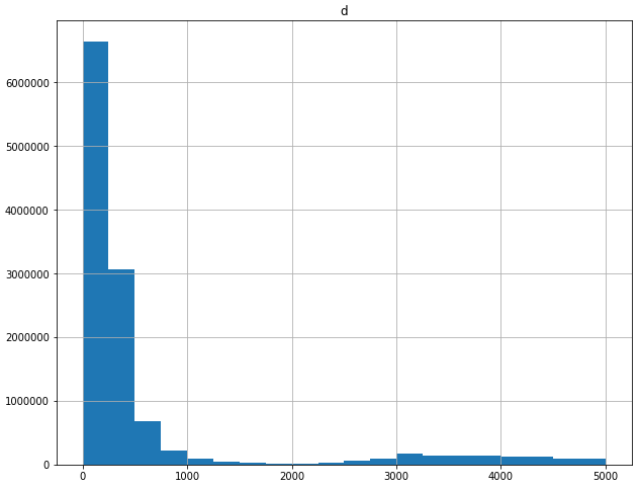

图 3.5 目标栅格与站点栅格距离统计直方图

从图中我们可以发现，总体分布较为连续，不存在明显异常的点。

7. 为后续训练及验证使用，本文按 8:2 的比例对数据集进行划分，取 80%为训练集，20%为测试集，划分后对两数据集进行了特征分布比对，发现大致相似，因此认为该划分是有效的。

四、问题一的模型建立与求解

4.1 问题分析

本题目的场景中，小区接收从某个站点发送的信号，并由于多种因素使得在小区的各个位置信号出现强弱变化。该问要求我们寻找可能的影响因子，从而设计与平均信号接收功率(RSRP)标签强相关的特征，并阐述原因。

根据题目可以得知特征主要可以从基于数据集所给数据进行设计和基于传统经验信

道模型两方面进行设计。前者是因为在物理层面上信号强度的强弱和距离，频率等因素都息息相关，在给定基本特征的基础上，通过对这些基本属性进行运算，能够得到具有具体物理含义的与目标值强相关的特征。后者则是领域内的专业人士根据多年的测量数据和从业经验得到关于信号强弱与一些因素的拟合公式，其中公式中的各个因素往往和结果RSRP 有着强烈的相关性。因此针对该问题，本文主要根据以上两点来进行特征设计。

具体而言，在基于传统信道模型 Cost231-Hata 设计时，重点需要考虑的是模型所使用的参数，以及模型是如何使用参数的——如对参数取对数、取平方等。这样的特征对于非专家来说是很难理解其具体的物理意义的，并且通过机器学习的方式也很难直接提取这些非线性的特征组合，但其往往与问题目标有着强相关性。

基于数据集信息设计的特征是从数据集所提供的参数寻找可能造成影响的因素[6]。直观上，接受点与站点的距离等因素对 RSRP 的强弱有着直接的影响。 因此我们考虑目标栅格、天线、信号线之间的几何位置关系来建立与目标值 RSRP 强相关的特征。最终，将以上两个方面寻找到的特征进行结合，即得到特征集合。

4.2 特征设计

4.2.1 基于 Cost 231-Hata 模型的特征设计

传统的经验信道模型是从经验数据中获取固定的拟合公式，包括 Okumura-Hata 和其改进版 Cost 231-Hata 等传统的拟合模型。其中 Okumura-Hata 模型[1]适用于 150-1500 MHz 的带宽。基站和接收者的距离应该保持在 1 到 20km 之内。其中间路径损耗定义为：

$$PL_{urban}(dB) = 69.55 + 26.16 \times log(f) - 13.82 \times log(h_t) - a(h_r) + [44.9 - 6.55\,log(h_t) \times log(d)] \tag{4.1}$$

其中，接收者的相关系数为：

$$a(h_r) = \begin{cases} 8.29[log(1.54h_r)]^2 - 1.1 & f_c \le 300MHZ \\ 3.2[log(11.75h_r)]^2 - 4.97 & f_c \ge 300MHZ \end{cases} \tag{4.2}$$

其中f代表频率，h_t代表有效的基站高度，h_r代表有效的接收者高度，d 代表基站和接收者之间的距离。其中天线有效高度定义如下：

$$h_t = Height + Cell\ \ Altitude - Altitude \tag{4.3}$$

此外，郊区的道路损失定义为：

$$PL_{suburban} = PL_{urban}(dB) - 2\left[log\left(\frac{f}{28}\right)\right]^2 - 5.4 \tag{4.4}$$

对于开阔的乡村区域，道路损失函数则为：

$$PL_{urual} = PL_{urban}(dB) - 4.78[log(f)]^2 - 18.33 \times log(f) - 40.98 \tag{4.5}$$

可以发现，Okumura-Hata 模型直接通过公式计算得到无线传播损失值，极大的方便了信号覆盖等后续实际工作的开展。然而由于模型过于简单，该模型的使用受到了诸多限制。比如 Okumura-Hata 模型只适用于 150-1500 MHz，为了适用于更高频率的无线通信，COST 231-Hata 模型[3]对其进行了改进，使其能够适应 1500-2000 MHz。该模型要求发射台的高度在 30 到 200 米之内，接收者的高度在 1 到 10 米之内，路径损失定义如下：

$$\begin{aligned} PL(dB) = 46.3 + 33.9 \times log(f_c) - 13.82 \times log(h_t) - a(h_r) \\ +[44.9 - 6.55\, log(h_t)] \times log(d) + C_m \end{aligned} \tag{4.6}$$

其中，接收者的相关系数为：

$$a(h_r) = \begin{cases} 8.29[log(1.54h_r)]^2 - 1.1 & f_c \le 300MHZ \\ 3.2[log(11.75h_r)]^2 - 4.97 & f_c \ge 300MHZ \end{cases} \tag{4.7}$$

场景纠正常数为：

$$C_m = \begin{cases} 0\text{dB} & \text{对于中型城市和郊区} \\ 3\text{dB} & \text{对于大城市} \end{cases} \tag{4.8}$$

相比于之前的 Okumura-Hata 模型，Cost 231-Hata 模型额外考虑了地形因素，这是因为信号在传播的过程中会发生绕射，衍射，反射等多种光学现象[8]，这些光学现象相互影响，使得接受点的信号损失很难计算得到[4]。同时，这些现象对于不同的地形通常对应着不同的损失值，比如信号通过楼层时的损失要比通过树木时的损失要小，此外现实中往往环境比较复杂[5]，通常是多种地形的叠加，所以考虑额外的地形影响会使得模型更加具有泛化性，能够在实际场景中取得不错的效果。

根据上述的经验模型可以发现，通过对基本信息进行固定的非线性组合，这些非线性的组合值就能够和最后的标签值 RSRP 存在着线性关系。因为线性关系是一种相对简单的关系，模型往往能够很精确地拟合这种线性映射，因此寻找和标志值有着线性关系的特征是非常重要的，这些经验公式为我们提供了很好的指导。

本文将 Cost231-Hata 中与预测值 RSRP 存在线性关系的表达式当作特征，从而在输入层增加特征的复杂度并且降低模型的训练难度。最终根据经验模型得到特征并带入问题一的数据集进行计算，结果如下：

表 4.1 基于 COST 231-Hata 模型的特征

索引	符号	值
1.1	$log(f_c)$	3.26
1.2	$log(h_t)$	1.94
1.3	$log(d)$	2.75
1.4	$log(h_t)\, log(d)$	5.34
1.5	$log(h_r)$	0

4.2.2 基于数据集信息的特征设计

除了专家经验外，物理上距离等因素必然对信号传播的损失起到决定性的作用，距离信号塔越远的地方接收的信号强度越小。本题中每个小区都有且仅有一个信号塔，并且每个信号塔上只存在一根天线。由于信号从天线发出，当信号塔的高度和天线的偏转角度固定后，必然存在一条虚拟的信号路径（信号线），其代表信息传播过程中离发射点相同距离中信号最强的集合，该信号线的角度通常和天线的偏转角度保持一致。因此我们可以认为，目标栅格的信号强弱与基站的相对位置信息，信号线的角度等因素都有着非常强的关联性。

根据题目提供的信号站位置等基本信息，我们可以通过信息之间的几何关系求得目标栅格接收端到发射基站底的距离、到发射天线的距离、到信号线的距离、到信号线在地面投影的距离、与信号线的高度差等信息，这些距离信息都可以作为模型输入的特征。另一方面，由于信号线可以看作是一空间射线，与天线到接收端的线段存在夹角，这一夹角或与信号损失相关。

基于数据集的位置和角度信息，本文提出以下 7 个基于几何位置的特征。

首先建立空间几何模型。由于我们仅需要考虑站点与目标栅格之间的距离关系，因而可以将站点发射机抽象为一条线段，天线抽象为线段的顶端，从天线发射出一条信号线，如下图建立三维直角坐标系来考虑相对位置关系：

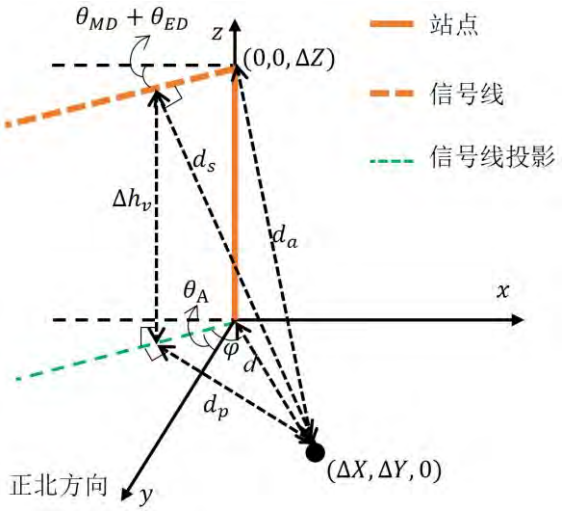

图 4.1 几何特征示意图

图中，θ_{MD}表示小区发射机垂直机械下倾角，θ_{ED}表示小区发射机垂直电下倾角，θ_A表示小区发射机水平方向角，φ表示目标栅格到原点连线与信号线夹角，$(\Delta X,\Delta Y,0)$表示目标栅格的坐标，$(0,0,\Delta Z)$表示天线坐标，Δh_v表示目标栅格与信号线的相对高度，d表示目标栅格与小区发射机栅格之间的距离，d_a表示目标栅格与天线的距离，d_s表示目标栅格与信号线的距离，d_p表示目标栅格与信号线在xoy平面投影直线的距离。

在该坐标系下，目标栅格坐标为：

$$(\Delta X,\Delta Y,0)=(Cell\ \ X-X,Cell\ \ Y-Y,0) \tag{4.9}$$

天线z轴坐标值为：

$$\Delta Z = Height + Cell\ \ Altitude - Altitude - Building\ \ Height \tag{4.10}$$

信号线方程l_s为：

$$x = \frac{\mathrm{y}}{\tan\theta_A} = \frac{z - \Delta Z}{\tan(\theta_{MD} + \theta_{ED})} \tag{4.11}$$

信号线在xoy平面投影的直线方程l_p为：

$$x\tan(\theta_{MD} + \theta_{ED}) + y = 0 \tag{4.12}$$

由于θ_{MD}、θ_{ED}、θ_A已知，即可构造出多个特征的模型：

$$\begin{cases} d = \sqrt{\Delta X^2 + \Delta Y^2} \\ d_a = \sqrt{\Delta X^2 + \Delta Y^2 + \Delta Z^2} \\ d_p = \dfrac{|\Delta X\tan\theta_A + \Delta Y|}{\sqrt{(\tan\theta_A)^2 + 1}} \\ \Delta h_v = \Delta Z - \Delta X\tan(\theta_{MD} + \theta_{ED}) \\ d_s = \sqrt{\left(\Delta Z - \tan(\theta_{MD} + \theta_{ED})\sqrt{d^2 - d_p^2}\right)^2 + d_p^2} \\ \varphi = \arctan\dfrac{\Delta X}{\Delta Y} + \theta_A \end{cases} \tag{4.13}$$

将表 1.2 中数据代入，最终可求得基于几何信息的特征，汇总后见下表 4.2

表 4.2 基于数据集几何信息的特征

索引	符号	值	含义
1.6	ΔZ	87	目标栅格与小区发射机栅格之间的高度差
1.7	d	565.69	目标栅格与小区发射机栅格之间的距离
1.8	d_a	572.34	目标栅格与天线的距离
1.9	d_p	565.69	目标栅格与信号线在xoy平面投影直线的距离
1.10	d_s	572.34	目标栅格与信号线的距离
1.11	Δh_v	59.03	目标栅格与信号线的相对高度
1.12	φ	90°	目标栅格与信号线夹角

另外，数据集还给出了一些发射机的工程信息、发射机栅格及目标栅格的地貌信息，

这些信息通常也直接和 RSRP 相关联。因此本文同样选取这些信息作为特征，罗列见下表 4.3：

表 4.3 基于数据集其它信息的特征

索引	符号	值	含义
1.13	f	1800	小区发射机中心频率
1.14	P_t	18.2	小区发射机发射功率
1.15	C^b	11	小区发射机所在栅格的建筑物高度
1.16	C	1	目标栅格的地物类型
1.17	h_b^b	11	小区发射机所在栅格的建筑物高度

4.3 小结

在本节中我们首先对问题进行了深入的分析，了解到由于实际环境的复杂性，试图直接找到 RSRP 与基本信息的映射关系是不可能的。所以本文同时根据以往的经验模型和几何特征来进行建模，以便于学习更有效的更具泛化性的映射函数。对于经验模型，我们返现一些基本信息的非线性组合往往和预测值存在着线性关系[2]，同时模型很容易学习线性映射，所以基于经验模型，我们将公式中的非线性组合当作特征进行输入。此外，根据通信知识和物理知识可以知道，几何因素同样对于 RSRP 起到至关重要的作用。为此，本文结合提供的基本几何信息，得到了更加丰富的几何特征，这些特征从各方面衡量了目标点与发射点的几何关系。最终结合两者得到关于问题一的特征及取值如下表，共计 17 项。

表 4.4 问题一的特征设计

符号	值	符号	值	符号	值
$log(f)$	3.26	d	565.69	f	1800
$log(h_t)$	1.94	d_a	572.34	P_t	18.2
$log(d)$	2.75	d_p	565.69	C^b	11
$log(h_t)\,log(d)$	5.34	d_s	572.34	C	1
$log(h_r)$	0	Δh_v	59.033	h_b^b	11
ΔZ	87	φ	90°		

五、问题二的模型建立与求解

5.1 问题分析

在本设问中，要求使用多种方法构造特征，并从多个层面量化判断所构造的特征是否

合适，形成排序并说明理由。因此本文工作也分为特征设计与特征排序两部分。

特征设计中，可以从传统信息传播模型参数、数据集信息、算法生成等三个方面出发进行构造。在基于传统信道模型的特征设计过程中，可以参考多方文献[9]，从经验模型、理论模型、改进型经验模型等多种模型中寻找合理的参数。在基于数据集的特征设计过程中，可以类似第一问中工作，基于单条数据的信息求取几何位置的信息特征，也可以纵览整个数据集，获得基于全局信息的特征。在算法生成方面，可以从单个可能造成影响的因素出发，考虑这些因素组合是否会产生更多作用，组合方式可以考虑相乘、相除、取对数、取指数等。

进行特征选择时，我们可以从多个角度为特征评分或排名。首先观察特征是否发散，若方差接近于 0，说明数据集在这个特征上基本没有差异，则该特征对于样本区分作用不大。另外可以计算特征与目标的相关性，包括线性相关性、等级相关性等。还可以考虑从机器学习的角度出发，使用随机森林等方法对特征进行评分排序。

5.2 特征设计

5.2.1 基于传统信道模型的特征设计

Stand Propagation 模型是经过 Hata 道路损失模型改进而来的。接收方的信号强度可以表示为：

$$P_r = P_t - \begin{Bmatrix} K_1 + K_2\log(d) + K_3\log(h_t) + K_4 DiffractionLoss \\ +K_5\log(d)\log(h_t) + K_6 h_r + K_7\log(h_r) + K_{clutter} f_{clutter} + K_{hill} \end{Bmatrix} \tag{5.1}$$

其中$K_1, K_2, K_3, K_4, K_5, K_6, K_7, K_{clutter}$为乘积因子。$P_r$为接收方的信号强度，$P_t$为基站的信号强度，$d$为两者之间的距离，$h_t$为有效的发射天线高度，$h_r$为有效的接收方天线高度，$f_{clutter}$为因地物所引起的平均加权损耗。$DiffractionLoss$为经过有障碍路径引起的衍射损耗，$K_{hill}$为山区的修正因子[10]。

资料表明，无线电基站以及接收者周围的地物对整个传播过程有着非常显著的影响。地物指的在地球表面的物体而不是实际的地形，其中就包括建筑植物等等。通常情况下离信号塔越近的地物对传播的影响也就越大。对于林区以及城市区域，在地物之间的损耗主要由衍射损失主导[5]。所以我们引入下面公式来估算单刃衍射损失[7]。

$$J(v) = 6.9 + 20\,log\left(\sqrt{(v-0.1)^2+1} + v - 0.1\right) \tag{5.2}$$

我们认为$J(-0.78) \approx 0$，所以当$v \le -0.78$时，我们把$J(v)$设置为 0。

当信号塔和接收端附近的地形比较高时，还需要加入额外的损失A_h。如果信号端比周围地形高时，即$h \ge R$，那么A_h就为 0。否则的话，根据地物种类的不同，则需要计算不同的A_h。其中对于海洋等空旷的区域，计算公式如下：

$$A_h = -(21.8 + 6.2\,log(f))\,log\left(\frac{h}{R}\right) \tag{5.3}$$

对于城镇，树林这样比较密集的区域，计算公式如下：

$$A_h = J(v) - 6.03 \tag{5.4}$$

其中v为：

$$v = \sqrt{(R-h)\arctan\left(\frac{R-h}{w_s}\right)f} \tag{5.5}$$

其中f为频率(GHz)，w_s代表街道的宽度，默认为 27 米。

通过上述经验模型，结合在第四章中已介绍过的 Cost 231-Hata 模型，最终我们得到额外的特征如下：

表 5.1 基于经验模型生成的特征

索引	符号	索引	符号
2.1	$log_{10}(f_c)$	2.6	v
2.2	$log(h_t)$	2.7	$log\left(\sqrt{(v-0.1)^2+1}+v-0.1\right)$
2.3	$log_{10}(d)$	2.8	$log(\frac{h}{R})$
2.4	$log(h_b)\,log(d)$	2.9	$log_{10}\,h_m^2$
2.5	$log(h_r)$	2.10	$log_{10}\,h_m$

表中涉及到了多个对某项距离取对数函数的特征，当距离为 0 时，其对数值不存在。对比专家传统信道模型，不难发现模型往往是这些参数的线性组合，某项对数参数对应着某一项信号损耗值。很自然的，当距离为 0 时，该项信号损耗也将不存在，因此在本文特征工程中，若对数函数值不存在，则令其为 0。

5.2.2 基于数据集信息的信道特征设计

在进行基于数据集信息的特征时，我们延续前一问的思路，首先建立基于几何位置信息的特征。在前一问中，数据集信息具有具体的数值，可以直接进行作图分析；而在本问环境下，应考虑更为一般的情形，建立适用于所有数据集信息的几何特征模型。

为了将复杂问题简单化，我们将三维空间投影到二维空间进行处理，再获取三维空间的距离信息，示意图见下图 5.1，其中，图 5.1.a为三维直角坐标系下的示意图，图 5.1.b为投影到xoy平面的的示意图，图 5.1.c为投影到xoz投影的示意图。

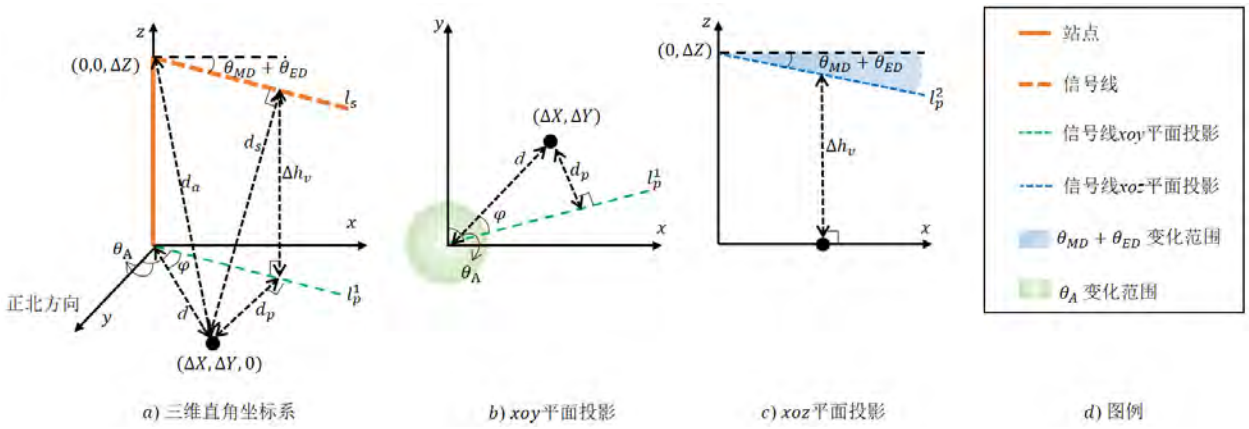

图 5.1 几何特征示意图

图中所有符号同上问，θ_{MD}表示小区发射机垂直机械下倾角，θ_{ED}表示小区发射机垂直电下倾角，θ_A表示小区发射机水平方向角，φ表示目标栅格到原点连线与信号线夹角，$(\Delta X, \Delta Y, 0)$表示目标栅格的坐标，$(0,0,\Delta Z)$表示天线坐标，Δh_v表示目标栅格与信号线的相对高度，d表示目标栅格与小区发射机栅格之间的距离，d_a表示目标栅格与天线的距离，d_s表示目标栅格与信号线的距离，d_p表示目标栅格与信号线在xoy平面投影直线的距离。

首先求取各点坐标与直线方程。

在该坐标系下，目标栅格坐标为：

$$(\Delta X, \Delta Y, 0) = (Cell\ \ X - X, Cell\ \ Y - Y, 0) \tag{5.6}$$

天线z轴坐标值为：

$$\Delta Z = Height + Cell\ \ Altitude - Altitude - Building\ \ Height \tag{5.7}$$

由图易见，对信号线投影线求方程后，各特征的建模就十分简单了。

信号线在xoy平面投影方程如下：

$$l_p^1 : Ax + By = 0 \tag{5.8}$$

其中，

$$\begin{cases} A = 1, B = 0 & \theta_A = 90^o, 270^o \\ A = 0, B = 1 & \theta_A = 0^o, 180^o \\ A = \dfrac{1}{\tan\theta_A}, B = 1 & \text{其它} \end{cases} \tag{5.9}$$

据此，就可以利用解析几何，对图 5.1.a 中的各项特征进行建模构造：

$$\begin{cases} d = \sqrt{\Delta X^2 + \Delta Y^2} \\ d_a = \sqrt{\Delta X^2 + \Delta Y^2 + \Delta Z^2} \\ d_p = \dfrac{|A\Delta X + B\Delta Y|}{\sqrt{A^2 + B^2}} \\ \Delta h_v = \Delta Z - |\Delta X| \tan(\theta_{ED} + \theta_{MD}) \\ d_s = \sqrt{(\Delta Z - \tan(\theta_{ED} + \theta_{MD})\sqrt{d^2 - d_p^2})^2 + d_p^2} \\ \varphi = \begin{cases} \arctan\left|\dfrac{A - \frac{\Delta Y}{\Delta X}}{1 + \frac{A\Delta Y}{\Delta X}}\right| & 1 + \dfrac{A\Delta Y}{\Delta X} \neq 0 \\ 90^o & 1 + \dfrac{A\Delta Y}{\Delta X} = 0 \end{cases} \end{cases} \tag{5.10}$$

整理汇总，获得几何位置特征表如下表 5.2：

表 5.2 基于几何位置的特征设计

索引	特征	值
2.11	ΔZ	$h_h^b + h_a^b - h_a - h_b$
2.12	d	$\sqrt{\Delta X^2 + \Delta Y^2}$
2.13	d_a	$\sqrt{\Delta X^2 + \Delta Y^2 + \Delta Z^2}$
2.14	d_p	$\dfrac{\lvert A\Delta X + B\Delta Y\rvert}{\sqrt{A^2 + B^2}}$, $\begin{cases} A = 1, B = 0 & \theta_A = 90^o, 270^o \\ A = 0, B = 1 & \theta_A = 0^o, 180^o \\ A = \dfrac{1}{\tan\theta_A}, B = 1 & \text{其它} \end{cases}$
2.15	d_s	$\sqrt{(\Delta Z - \tan(\theta_{ED} + \theta_{MD})\sqrt{d^2 - d_p^2})^2 + d_p^2}$
2.16	Δh_v	$\Delta h_v = \Delta Z - \lvert\Delta X\rvert \tan(\theta_{ED} + \theta_{MD})$
2.17	φ	$\varphi = \begin{cases} \arctan\lvert\dfrac{A - \Delta Y/\Delta X}{1 + A\Delta Y/\Delta X}\rvert & 1 + \dfrac{A\Delta Y}{\Delta X} \neq 0 \\ 90^o & 1 + \dfrac{A\Delta Y}{\Delta X} = 0 \end{cases}$

另一方面，类似于第一问，还可以根据数据集给出的发射机的工程信息、发射机栅格及目标栅格的地貌信息构造特征，罗列见下表 5.3：

表 5.3 基于数据集其它信息的特征

索引	符号	含义
2.18	f	小区发射机中心频率
2.19	P_t	小区发射机发射功率
2.20	C^b	小区发射机所在栅格的建筑物高度
2.21	C	目标栅格的地物类型
2.22	h_b^b	小区发射机所在栅格的建筑物高度

5.2.3 基于算法的生成特征

通过实践模型可以发现，预测值 RSRP 和一些非线性特征有着比较大的相关关系，所以在输入数据中增加一些原先特征的组合，可以很好地提高模型的性能。因此除去上述已有的特征外，我们通过对特征进行多项式组合得到更多的非线性特征，便于本文探索一些与结果有着明显关系的特征。例如对于特征(x_1, x_2, x_3)，设定结果的度为 2，其经过多项式组合后的结果为：

$$(x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3) \tag{5.11}$$

此外，常见的非线性特征还包括基于对数函数的特征变换：

$$(log(x_1), log(x_2), log(x_3) \tag{5.12}$$

和基于角度的特征变换：

$$(tan(x_1), sin(x_2), cos(x_3)) \tag{5.13}$$

我们通过对资料提供的基础特征进行上述三种变换，得到生成的额外非线性数据特征。排序不符合物理意义的组合后，基于算法生成的模型特征如下：

表 5.4：基于算法生成的特征

索引	符号	索引	符号
2.23	$C^b h_b^b$	2.27	$d_a d_s$
2.24	$log_{10}(d_a)$	2.28	$\sin(\varphi)$
2.25	$log_{10}(\Delta h_v)$	2.29	$\cos(\varphi)$
2.26	$log_{10}(d_s)$	2.30	$\cos(\varphi)\, d_s$

5.4 特征选择

通过上述步骤，我们获得了与任务相关的大量特征，选择合适的特征来进行训练对于最后的结果而言是至关重要的，所以如何去度量特征和预测值 RSRP 之间的相关性成为亟待解决的问题。

在实际度量过程中，由于单种度量方式存在诸多限制，所以我们需要通过多种度量方式来综合进行排序。

如果一个特征不够发散，即对应的方差趋近于 0，那么样本则在该特征上没有表示出差异性，对样本的区分没有作用，所以首先计算各个特征的方差信息：

$$\sigma^2 = \frac{\sum\big(X - E(X)\big)^2}{N} \tag{5.14}$$

其中$E(X)$为样本的期望。

此外，统计学中常用的衡量相关性的方式还有皮尔森相关性系数以及 Spearman 相关性系数。我们使用皮尔森相关性系数来计算变量的之间的相关程度，其定义如下：

$$\rho_{X,Y} = \frac{E(X,Y) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}} \tag{5.15}$$

然而皮尔森系数只能够估算两个变量之间的线性相关性，并且需要这些变量大致符合正态分布，所以引入 Spearman 相关性系数，其计算公式如下：

$$\rho_{X,Y} = 1 - \frac{6\sum_{i=1}^{N}(X_i' - Y_i')^2}{N(N^2 - 1)} \tag{5.16}$$

其中X_i'和Y_i'分别的X_i和Y_i在各自数组中的排序索引。

除了统计学种常用的特征相关性分析之外，目前采用机器学习技术得到特征的相关重要程度也逐渐成为了一种主流的方式。本文采用随机森林方式进行模型特征的重要性评估，核心思想是得到每个特征在决策树上的贡献值进行排序。对于每一颗决策树，都可以根据数据得到误差值$Error_1$。然后对样本的特征进行扰动，得到新的误差值$Error_2$。如果扰动之后的准确率大幅度下降，则说明该特征对于结果有着非常大的影响，重要程度就越高。所以我们可以通过以下公式衡量特征的重要性：

$$P = \frac{\sum(Error_1 - Error_2)}{N_{tree}} \tag{5.17}$$

得到这些重要程度之后，我们对上一小节构造的特征并上数据集原始特征作为特征集，进行整体的重要性程度排名，最后按四个排名的均值排序，取排前 10 列各项评分情况表如下，全表见附录：

表 5.4 前 10 位特征各项评分

符号	方差	Pearson	Spearman	随机森林	平均排名
d	1090.409	0.185	0.394	0.11	4.75
$\sin(\varphi)$	1109.382	0.175	0.332	0.141	6
$log_{10}(d)$	0.5	0.339	0.394	0.08	7.75
v	13.233	0.303	0.342	0.035	8.5
Δh_v	135.607	0.143	0.283	0.125	9.25
d_p	780.733	0.167	0.291	0.066	9.5
$log_{10}(d_a)$	0.491	0.337	0.394	0.042	9.75
d_a	1089.617	0.185	0.394	0.025	10
$log_{10}(d_a)$	1.379	0.339	0.393	0.031	10
$d_a d_s$	1.014	0.323	0.367	0.037	10.25

而后取平均排名的倒数作为最终得分进行排序，汇总如下表。

表 5.5 特征得分排序表

排序	特征	该特征与目标的相关性	排序	特征	该特征与目标的相关性
1	d	0.211	19	$log_{10} h_m^2$	0.051
2	$log_{10}(d)$	0.167	20	p_t	0.050
3	Δhv	0.129	21	h_a	0.048
4	v	0.118	22	$log_{10} h_m$	0.047
5	d_p	0.108	23	$log_{10}(h_b)$	0.047
6	d_a	0.105	24	$log_{10}(h_h^b)$	0.046
7	$log_{10}(h_b)\, log_{10}(d)$	0.103	25	$log_{10} h_a^b$	0.045
8	v_2	0.100	26	$log_{10} h_a$	0.043
9	d_s	0.100	27	φ	0.041
10	h_b	0.098	28	$log(f_c)$	0.040
11	C^b	0.078	29	$C^b h_b^b$	0.040
12	ΔZ	0.074	30	$log_{10}(d_a)$	0.040
13	p_t	0.070	31	$log_{10}(\Delta h_v)$	0.036
14	v_4	0.057	32	$log_{10}(d_s)$	0.035
15	h_b^a	0.057	33	$d_a d_s$	0.035
16	$log_{10}(h_b^b)$	0.056	34	$sin(\varphi)$	0.034
17	v_3	0.056	35	$cos(\varphi)$	0.032
18	C	0.052	36	$cos(\varphi)\, d_s$	0.029

其中，v为$\sqrt{f \times (R-h) \times arctan(R-h)/d}$其中 R为接收点高度，h为基站高度。v_2为$log_{10}\left(\sqrt{(v-0.1)^2+1}\ +v-0.1\right)$。$v_3$为$-log_{10}(\frac{h}{R})$。$v_4$为$v_3 \times \log_{10}(f)$。剩余变量可以在前文找到定义。

5.3 小结

本章首先详细地分析了特征选择的重要性，输入强相关的特征会对模型起到积极的作用，输入不相关的特征则会使得模型难以学习真正的映射。相较于问题一中的单一数据集和 Cost 231-Hata 模型，我们进一步分析了 Stand Propagation 模型，并通过查阅文献得到了衍射损失等强相关特征，同时将第一问的几何特征推广，使其能够适应各种位置下的几何特征计算，最后本文还通过尽可能枚举非线性组合的方式去寻找有效特征。在得到大量特征之后，本文分别从统计学衡量指标，如方差，Pearson 相关性系数，Spearman 相关性系数来度量特征与目标值的相关性，同时从机器学习的方法，如随机森林，去衡量各个特征的重要程度。最后综合各种指标的度量，对特征进行排序完成特征筛选。

六、问题三的模型建立与求解

6.1 问题分析

通过上述的特征提取与筛选，我们得到了一些与目标值 RSRP 相关的有效特征。根据经验公式等信息可以发现，这些特征大多数都和目标值之间存在着线性关系。目前，随着神经网络的快速发展和计算力的显著提高，深度学习方法在分类问题和回归问题上都取得了显著的效果。相较于传统的机器学习算法，神经网络强大的拟合能力能够学习出非常复杂的映射关系，所以对于无线传播这一复杂场景的预测问题，本文通过训练多层神经网络的方式来尽可能地拟合特征与 RSRP 的映射关系。

6.2 模型的建立

基于 RSRP 和特征为线性关系的想法，本文首先构建了一个三层的神经网络，其结构示意图如下：

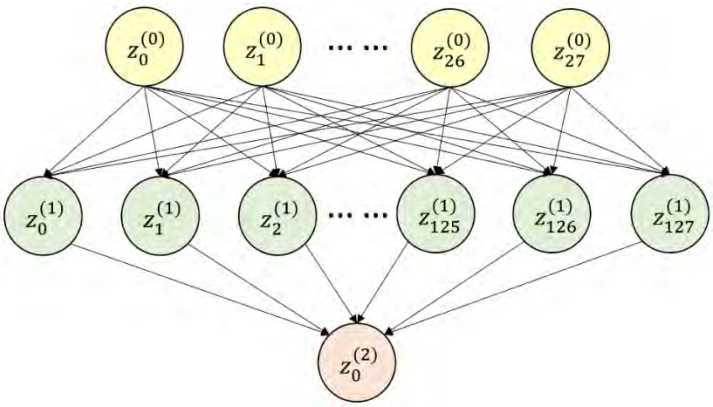

图 6.1 三层神经网络示意图

其中输入层$Z^{(0)}$的维度为d，即输入特征的个数。中间层的维度为固定的 128 层，最后再映射到维度为 1 的$Z^{(2)}$，即特征向量相应的 RSRP 预测值$\hat{p}$。然后计算 RSRP 预测值和真实值p之间的平方损失误差作为模型的损失函数：

$$L_{RMSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{p}^{(i)} - p^{(i)}\right)^2 \tag{6.1}$$

然而在实际的无线传播过程中，整体的环境因素过于复杂，各种因素组合情况千变万化，很难通过有限的指数、多项式等人为组合考虑到所有可能情况。然而往往这种非线性组合对于最后的 RSRP 预测值有着重要的影响。因此需要在网络中拟合这些非线性组合用来提高模型的性能。由于组合的方式无法事先知道，本文通过在网络增加非线性层，根据损失函数采用梯度下降的方式自动学习非线性组合。

此外，可以发现验证集的损失值和训练集的损失值基本持平，这说明网络的拟合程度不够，没有能够很好的拟合整个训练集。为了解决该问题，一种常见的做法就是增加模型的参数量，提高模型的复杂度。因此，我们增加了模型的层数和每层节点的个数以此来提高模型的拟合能力。

为了提高拟合非线性组合的能力，我们加入了 leaklyRelu 层以及 BatchNormalization 层。其中 leaklyRelu 能够对中间值进行非线性变换，对于输入x，其输出$f(x)$定义如下：

$$f(x) = \max(0.01x, x) \tag{6.2}$$

通过多层 leaklyRelu 的叠加，就可以很好地拟合非线性组合。此外在实际训练过程中，随着网络层数的增加，通常会遇到梯度消失的情况，这是因为在训练过程中输入值的分布逐渐朝着非线性函数的两端靠近，导致梯度变小网络收敛越来越慢，为了解决这个问题，我们在模型中添加了 BatchNormalization 层，将经过非线性映射后的值x拉回到标准正态分布：

$$x = \frac{x - E(x)}{\sqrt{Var(x)}} \tag{6.3}$$

综上，我们的模型结构从三层全连接网络扩展到五层全连接网络，并在每次全连接后加入非线性激活函数层以及批正则化层。

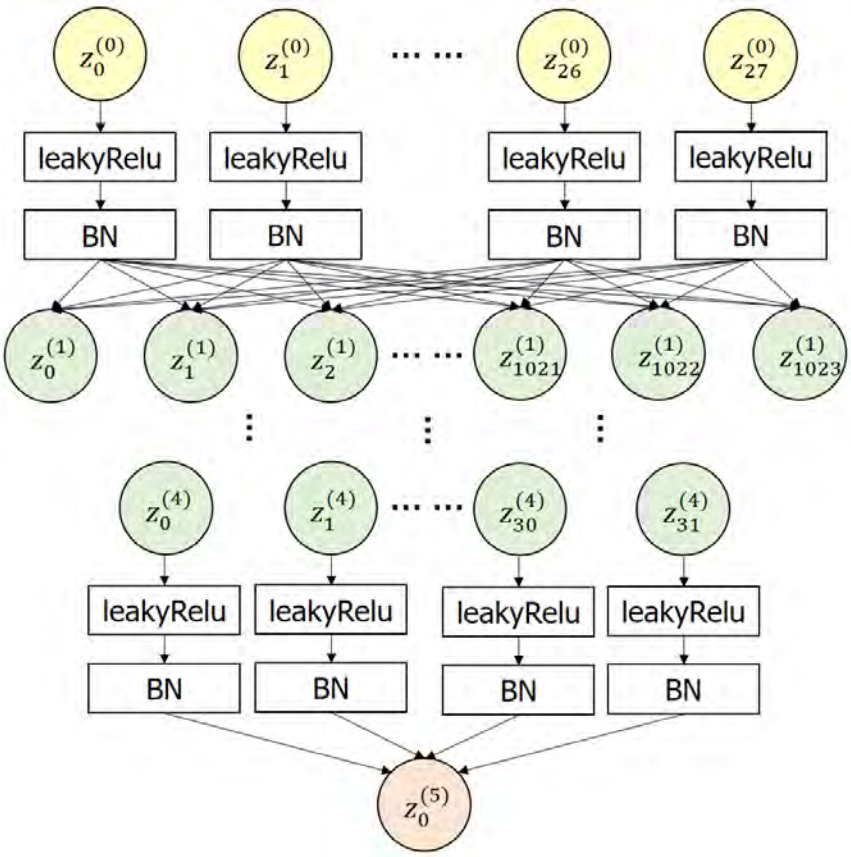

图 6.2 四层全连接神经网络示意图

6.3 模型的求解

在实际训练过程中，我们采用 Adam 优化器进行梯度的更新。相比于传统的随机梯度下降，Adam 通过计算梯度的矩阵来为各个参数设置独立的学习速率，对变化频繁的参数采用大的步长进行学习，对于稀疏的参数则采用小步长进行学习。其第 k 步梯度g_k更新如下：

$$g_{k+1} = g_k - \alpha \frac{g_k}{\sqrt{\sum_{i=1}^{k} g_k^2}} \tag{6.4}$$

同时将整个数据集划分成为 80%的训练集以及 20%的验证集，通过在验证集上运行模型来观察实际的泛化能力。

对于三层全连接模型，其损失函数变化如下：

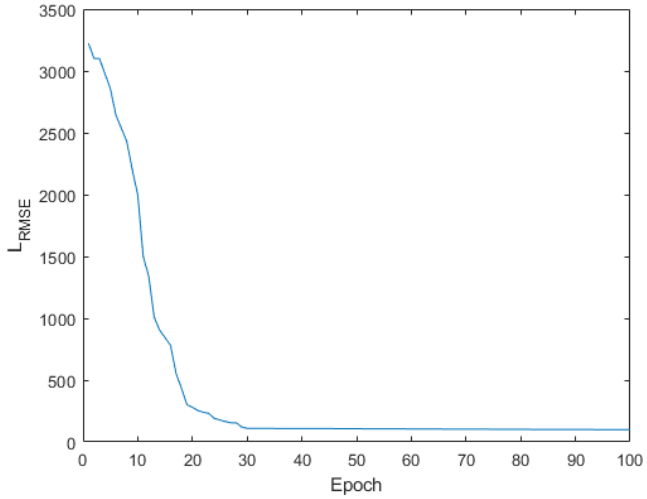

图 6.3 三层全链接神经网络损失函数变化

通过损失函数可以发现，在验证集上的损失值逐渐降低最终趋于稳定，最终在 99 左右维持稳定。这说明我们的模型有着良好的泛化性能。然而在实验中发现在训练集上的损失值大致和验证集上的损失值持平，并且保持较高的数值。这说明模型没有足够好地拟合训练集。为此，我们训练更深的非线性网络，其损失函数如下所示：

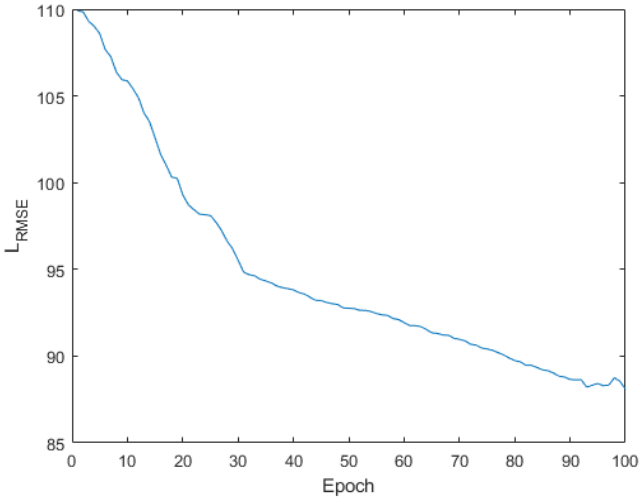

图 6.4 五层全链接神经网络损失函数变化

相比于之前简单的三层全连接网络，可以看到该网络在验证集上有着更好的表现，这说明该网络更好地拟合出了预测 RSRP 所需要的非线性特征。此外该模型的初始化损失值要比之前的模型显著降低，这是因为参数量的增加导致模型的拟合能力提高。

6.3 模型的效果分析

为了能够更好地理解模型预测 RSRP 效果的好坏，本文结合位置信息将预测得到 RSRP 值与真实的 RSRP 值进行可视化对比，如下图 6.5 所示：

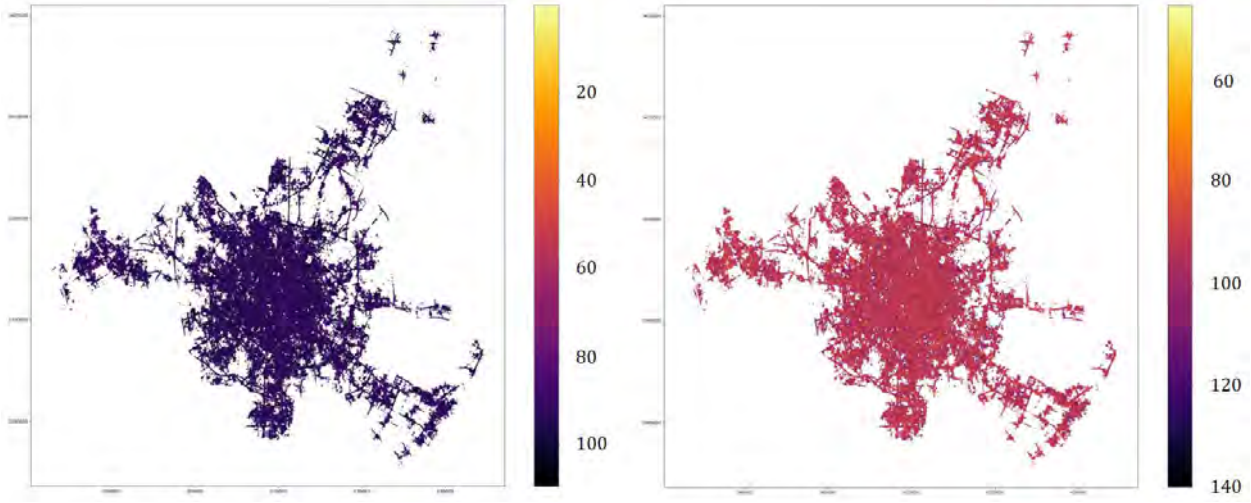

图 6.5 验证集预测值与真实值可视化对比

通过上图可以发现，虽然两图再使用同一色图的情况下，呈现出了不同的颜色，但是对比图边的色块，我们可以发现，两图的颜色变化呈现相似的趋势。然而两者的颜色没有保持相同，这是因为在真实值中存在一些过于极端的数据点，导致整体数值取值范围比预测的取值范围更大，导致作图时的颜色映射也不同，而该模型的预测值事实上和真实值大致落在相同区间。

为了消除极端数据点对可视化结果的影响，我们去掉区间两端 2.5‰的极端值，重新对预测值和真实值进行可是话操作，效果见下图 6.6，大图参见附件。

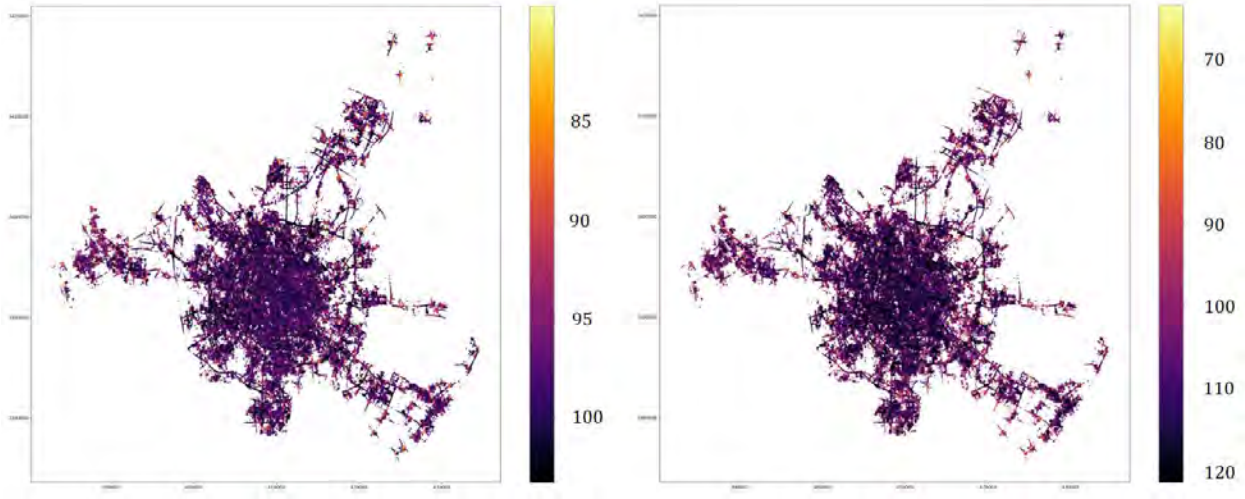

图 6.6 验证集预测值与真实值去端后可视化对比

从上图可以很明显地看出两者的分布基本相似，这进一步说明了本文提出的模型学习到了有意义的映射函数，能够很好地根据提供的基本数据，预测得到相应的 RSRP 值。

此外，本文将模型上传至华为云进行测试，在官方提供的测试集上也取得了非常不错的得分，这充分体现了本题提出模型的有效性和泛化能力。截至目前，该模型在线上获得了前 3%的成绩。

七、模型的评价与改进

本文针对无线智能传播这一问题进行了细致的分析，了解到由于环境的复杂性，导致精确计算接收区域的平均信号接收功率 RSRP 是不可行的。然而在实际情况中，RSRP 对于选择基站站址，选择基站覆盖区域等问题起到非常重要的作用。为此建立一个能够适应各种环境并且预测得到 RSRP 的模型成为亟待解决的问题。

首先基于提供的大量数据集，本文通过采用可视化的方式对大量数据进行了系统全面的分析。通过大量的图像信息，我们迅速地对各种数据的分布有了全局的了解，这对后续的特征选择有着很好的指导作用。此外，通过可视化的方式，能够明显看出某些数据不符合整体的数据分布，便于我们快速排查出不合理的异常点信息。

本文进一步分析了模型的好坏主要取决于选取特征的好坏。与预测值相关性强的特征往往能够带来更好的预测效果。为此我们主要从两个角度进行考虑，首先通过经验模型，即以往的专家经验，得到了一系列与 RSRP 存在线性关系的非线性变量组合，然后将这些非线性组合作为特征放入模型，从而提高模型的复杂性和拟合效果。此外，从通信和物理知识的角度下，本文也全面考虑了影响 RSRP 的物理因素，通过对目标栅格和发射台栅格的相对位置加以利用，得到基于空间几何结构的物理特征。为了能够对特征与预测值的关系进行更好的探索，我们还通过算法将基本信息进行非线性组合当作我们的候选特征。

基于以上方式得到许多候选特征后，本文通过利用统计学的相关性分析方式以及机器学习的分析方式，对每个特征按照重要性程度进行排序，最终筛选出与目标值最为相关的一些特征放入模型。最后我们将重要的特征集合放入到构建好的深度神经网络中进行训练，把 RSRP 的预测值和实际值的均方误差当作损失，通过梯度下降的方式更新模型的参数，从而拟合特征与 RSRP 的映射关系。实验结果表明，本文提出的模型学习到了与真实 RSRP 值相似的分布，并且能够快速收敛，在验证集和测试集上表现出了很好的泛化能力。

最后，本文列举该模型之后可能改进的方向：

1. 由于该数据集主要为小区数据，绝大部分地形为街道，楼层等。这种情况下信号传播的损失主要以衍射损失为主。为了能够使得模型能够使用更多的地形环境，同时考虑反射损失和绕射损失将会提高模型整体的泛化能力。
2. 在数据可视化的过程中，我们可以看到存在严重的数据不均衡问题，这使得模型可能会在数据量较小的情况下效果不佳。可以通过采样更多数据点的方式来缓解该问题，或者查阅更多的资料分析每种情况下信号损失的主要因素，根据以往的先验专家经验来处理该情况。
3. 通过实验可以看到损失值仍然在缓慢下降，因此采用更深层的深度网络并且训练足够长的时间可能会产生更好的效果。

参考文献

[1] Phillips C, Sicker D, Grunwald D. A survey of wireless path loss prediction and coverage mapping methods[J]. IEEE Communications Surveys & Tutorials, 2012, 15(1): 255-270.

[2] Durgin G, Rappaport T S, Xu H. Measurements and models for radio path loss and penetration loss in and around homes and trees at 5.85 GHz[J]. IEEE Transactions on communications, 1998, 46(11): 1484-1496.

[3] Dalela C, Prasad M, Dalela P K. Tuning of COST-231 Hata model for radio wave propagation predictions[J]. Computer Science and Information Technology (CS & IT), DOI, 2012, 10: 255-267.

[4] Singh Y. Comparison of okumura, hata and cost-231 models on the basis of path loss and signal strength[J]. International journal of computer applications, 2012, 59(11).

[5] Peh, Beng Yeow. Characterisation of Clutter Loss Using Fish-eye Lens for the Prediction of DTV Coverage. Diss. Nanyang Technological University, School of Electrical and Electronic Engineering, 2000.

[6] Rappaport, Theodore S. Wireless communications: principles and practice. Vol. 2. New Jersey: prentice hall PTR, 1996.

[7] Rogers, Steven R. "Diffusion analysis of track loss in clutter." IEEE Transactions on Aerospace and Electronic Systems 27.2 (1991): 380-387.

[8] Hata, Masaharu. "Empirical formula for propagation loss in land mobile radio services." *IEEE transactions on Vehicular Technology* 29.3 (1980): 317-325.

[9] Zhang, Yue Ping, Guo Xin Zheng, and J. H. Sheng. "Radio propagation at 900 MHz in underground coal mines." *IEEE transactions on antennas and propagation* 49.5 (2001): 757-762.

[10] Almers, Peter, et al. "Survey of channel and radio propagation models for wireless MIMO systems." *EURASIP Journal on Wireless Communications and Networking* 2007.1 (2007): 019070.

附录

符号	方差	Person	Spearman	随机森林	排名
d	1090.409	0.185	0.394	0.11	4.75
$log(d)$	0.5	0.339	0.394	0.08	7.75
$\_hv$	135.607	0.143	0.283	0.125	9.25
v	13.233	0.303	0.342	0.035	8.5
dp	780.733	0.167	0.291	0.066	9.5
da	1089.617	0.185	0.394	0.025	10
$log(hb)log(d)$	1.379	0.339	0.393	0.031	10
v_2	0.337	0.268	0.342	0.035	13.5
ds	787.671	0.17	0.001	0.077	14.25
$Building\_Height$	14.499	0.045	0.048	0.009	17.75
$Cell\_Clutter\_Index$	3.245	0.018	0.018	0.035	20
$\_Z$	18.031	0.038	0.041	0.014	18
pt	2.514	0.014	0.019	0.043	19.25
v_4	0.033	0.039	0.04	0.03	19.75
$Cell\_Altitude$	11.034	0.008	0.002	0.03	21.5
$log\ \_Building\_Height$	0.475	0.077	0.048	0.008	21.5
v_3	0.01	0.039	0.04	0.028	21.75
$Clutter\_Index$	3.042	0.029	0.006	0.021	22
$log10\_hm\_2$	0.082	0.034	0.036	0.015	23.5
pt	9.515	0.002	0.003	0.027	24.75
$Altitude$	10.997	0.007	0.002	0.016	24.5
$log10\_hm$	0.015	0.034	0.036	0.015	24.75
$log(hb)$	0.012	0.004	0.003	0.048	25
$log\ Height$	0.31	0.003	0.003	0.027	27.5
$log\ \_Cell\_Altitude$	0.01	0.007	0.002	0.029	28.5
$log\ \_Altitude$	0.01	0.006	0.002	0.016	31.25
fa	1.768	0.008	0	0	29.75
$log(Frequency\_Band)$	0.001	0.006	0.001	0.004	34.5
$C^b h_b^b$	14428.27	0.019	0.022	0.137	12.75
$\log(d_a)$	0.491	0.337	0.394	0.042	9.75
$\log(\Delta h_v)$	0.626	0.265	0.294	0.012	17.5
$\log(d_s)$	0.537	0.299	0.332	0.006	17.5
$d_a d_s$	1.014	0.323	0.367	0.037	10.25

$\sin(\varphi)$	1109.382	0.175	0.332	0.141	6
$\cos(\varphi)$	9.388	0.021	0.031	0.022	20.75
$\cos(\varphi)\, d_s$	0.007	0.029	0.031	0.003	28.5