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	北京师范大学
参赛队号	**19100270008**
队员姓名	**1.** 赵子鸣
	2. 王妍
	3. 李本继

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　基于神经网络的无线电波传播损耗预测模型

摘　　要：

2019 年 6 月，工信部正式向中国电信、中国移动、中国联通、中国广电发放 5G 商用牌照，标志着中国的 5G 商用元年正式到来，无线电波通信技术将在未来几年迎来前所未有的高速发展。为更好地满足用户不断增长与丰富的的市场需求，运营商需要部署大量基站来提高无线电波传播质量，而准确、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。

本文以观测点的平均信号接收功率（RSRP）的 PCRR 及 RMSE 指标为优化目标，旨在准确预测观测点的 RSRP 值。针对这一预测类优化问题，本文从模型分析和数据分析的角度对特征进行提取，并基于这些特征先后建立两个神经网络网络。

问题一本质上是模型分析任务。首先，运用文献综述法，针对关无线电波传播的 36 个模型的特点及适用范围进行汇总分析，学习其特征及特征引入方式；其次，重点研究了 Okumura-Hata 模型、Cost 231-Hata 模型、SPM 模型，对变量公式及各项参数的选择条件进行剖析，分析其作为模型训练特征矫正项的可能性；最后，对比已有变量名与三种经典统计模型特征变量，选取具有共性且合理的特征（**链路距离、基站发射机绝对高度、观测点绝对高度、载波频率、场景纠正**）作为我们模型的备选特征。

问题二本质上是数据分析任务。首先，对题目提供的 12011833 条数据的数据集进行数据清洗与可视化处理；其次，通过重复随机抽样、取平均值的方式，对 16 个变量进行因子分析，得到 6 个解释性极佳的因子，并基于因子分析结果对对变量进行组合与转化；最后，基于已有变量组合和无线电波的传播特性，提取数据中有效统计特征（**波频影响因子、自由空间的距离损耗、观测点密度损耗、建筑物密度损耗、平均建筑物高度影响因子、平**

均地形影响因子、直射因子、反射因子、散射因子、绕射因子、经典统计模型模型损耗校正项），并通过 Pearson 相关系数、Kendall 相关系数、Spearman 相关系数测试其余 RSRP 值的相关性，经过显著性筛选最终得到 19 个有效特征。

问题三本质上是工程实践任务。首先，本文在线下通过单 GBDT 模型对上述特征尝试进行了训练与线下指标评估，取得了 9.1276 的线下 RMSE。其次，由于传统 GBDT 模型并不能很好拟合出本问题的目标。本文用 GBDT 的预测输出与经验特征、原始数据特征、统计特征进行了融合，作为多层神经网络的输入，进行融合模型的训练，它会把 PCRR 和 RMSE 两个指标同时作为优化目标，进行多目标学习，整体预测无线电波信号强度。最终，本文建立的基于 AI 的无线电波信号强度预测模型，**在线下取得了 22.5834% 的 PCRR 值和 8.5807 的 RMSE**，并在华为云 ModelArts 平台也取得了优异的成绩。

在研究过程中，我们还发现，传统经验模型的特征对模型帮助不大，原始数据中的位置特征和统计特征对模型帮助很大，而特征中变量的最优组合方式仍有待深入研究。

关键字：无线智能传播　特征工程　多目标融合模型　神经网络　TensorFlow

目录

1. 问题重述

1.1 问题背景

图 1　5G 时代的到来

2019 年 6 月，工信部正式向中国电信、中国移动、中国联通、中国广电发放 5G 商用牌照。随着中国 5G 商用元年的到来与市场前景的扩大，需要建立大量的信号传播基站。据工信部透露，中国在 2019 年将新建 5 万个 5G 基站，2020 年中国将新建达到 68 万个[1]。由于在实际生产实践中，基站具有相当高的经济成本，因此，对于无线电基站的选址至关重要，如何合理的安排基站位置使得该基站的信号范围覆盖更广、信号强度更高、经济成本更低、性价比更高具有重要的研究价值。然而，研究基站选址问题的基础是要在不同的观测点对无线电波信号进行准确的预测，这对于通信技术的发展，乃至 5G 网络的科学部署有着非常重要的意义[2]。

1.2 问题描述

在本次数学建模赛题本质上是一个学习优化问题，要求建立无线电波信号值的预测模型，任务主要包括模型分析、数据清洗、数据处理、特征选择、特征组合、特征提取、特征相关性分析、模型建立和模型比较与测试。在各个任务中，特征选择和模型建立的关键目标均为参考信号接收功率 RSRP 值，在基站发射机发射功率已知的条件下，本质上是对无线电波信号传播损失进行特征提取与预测。由于数据已经给出了各个观测点的 RSRP 实际值，该问题为有监督学习问题。

对于本题目任务的理解如图2所示，基站发射机在特定的地理位置的一定的高度下，以一定的功率，主要向特定水平和铅垂方向，发出特定频率的无线电波信号，在不同的地形及建筑物高度环境下，通过直射、透射、绕射、散射、反射、折射等传播方式，传向一定区域范围所有信号接收点。本题目给出的训练数据就是从所有信号可接收点钟随机选取得到的[3]。

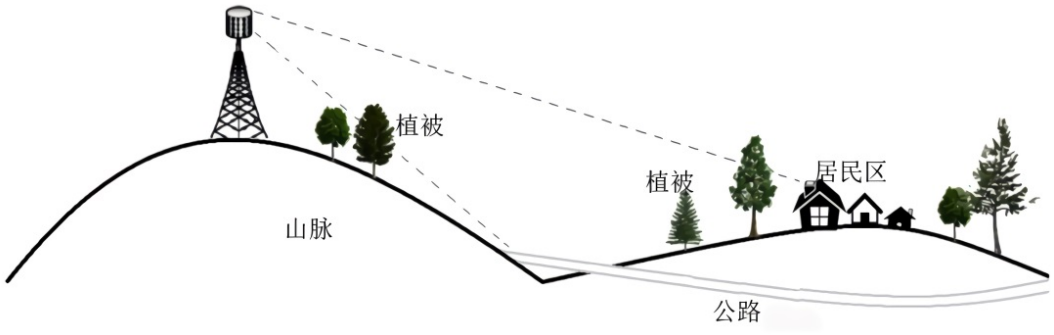

图 2　任务示意图

在问题一中，要求根据 Cost 231-Hata 模型以及题目所提供的数据信息选择合适的无线电波传播特征，并分析其合理性和现实意义。为了保证模型特征提取的多样性与可靠性，本文在选取 Cost 231-Hata 模型特征的基础上，对大量已有文献中的模型进行分析比对，并重点对 Okumura-Hata 模型与 Standard Propagation Model（SPM）模型的进行选择。

在问题二中，要求在模型分析的基础上，从题目提供的包含 18 个变量、12011833 个观测点数据的数据集中进行特征提取，在对特征进行合理量化、排序的基础上计算这些特征与目标的相关性，并给出特征的数值计算方法与现实意义。

在问题三中，要求基于已经选择的具有现实意义的、与目标相关性较强的有效合理的特征，通过多种方法建立基于 AI 的无线电波信号强度预测模型，即对不同地理位置的 RSRP 值进行准确预测，其优化目标包括将 PCRR 值（综合考虑准确率和召回率目标）控制在 20% 以上与最小化 RMSE 值（均方根误差）。

2. 符号与变量说明

2.1 符号说明

符号	原数据变量	中文含义	单位
Id	Cell Index	小区编号	
X_C	Cell X	基站坐标	
Y_C	Cell Y	基站坐标	
H_M	Height	基站发射机相对地面的高度	米
A_C	Cell Altitude	基站海拔高度	米
H_C	Cell Building Height	基站所在格点的建筑物平均高度	米
CI_C	Cell Clutter Index	基站地物类型索引	
α	Azimuth	基站发射机水平方向角	度
θ_1	Electrical Downtilt	基站发射机垂直电下倾角	度
θ_2	Mechanical Downtilt	基站发射机垂直机械下倾角	度
f	Frequency Band	基站发射机中心频率	Mhz
Pr	RS Power	基站发射机发射功率	dBm
X	X	观测点坐标	
Y	Y	观测点坐标	
A	Altitude	观测点海拔高度	米
H	Building Height	观测点所在格点的建筑物平均高度	米
CI	Clutter Index	观测点地物类型索引	
R	RSRP	平均信号接收功率	dBm

注：这里只列出论文各主要通用符号，个别模型单独使用的符号在首次使用时会进行说明。

符号	新生成特征含义	单位
Y_f	波频影响因子	
N_i	第 i 个基站的所有观测点的数目	
L_D	自由空间的距离损耗	
L_N	观测点密度损耗	
L_B	建筑物密度损耗	
$\overline{W_H}$	同一个基站的观测点，H 的平均值	
$\overline{W_{CI}}$	同一个基站的观测点 CI 的平均值	
Pr^*	发射功率修正项	dBm
N_i	每个基站的观测点数量	
M_i	同一个基站的观测点的 HC 不为 0 的数量	
φ	水平方向，基站观测点连线与正北方向的夹角	度
ϕ	垂直方向，基站观测点连线与水平方向的夹角	度
λ_{11}	完全直射因子	
λ_{12}	相对直射因子	
D	距离中心直射线的最短距离	米
$\lambda_{1,2,3}$	反射、散射、绕射因子	
L_{ok}	Okumura-Hata 模型矫正项	
L_{cost}	Cost 231-Hata 损耗矫正项	
L_{SPM}	SPM 损耗矫正项	

2.2 变量解释

在本题的任务中，最核心的指标，也是模型最终要预测的指标为 RSRP 值。因此，有必要对平均信号接收功率 RSRP 值和电波信号传播损耗 L 及其单位进行说明。

平均信号接收功率（RSRP），是无线信号强度（RSSI）的关键参数，由于无线信号一般为 mW 级别，不便于计算与表示，因此将其转换为 dBm 表示。dbm 是一个表示功率绝

对值的单位，他的计算公式为

$$0\mathrm{d}bm = 10\lg(1\mathrm{mW}) \tag{1}$$

即 1mW = 0dBm，小于 1mW 无线信号就是 dBm 表示为负数。在实际传输过程中，信号接收方是很难达到接收功率 1mw 的，因此 RSRP 常常以负值的 dBm 数出现。

一般而言，dbm 值越大，信号强度越高，接收效果越好，但考虑到实际应用中的经济成本，当一个区域接收到的 dbm 值介于 0 到-50dbm 之间，或者介于 0 到-70dbm 之间时，认为该区域信号值良好。当接收到的无线信号小于-70dbm 则会出现传输不稳定，速度缓慢的现象，此时无线网络就无法正常使用。但在本次比赛中，弱覆盖判决门限 P 的值定为-103 dBm，即当一个区域接收到的 dbm 值介于 0 到-103dbm 之间时，认为该区域信号值良好[4]。

特别需要注意的是，电波信号传播损耗值为发射功率与平均信号接收功率的差值，即为 S =Pr-R，其单位为 dB 或无单位，是一个表征相对值的值，纯粹的比值，只表示两个量的相对大小关系[5]。

3. 模型假设

- 假设观测点相对独立，RSRP 测量值不受到文件中所给发射机以外的发射机的影响；
- 假设 RSRP 测量值在观测时，不会受到偶然因素的干扰，例如：飞机飞过遮挡信号，考场考试屏蔽周围信息、雷电天气等；
- 假设同一观测点测量接收到的两个发射机发出信号的 RSRP 值不会叠加，且不相互干扰。
- 假设每个发射机除了发射功率（RS Power）和中心频率（Frequency Band) 外，其余设备自身参数，如 CPU 型号、内存大小等，均一致；
- 假设基站发射机发射角度固定，定向发射信号；
- 假设基站发射机在水平和竖直方向具以面的形式发射信号，形成圆锥体发射；
- 由于房屋结构信息无法量化，假设数据集中的观测点以及站点发射机都不考虑在室内的情况；
- 在考虑地面反射时，假设反射点的海拔高度与基站高度一致；
- 在考虑反射、散射、折射、透射时，假设同一基站区域内，地形和高度一致，可用区域内平均值代替。

4. 模型的建立与求解

基于上述符号和假设，后文将按照如图3所示流程建立数学模型。

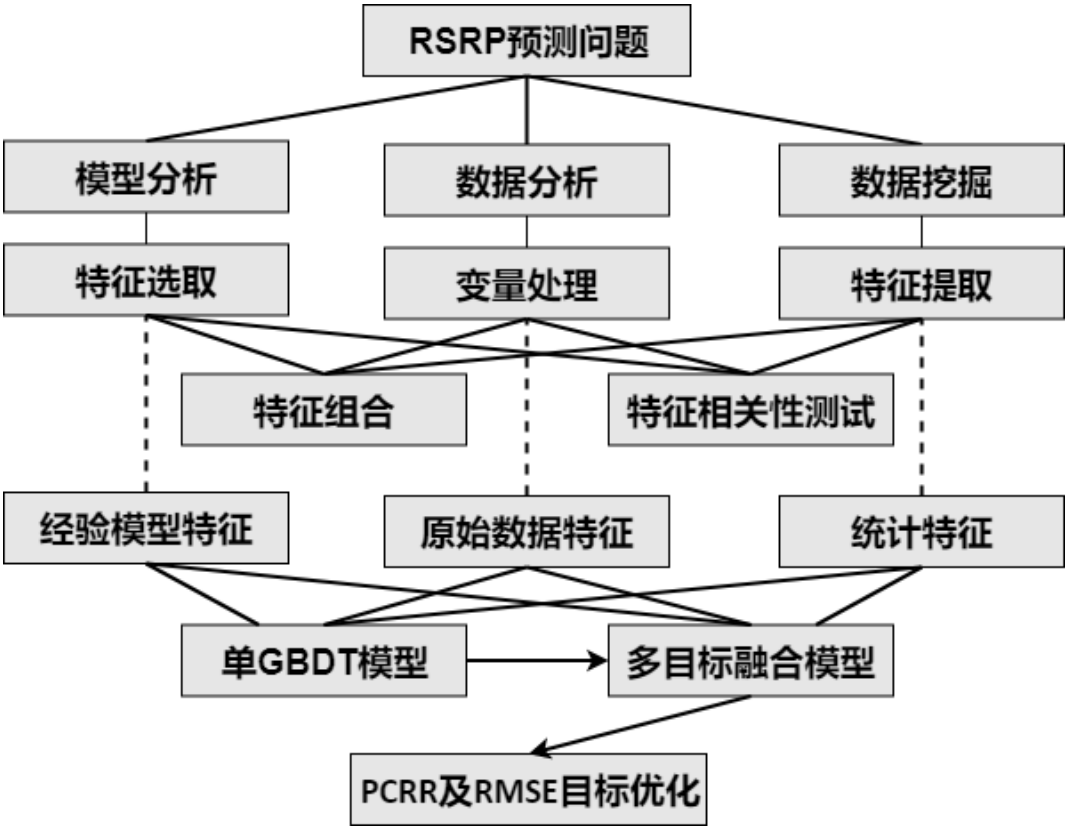

图 3　研究思路图

4.1 问题一建模与求解

问题一属于基于模型分析与特征选取问题，对于解决此类问题文献综述法和模型比较的方法。

针对问题一，首先，通过阅读大量文献，对现在已有的模型进行分析与比较，主要针对每一种模型的特点及适用范围进行汇总，找出各种模型使用的共有特征及其作用；其次，本文重点分析了 Okumura-Hata 模型、Cost 231-Hata 模型、Standard Propagation Model（SPM）模型，并对其公式及各项参数与已有变量名进行详细比对，选取适合本任务的有效特征，基于已有变量给出计算公式，并阐述其合理性与现实意义。

4.1.1 常见无线智能传播模型

近年来，由于通讯技术的快速发展，无线电波应用越来越广泛，出现了大量研究无线电波传播模式的文献与模型，根据辐射源覆盖范围的的不同，可分为大区模型、小区模型、微区模型；根据传播模式性质的不同，可分为：经验模型、半经验模型、确定性模型；根

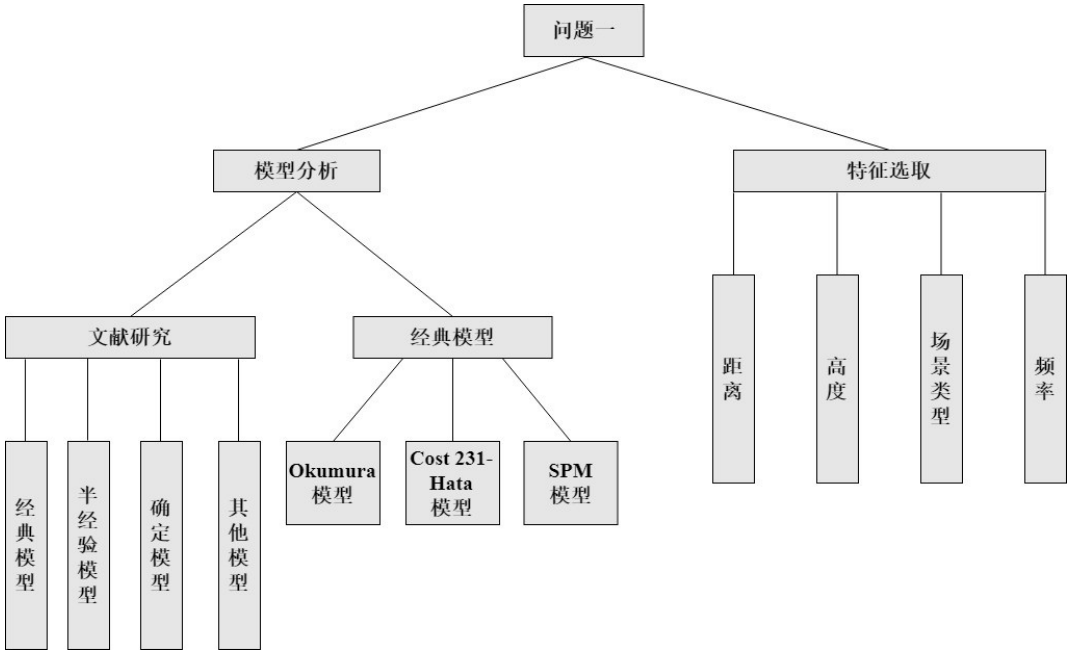

图 4　问题一思路图

据分析电波传播方向的不同，可分为正演模型、反演模型。[6] 对于常见文献中的模型，本文对其核心的模型思想、模型特点及适用范围进行提炼得到表1。

通过阅读大量文献及表可知[7]，最早的无线电波传播模型均为基于统计规律的经验模型或半经验模型，而现代技术的发展促进了确定性模型的产生于发展。虽然确定性模型预测较为准确，但由于其较高的算法复杂度及经济成本，使用并不便利。因此，本文希望通过现在新兴的机器学习与深度学习的方式，建立现代深度学习预测模型。

首先，科学研究需要站在巨人的肩膀上，因此，本文对已有的三种经典统计模型（Okumura-Hata 模型、Cost 231-Hata 模型、SPM 模型）进行详细分析，并希望通过其预测结果数值，指导神经网络学习进程。

- **1.Okumura-Hata 模型**

Okumura-Hata 模型是基于 Okumura 模型进行改进的模型[8]。首先，该模型对复杂的地形地物环境进行了一个系统的划分。其预测方法是把“准平滑地形”作为分析和描述传播模型的基准。准平滑地形指表面起伏高度小于 20m 的地形，也称为中等起伏地形，其起伏缓慢。对应地，不规则地形指准平滑地形以外的地形，包括但不限于连绵起伏山地、丘陵地带、坡形地形、海岸线等地形。对于准平滑地形，按照地物的密集程度可以分为三类：开阔地、郊区环境、市区环境。

Okumura-Hata 模型可以通过修改参数来适应其他地形地区，例如：可以通过修改参数

表 1　移动通信中电波传播预测模型的改进研究

类别	模型		模型思想、模型特点	适用范围
经验模型 (统计模型）	Okumura 模型		“准平坦地形”作为分析和描述传播特性的基准	市区、郊区、开阔区
	Hata 模型		Okumura 预测方法的公式表达	
	Okumura-Hata 模型		Okumura 模型可以通过修改参数使之适应其他地形地区	城市或者准城市地形
	Cost 231-Hata 模型		对较高频段的 Okumura 传播曲线进行改进	基站天线高度高于邻近的建筑物的屋顶高度
	CCIR 模型		在 Okumura-Hata 模型的基础上，增加了地形对信号传播损耗的影响	高密度城区
	Lee 模型		刃形绕射	市区、郊区
	Ibrahim&parsons 模型		引入建筑物密度、建筑物高度及地面情况的关系特征	平坦地面
	McGeehan Griffiths 模型			
	Atifi&parsons 模型			
半经验模型	Ikegami 模型		入射射线和水平射线所形成的夹角很大时很准确	均匀市区环境
	Walfish&Bertoni 模型		绕射屏模式	均匀市区和郊区
	Xia&Bertoni 模型		Walfisch—Bcrtoni 模式的改进	均匀市区和郊区
	Cost 231-Walfish-Ikegami 模型			
	SUI 模型		适合更小的蜂窝，更低的基站高度，更高的频率（可扩展到 3.3GHz），以及更丰富的传播场景	
	SPM 模型（室外蜂窝网络传播模型）		该模型因为增加了地貌相关因子、山区校正因子、衍射损耗因子等，被广泛应用于网规网优工具中	
确定性预测模型 （理论模型）	射线跟踪模型	SBR／镜像法	简单环境	预测准确 算法复杂度高
		双射线模式	只考虑直达射线和地面反射射线	
		多射线模式	四射线模式由直达射线、地面反射射线和两条被建筑物墙壁反射一次的射线	
		几何绕射理论（GTD）	绕射	
		一致性绕射理论（UTD）	绕射	
		物理光学（PO）	绕射	
		Volcano 模型		
	FDTD 模型			
	矢量抛物线方程模型			
	快速远场近似模型			
	多隙缝波导模型		该模式可以解释该模式可以解释实验中在市区条件下没有观察到电波拐点	低于屋顶平面的发射天线
	Lund 大学模型			
	神经网络模型			
其他模型	对数距离路径损耗模式		平均路径损耗是距离的 n 次幂的函数	蜂窝小区环境
	衰减因子模式		预测同一楼层或通过不同楼层的传播路径损耗	
	WCDMA 无线网络规划模型		反演模型	
	估计接收信号的特征方法	MERS	依赖于波到达方位角的概率密度分布(APD)和辐射类型	
	Kurner, T.，Meier, A.(2002)		考虑了在密集城市区域中所有相关传播现象（植被影响、多路径传播）	密集城市区域 1.8Ghz
	垂直平面模型	VPM		
	多路径模型	MPM		
	植被模型	VegMod		

将其频率范围扩大到 3000MHz。修改得到的模型为 Okumura-Hata 模型，它适用于①发射机发射的信号频率范围为 150-3000MHz；②信号的传播距离范围为 1-100KM；③发射机的天线高度范围为 30-1000m；④城市或者准城市地形。其公式为：

$$\mathrm{Lb} = 69.55 + 26.16\lg f - 13.82\lg(h_b) - \alpha(h_m) + (44.9 - 6.55\lg(h_b))\lg d \tag{2}$$

其中，f：工作频率（MHz）；h_b：基站天线有效高度（m）；h_m：移动台天线有效高度（m）；d：移动台与基站之间的距离（km）；$alpha(h_m)$：移动台天线高度因子。在不同的地形地区，可以通过调整 $alph$、K 使公式的预测效果更好。

对于大城市

$$\begin{aligned} \alpha(h_m) &= 8.29[\lg(1.54h_m)]^2 - 1.1dB \quad (f \le 300MHz) \\ \alpha(h_m) &= 3.2[\lg(11.75h_m)]^2 - 4.97\mathrm{dB} \quad (f \ge 300MHz) \end{aligned} \tag{3}$$

当 h_m 在 1.5-4m 之间，上面两式基本一致。

对于中小城市（除大城市以外的其它所有城市）

$$\alpha(h_m) = (1.1\lg f - 0.7)h_m - (1.56\lg f - 0.8) \tag{4}$$

对于郊区

$$\mathrm{Lb}_s = \mathrm{Lb}_{\text{市区}} - 2[\lg(f/28)]^2 - 5.4 \tag{5}$$

对于开阔地

$$\mathrm{Lb}_g = \mathrm{Lb}_{\text{市区}} - 4.78(\lg f)^2 + 18.33\lg f - 40.94 \tag{6}$$

- **2.Cost 231-Hata 模型**

由于 3G、4G 时代所使用的频段、覆盖的地区等因素与以往不同，因此无法使用 Okumura-Hata 模型进行信号覆盖预测。鉴于此，欧洲研究委员会 COST 231 传播模型小组建议，将 Okumura-Hata 模型作为基础，引入多个修正参数，使得其支持的频率扩展到 2GHz，因此得到了新的 Cost 231-Hata 模型。其定义如下：

$$\mathrm{PL} = 46.3 + 33.9\lg f - 13.82\lg(h_b) - \alpha(h_m) + (44.9 - 6.55\lg(h_m))\lg d + \mathrm{C}_m \tag{7}$$

其中 PL 定义为传播路径损耗（dB）、f 载波频率（MHz）、h_b 为基站天线有效高度（m）、h_m 为用户天线有效高度（m）、α 用户天线高度纠正项 (dB)、d 为链路距离（km）以及 C_m 场景纠正常数（dB）。其中 C_m 针对地区的一个修正项，在用于大城市中心路径的损耗测试中其应设置为 3dB，在其他区域时设置为 0dB。C_m 取值同上。

Cost 231-Hata 模型的适用范围如下：

- **3.Standard Propagation Model（SPM）模型**

表 2　Cost 231-Hata 模型的适用

限制条件	范围
频率（MHz）	1500-2300
链路距离（Km）	1-20
接收天线有效高度（Km）	1-10
发射机天线的有效高度（m）	30-200

SPM 模型即标准传播模型，是建立在 Cost 231-Hata 模型的基础之上的半经验模型，用于 150MHz-2000MHz 频率的无线电通信的传播损耗预测中。其适用范围较广，对环境方面没有限制，广泛适用于 3G 及 4G 通信系统在城市中的传播情况预测。该模型的基本公式如下：

$$\begin{aligned} L_1 = {} & K_1 + K_2 \lg d + K_3 \lg H_{\text{Txeff}} + K_4 \text{Diff\_}loss \\ & + K_5 \lg H_{\text{Txeff}} \lg d + K_6 H_{\text{Rxeff}} + K_{\text{clutter}} f(\text{clutter}) \end{aligned} \tag{8}$$

其中，K_1 是与频率相关的因子，单位为 dB；K_2 是距离衰减因子；d 是发射机和接收机之间的距离，单位为 m；K_3 是接收天线高度相关的因子；H_{Txeff} 是发射天线的有效高度；K_4 是与衍射相关的因子，其值必须为正；Diff_$loss$ 是衍射而引起的损耗，单位为 dB；K_5 是与发射天线有效高度和距离相关的乘性因子；K_6 是 H_{Rxeff} 的乘性因子；Rxeff 为移动台接收天线高度；K_{clutter} 是 f_{clutter} 的乘性因子，f_{clutter} 是由杂散引起的加权平均损耗。

从此模型的公式中可以看到，信号的路径损耗与基站发射频率、天线与基站的链路距离、接收天线的高度等相关，同时，为了更好的拟合实际数据，改模型还加入了电磁波的衍射项、地区地貌特征修正项，使其能够更好地进行预测。

综上所述，本文将以上三个经典模型的优点和缺点进行汇总如表4。

通常，按照城市地形区域：在建筑密集的大城市区域，可以优先选择 SPM 模型；小城市或城郊则优先选择 Cost231-Hata 模型；在较为开阔的区域，可以优先选择 Okumura-Hata 模型。按照频率要求和覆盖范围：频率较低、覆盖比较广的蜂窝网络优先选择 Okumura-Hata 模型、Cost231-Hata 模型或 SPM 模型进行优化修正；频率较高、覆盖面积适中的传播模型优先选择 Cost231-Hata 模型；频率高、覆盖面积小（如短距离通信）的情况优先选择 SPM 模型。特别地，如果地物类型单一且便于统计，使用 SPM 模型来添加地物损耗也是合理的选择。

表 3　经典统计模型优缺点汇总表

模型	优点	缺点
Okumura-Hata 模型	1.城市中能够简单精确预测	1.不同地形变化大，预测缓慢 2.不适用于覆盖半径大约 1km 的个人通信模式 3.不适用于 3G、4G 通信频段
Cost 231-Hata 模型	1.特别适用于宏蜂窝场景 2.适用于 3G、4G 通信频段	1.不适用于覆盖半径大约 1km 的个人通信模式
Standard Propagation Model（SPM）模型	1.预测结果更加精确 2.特别适用于 1 公里到 20 公里的小区 3.适用范围广 4.适用于 3G、4G 通信频段	1.是一种统计模型，需要进行大规模测试数据。

4.1.2 模型特征提取

本文基于大量文献及上述三个主要的经典模型，提取具有代表性和共性的特征如下表4所示。

通过比对三个经典统计模型的特征参量与文件中所包含的变量，可以得到如下五个基本特征。

①**链路距离 d**，即基站与观测点的水平位置距离。

$$d = \mathrm{sqrt}\left((\mathrm{X_C} - \mathrm{X})^2 + (\mathrm{Y_C} - \mathrm{Y})^2\right) \tag{9}$$

特征 d 在 Okumura-Hata 模型、Cost 231-Hata 模型、SPM 模型均存在，并且 $\lg(d)$ 与电波信号传播损耗呈线性关系。大量文献中也表明，距离越远，平均信号接收功率越小，本质上是由于无线电波在传播过程中可能收到空气中大量粒子以及物体的影响，距离越远，无线电波在传播的过程中可能遇到阻碍的几率就越大，损耗可能越多，这符合实际情况。

②**基站发射机绝对高度** h_b，即基站发射机距地面高度与基站海拔高度之和。即：

$$h_b = \mathrm{H_M} + \mathrm{AC} \tag{10}$$

③**观测点绝对高度** h_b，即观测点的海拔高度。即：

$$h_m = \mathrm{A} \tag{11}$$

特征②和③在 Okumura-Hata 模型、Cost 231-Hata 模型、SPM 模型均存在, 且 $\lg(h_b)$、$\lg(h_m)$ 均与电波信号传播损耗呈线性关系。此外，本文还发现 $\lg(h_m) * \lg(d)$ 也与电波信号传播损耗呈线性关系。

表 4　模型特征比较提取表

特征	变名	Okumura-Hata 模型	Cost 231-Hata 模型	SPM 模型
基站坐标	Cell X	链路距离 d	链路距离 d	K2 是距离衰减因子
基站坐标	Cell Y			
观测点坐标	X			
观测点坐标	Y			
基站发射机中心频率	Frequency Band	载波频率 f	载波频率 f	K_1 是与频率相关的因子
基站发射机相对地面的高度	Height	基站天线有效高度 ℏb	基站天线有效高度ℏb	H Txeff 是发射天线的有效高度
基站海拔高度	Cell Altitude			
观测点海拔高度	Altitude	用户天线有效高度 ℏm	用户天线有效高度ℏm	K3 是接收天线高度相关的因子
基站地物类型索引	Cell Clutter Index	场景（地区）纠正常数 K、α	场景（地区）纠正常数 K、α	K4 是与衍射相关的因子
观测点地物类型索引	Clutter Index			
基站所在格点的建筑物平均高度	Cell Building Height			
观测点所在格点的建筑物平均高度	Building Height			
小区编号	Cell Index			
基站发射机水平方向角	Azimuth			
基站发射机垂直电下倾角	Electrical Downtilt			
基站发射机垂直机械下倾角	Mechanical Downtilt			
基站发射机发射功率	RS Power			

④**载波频率** f，即基站发射机中心频率 f，且 $\lg(f)$ 常常与 $lg(\mathrm{A})$ 结合来影响电波信号传播损耗。

⑤**场景纠正常数 K**，即区域建筑物和地形对该观测点的电波信号传播损耗的影响。由于不同模型对这一特征的量化方式不同，因此目前仅了解其大致的函数关系。可以发现，该特征影响的量化值一般与 $\lg(f)$ 及 $\lg(A)$ 呈线性关系。即：

$$\mathrm{K} = \mathrm{F}(\lg(f), \lg(A)) \tag{12}$$

综上所述，通过问题一的分析，本文可以初步得到特征 lg(d)、lg(h_b)、lg(h_m)、lg(h_m)$*$lg(d)、F(lg(f)，lg(A))。同时，现有特征仍存在以下三点不足：

1. 仅仅分析了三个具有代表性的经典统计模型，且这些模型都是在很早之前提出，虽然有一定借鉴意义，但对于现在无线电波传播损耗的预测必然存在较大误差。

2. 数据集中的大量变量并未使用，这些变量的合理组合对于模型预测效果可能有较大帮助。

3. 有关区域范围的建筑物和地形因素的量化方法，仍然有待深入研究。

对于以上三点不足，文本对于将在问题二的解决过程中进行确定与解释。

4.2 问题二建模与求解

问题二属于数据分析与特征提取问题，对于解决此类问题采用因子分析和相关分析。

针对问题二，首先需要对数据进行可视化，对于已有数据各个变量的基本统计性质具有初步的了解，进一步，对数据进行清洗，去除收集时存在的非偶然性误差，以及无关的具有多重共线性的变量。然后，本文在数据采样的基础上，使用因子分析的方法对变量进行组合，对部分分类变量进行合理的排序与数值转化。最后，基于问题一中模型分析与因子分析结果，对已有变量进行组合形成特征，并对特征进行数值化表达与相关性测定。

4.2.1 数据介绍及可视化

本文所使用的数据全部来自华为云 AI 竞赛平台，数据包含 train_set 和 test_set 两个文件夹(https://developer.huaweicloud.com/competition/competitions/1000013923/circumstances)，其中 train_set 包含来自 4000 个文件的 18 个变量的数据，其中每个文件中大约包含 3000 条数据，共计 12011833 条，变量包括：Cell Index、Cell X、Cell Y、Height、Cell Altitude、Cell Building Height、Cell Clutter Index、Azimuth、Electrical Downtilt、Mechanical Downtilt、Frequency Band、RS Power、X、Y、Altitude、Building Height、Clutter Index、RSRP。

在数据中心，基站在 X 轴的坐标跨度范围为（384180，434540），在 Y 轴的坐标跨度范围为（3376325，3417960），观测点 X 轴的坐标跨度范围为（382930，434580），在 Y 轴的坐标跨度范围为（3375740，3418880），需要特别注意的是观测点和基站坐标为所在格点的左上角坐标，格点宽度为 5 米，示意图如图6所示。

进一步，本文将全部观测点的建筑物高度信息依据（X，Y）信息绘制在如图7所示的平面上，建筑物高度低于 9 米或不存在观测点的地方在图中表现为白色。途中，可以看到地图中心的六个呈环状的道路和两条隐约的河流，在中心城市的南部存在部分高于 100 以上的建筑群。

如图8所示，是观测点地物类型索引的直方分布图，可以看出观测点地物类型索引类型为市区开阔区域的占去大多数，道路、植被区、中高层建筑其次，可以看出观测点主要分

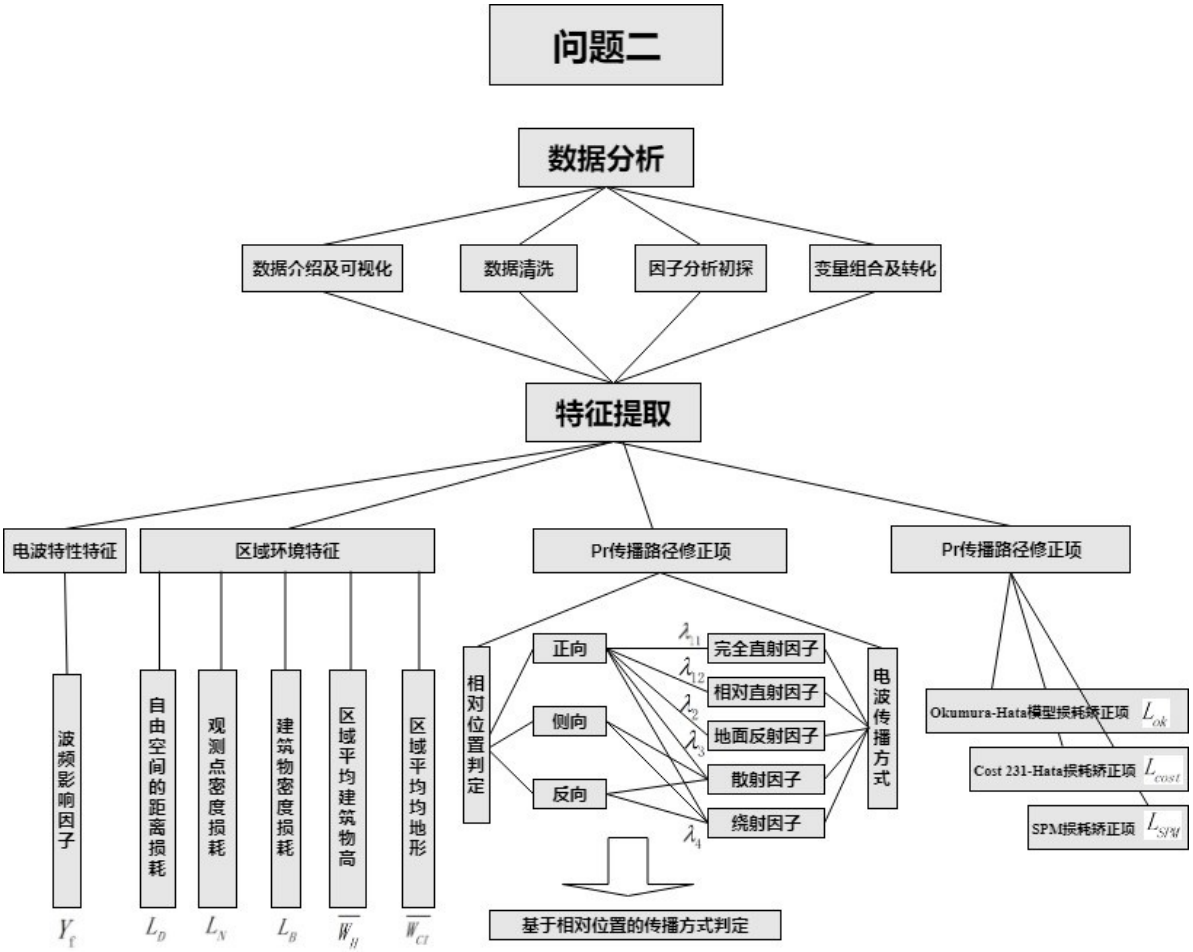

图 5　问题二思路图

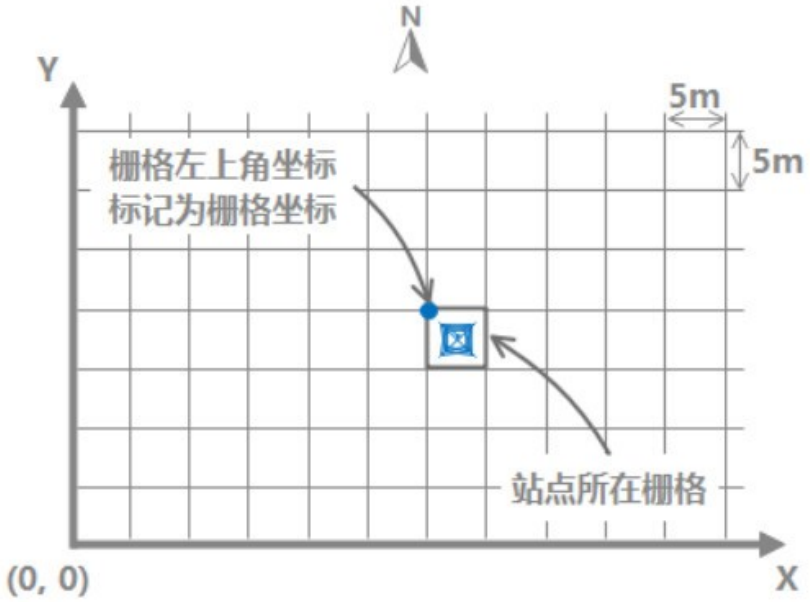

图 6　栅格化坐标示意图

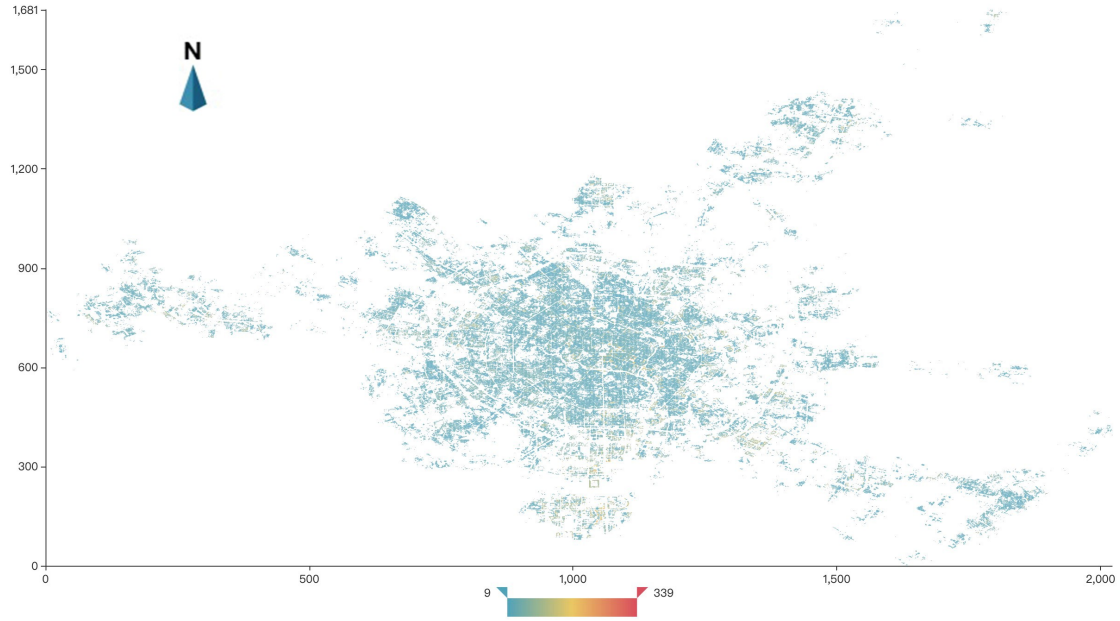

图 7　栅格化坐标示意图

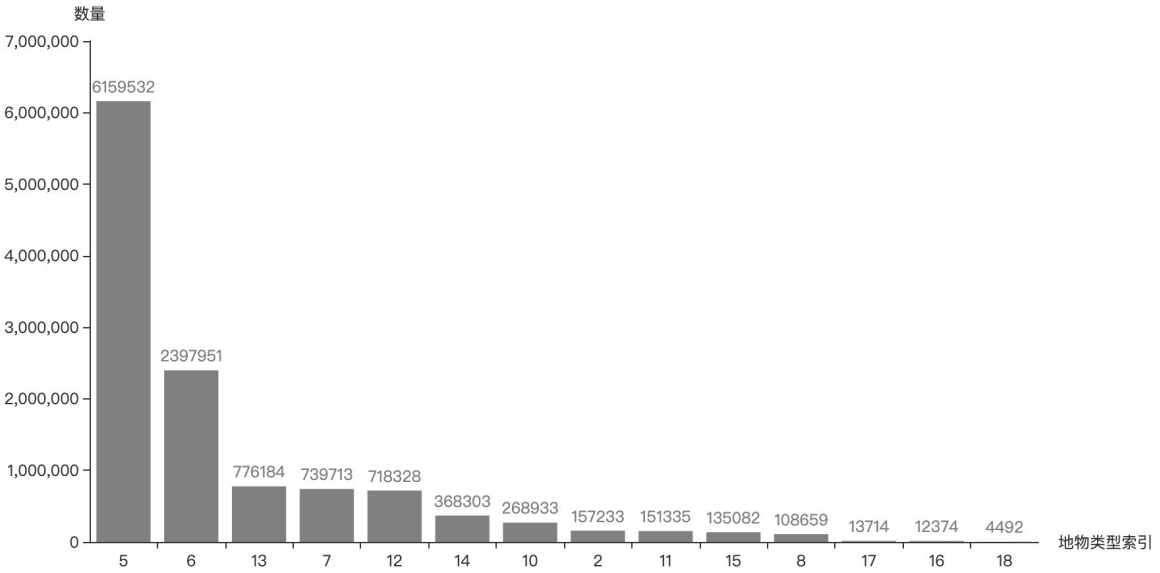

图 8　二维图 X Y 建筑物高度

布在城市中心区域。同时，可以发现观测点中不包含海洋（Clutter Index = 1）、湿地（Clutter Index = 3）、城郊开阔区域（Clutter Index = 4）、森林植被（Clutter Index = 9）、农村（Clutter Index = 19）、CBD 商务圈（Clutter Index = 20）这些地物类型。

4.2.2 数据清洗

对于部分异常数据 (建筑物高度和地物类型标识不符、空间距离与信号功率不符)，本文在模型训练前先进行了数据清洗。

1. 建筑物高度和地物类型标识不符

表 5　地物类型索引表

Clutter Index	含义	Clutter Index	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

如表5所示，题目中的 Table 4 给出了地物类型名称的编号含义，其中地物类型是隐含大量高度信息的。例如：当地物类型索引为 14 时，该格点的建筑高度仍存在大于 20 米的现象，同理，分析可知地物类型索引为 2、5、6、7、8 的区域不应该包含高于 20 米建筑建筑、地物类型索引为 10、11、12、13、14 的格点内的建筑高度不应该与地物类型索引所描述的内容相矛盾。

然而分析数据后，本文发现，存在部分变量矛盾的数据，因此有必要进行数据清洗，例如：小区编号为 2461901 的坐标为（411170，3395480）的观测点建筑物高度为 12 米，但地物类型索引为 10，与建筑物高度高于 60 米相矛盾。小区编号为 1231001 的坐标为（411140，3395880）的观测点建筑物高度为 156 米，但地物索引类型为 14，建筑物高度低于 20 相矛盾。

对于此类异常数据，本文对其进行数据清洗。

2. 空间距离与信号强度不符

大量文献资料表明，距离对于信号强度的影响是很大的，然而数据中存在部分空间距离远，信号强度却很大的点。为了去除这种异常点，首先，本文先定义了无线电波传播损

耗 L，其单位为 dB。

$$L = Pr - R \tag{13}$$

根据题目所给信息，本文定义了损耗大小的临界值 $L_0 = \overline{Pr} + 103 = 114.6$，其中 Pr 在 2097 个基站中的均值为 11.593514544587505，约为 11.6dbm。

其次，本文对空间距离 $L_D = \lg\left(\text{sqrt}\left((X_C - X)^2 + (Y_C - Y)^2 + ((H_M + A_C) - A)^2\right)\right)$ 进行排序，取其 0.2 分位数为 116.94443124834973，约为 116.9 米，与 0.8 分位数为 481.92281539655426，约为 481.9 米。

对于距离小于 116.9 米的观测点，若其无线电波传播损耗 L 大于 114.6，则认为是异常数据，应当剔除；同理，对于距离大于 481.9 米的观测点，若其无线电波传播损耗 L 小于 114.6，则认为是异常数据，应当剔除。

例如：小区编号为 1001701 的坐标为（424455，3376320）的观测点距离其基站的距离为 64.82 米，无线电波传播损耗为 150.2 大于 114.6。小区编号为 1032901 的坐标为（428735，3418170）的观测点距离其基站的距离为 212.84 米，无线电波传播损耗为 66.7 小于 114.6。

对于此类异常数据，本文对其进行数据清洗。

3. 无用变量的剔除

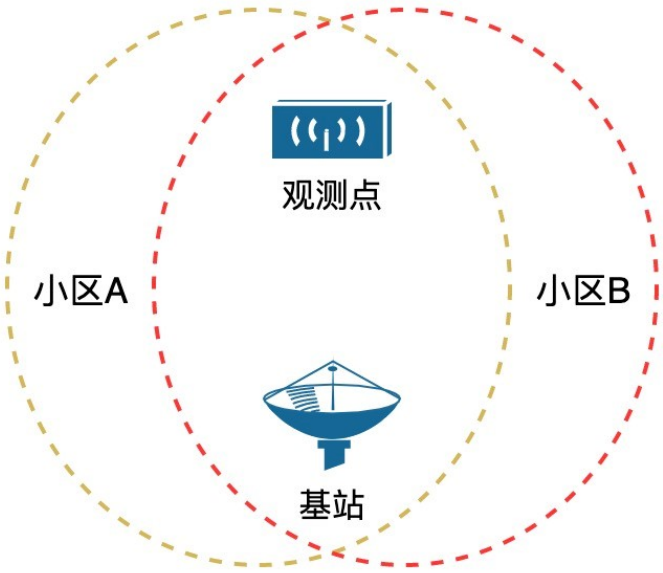

图 9　基站与小区关系图

对于数据进行基本的数据处理可以发现，题目所提供的数据共包含 4000 个小区和 2097 个基站，存在大量小区编号不同但是基站站点相同的数据，因此，小区编号与基站坐标不存在一一对应关系，因此，以小区为单位分析数据并不如以所属站点为单位分析数据更合理，因此，本文将数据的小区编号变量剔除。

4.2.3 因子分析初探变量组合

由于数据量过大，本文首先采用随机采样的方法获得样本大小为 1048574 数据，对剔除小区编号和 RSRP 值的 16 各变量进行因子分析。

1. 考察原有变量是否适合进行因子分析

首先，应该考察手机到的原有变量之间是否存在一定的线性关系，是否适合采用因子分析提取因子。这里，借助变量的相关系数矩阵、反应像相关矩阵、巴特利特球度检验和 KMO 检验方法进行分析，见表6。同时，由于数据中存在缺失值，采用剔除变异值处理缺失值。

表 6　KMO 和巴特利特检验

指标	数值
KMO 取样适切性量数	.588
近似卡方	11552718.086
自由度	120
显著性	.000

巴特利特球度检验统计量的观测值为 11552718.086，相应的概率 P 值接近 0. 如果显著性水平为 0.05，由于概率 P 值小于显著性水平，则应拒绝原假设，认为相关系数矩阵与单位阵有显著性差异。同时，KMO 值为 0.558，根据 Kaiser 给出的 KMO 度量标准可知数据适合进行因子分析。

2. 提取因子

在这里进行尝试性分析：根据原有变量的相关系数矩阵，采用主成分分析法提取因子并提取大于 1 的特征值。

表 7　总方差解释

在表7中，可以看出，第一列为是因子编号，以后三列组成一组，每组中数据项的含义一次是特征值、方差贡献率和累计方差贡献率。

第一组数据项（第二列-第四列）描述了因子分析初始解的情况。可以得到：第 1 个因子的特征值为 3.159，解释原有 17 个变量总方差的 19.743%（即 3.159/16*100%），累计方差贡献率为 19.743%；第 2 个因子的特征值为 1.781，解释原有 16 个变量总方差的 11.134%（1.781/16*100%），累计方差贡献率为 30.876%。其余数据含义类似。在初始解中由于提取

	初始特征值			提取载荷平方和			旋转载荷平方和		
成分	总计	方差百分比	累积 %	总计	方差百分比	累积 %	总计	方差百分比	累积 %
1	3.159	19.743	19.743	3.159	19.743	19.743	2.161	13.507	13.507
2	1.781	11.134	30.876	1.781	11.134	30.876	2.041	12.758	26.265
3	1.738	10.863	41.739	1.738	10.863	41.739	1.855	11.593	37.858
4	1.138	7.113	48.852	1.138	7.113	48.852	1.730	10.815	48.673
5	1.065	6.657	55.509	1.065	6.657	55.509	1.088	6.803	55.476
6	1.017	6.359	61.868	1.017	6.359	61.868	1.023	6.392	61.868
7	.994	6.211	68.079						
8	.977	6.104	74.183						
9	.962	6.011	80.195						
10	.937	5.856	86.050						
11	.869	5.429	91.480						
12	.814	5.088	96.567						
13	.467	2.918	99.485						
14	.065	.408	99.893						
15	.010	.060	99.953						
16	.007	.047	100.000						

提取方法：主成分分析法

了 16 个因子，在初始解中由于提取了 6 个因子，因此原有变量的总方差均被解释，累计方差贡献率为 100%。

第二组数据项（第五列到第七列）描述了因子阶的情况，可以看出，由于提取了四个因子，四个因子一共解释了原有变量总方差的 61.868%。总体上，原有变量的信息丢失较少，因子分析效果较理想。

第三组数据项（第八列到第十列）描述了最终因子阶的情况。可见，因子旋转后，总的累计变量的方差贡献率没有较大改变，也就是没有影响原有变量的共同度，但却重新分配了各个因子解释原有变量的方差，改变了各因子的方差贡献，使得因子更易于解释。

在图10中，横坐标为因子数目，纵坐标为特征值。可以看出：第 1 个因子的特征值很高，对解释原有变量的贡献最大。第六个以后因子的特征值小于 1，对解释原有的变量的贡献相对较小，已经可以成为可忽略的“高山脚下的碎石”，因此提取 6 个因子是合适的。

3. 因子的命名解释

这里，本文采用最大方差法对因子载荷矩阵进行蒸饺旋转以使因子具有命名解释性。指定按第一个因子载荷降序的顺序输出旋转后的因子载荷，并绘制旋转后的因子载荷图。

由表8可知，Cell X、X 在第一个因子上有较高的载荷，Cell Y、Y 在第二个因子上有较高的载荷，前两个因子主要包含了基站与观测点的距离信息，可解释为距离；Cell

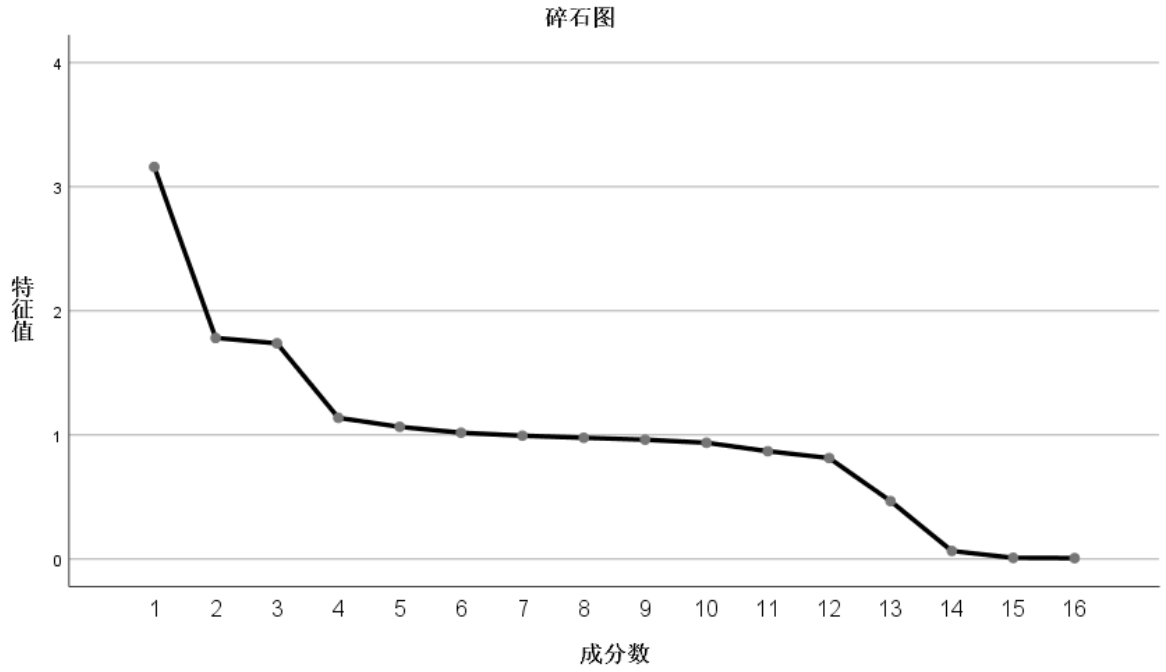

图 10

	成分					
	1	2	3	4	5	6
Cell X	.968	-.040	-.127	.005	.004	.012
Cell Y	-.046	.976	.159	.001	.009	-.012
Height	.043	-.010	.060	.834	.033	.044
Cell Altitude	-.333	.234	.828	.027	.045	.074
Cell Building Height	-.019	.018	-.151	.204	.424	-.168
Cell Clutter Index	.004	-.006	.060	-.044	.563	-.362
Azimuth	.072	-.054	.088	.022	.004	.703
Electrical Downtilt	.062	-.055	.025	.802	-.043	.078
Mechanical Downtilt	-.093	.073	-.209	.536	.051	-.126
Frequency Band	-.147	.086	-.330	-.009	-.012	.398
RS Power	.134	.053	-.236	.454	-.139	-.150
X	.967	-.040	-.128	.007	.004	.004
Y	-.044	.977	.160	.000	.012	-.012
Altitude	-.339	.231	.825	.026	.060	.080
Building Height	.010	-.017	-.056	-.060	.479	.362
Clutter Index	.004	.018	.066	-.012	.576	.144

提取方法：主成分分析法

表 8　旋转后的成分矩阵

Altitude、Altitude 在第三个因子上有较高的载荷，该因子主要包含了基站与观测点的高度信息；Height、Electrical Downtilt、Mechanical Downtilt、RS Power 在第四个因子上有较高的载荷，该因子主要包含了与发射机有关的信息；Cell Building Height、Cell Clutter Index、Building Height、Clutter Index 在第五个因子上有较高的载荷，该因子主要包含了与环境有关的信息；Azimuth、Frequency Band 在第五个因子上有较高的载荷，该因子主要包含了发射机水平方向角和发射频率信息。

4.2.4 变量组合与转化

1. 因子分析结果解释与变量基本组合

因子分析的结果对于变量之间的组合具有指导意义，不论从因子分析模型、还是从距离的现实意义角度，结合 Cell X 和 X 变量、Cell Y 和 Y 变量、Cell Altitude 和 Altitude 变量的组合都是合理的。第四组变量组合显示出 Height、Electrical Downtilt、Mechanical Downtilt、RS Power 四个变量的组合，虽然都是有关基站发射机的变量，但其特征结合方式并不是很顺理成章的，仍需要进一步分析。在第五组变量组合中，Cell Building Height、Cell Clutter Index、Building Height、Clutter Index 主要包括地物类型索引和格点建筑物高度，是影响电波传输方式的重要因素，其结合是合理的。第六组变量组合将 Azimuth、Frequency Band 两个变量结合起来，其原因仍有待分析。

2. 变量转化

本文进一步分析了地物类型索引对电波损耗的影响大小，并认为其平均影响可用反射率来量化，并依据其平均影响的大小对地物类型索引的地形校正因子 W_cI 进行排序。

首先，需要先定义无线电波传播损耗 L，其单位为 dB。

$$L = Pr - R \tag{14}$$

进一步本文通过对无线电波传播损耗 L 进行标准化与归一化，来对地物类型索引的地形矫正因子 W_cI 进行排序，数值计算结果如表9所示： 表中，$\overline{L}$ 为传播损耗均值，$\overline{L}^*$ 为标

表 9　矫正因子

准化后的传播损耗数值。W_{cI} 值越大，传播损耗越大，吸收的电波越多，反射的电波越少。

4.2.5 特征提取与数值计算方法

在因子分析初步探究小区数据集中变量关系后，可以发现，提取的六个主要因子的方差贡献率仅仅为 61.868%，方差中仍存在大量的未能被模型完全提取，因此，本文将结合

观测点地物类型索引	数目	$\overline{L}$	$\overline{L}^{*}$	W_{CI}
2	13748	103.5397934	0.041625576	0.455196
5	538331	103.1312859	0.004619031	0.451124
6	208968	103.5755472	0.044864494	0.455552
7	64145	103.1844933	0.009439069	0.451654
8	9522	100.4445085	-0.238775112	0.424345
10	23503	105.2859537	0.199809575	0.4726
11	13296	102.8906002	-0.017184594	0.448725
12	62606	102.68627	-0.035694787	0.446689
13	67864	102.6770363	-0.036531267	0.446597
14	32044	101.1273905	-0.176913094	0.431151
15	11892	99.84396485	-0.293178134	0.418359
16	1059	97.72405099	-0.485220328	0.39723
17	1210	97.90183471	-0.469114968	0.399002
18	386	94.9942487	-0.732512083	0.370021
样本最小值		57.87	-4.09558371	0
样本最大值		158.2	4.993272985	1

对于已有无线电波传播模型的分析结果与因子分析的变量组合，提炼出符合现实意义的有效特征。

（一）无线电波特性特征

大量文献表明，电波信号的传播首先与电波自身的性质有关，特别是电波的频率与波长，因此，本文首先引入波频影响因子作为特征。

1. 波频影响因子 Y_f

在 4.1 中研究的三个模型中表明，$\lg(f)$ 对于电波信号传播损耗 L 有较为显著的影响，因此，在模型中加入特征波频影响因子 Y_f，即：

$$Y_f = \lg(f) \tag{15}$$

（二）区域环境特征

除了无线电波本身的特性特征外，基站发射机所处的环境也是重要的影响因素，对于量化电波散射、绕射，乃至散射具有重要影响。

1. 自由空间的距离损耗 L_D

在 4.1 中研究的三个模型和因子分析的结果表明，基站与观测点之间的距离对于电波信号传播损耗 L 有较为显著的影响，因此，在模型中加入特征自由空间的距离损耗 L_D，即：

$$L_D = \lg\left(\mathrm{sqrt}\left((X_C - X)^2 + (Y_{C-}Y)^2 + ((H_M + A_C) - A)^2\right)\right) \tag{16}$$

在实际运算中，由于 lg(0) 不存在的限制，需要剔除基站与观测点位置相同的数据。本文的处理是当 $L_D < 1$ 时，默认 $L_D = 1$，认为该距离下不存在无线电波距离损耗。

2. 观测点密度损耗 L_N

由于数据观测点是随机选取的，因此，认为观测点位置的密度分布即为真实信号接收器分布的密度分布，大量文献表明，当无线电波传播过于密集时，其传波损耗值会相应增加，因此，本文考虑加入观测点密度损耗值 L_N，即：

$$L_{N=N_i} \tag{17}$$

其中，N_i 表示第 i 个基站的所有观测点的数目。如图11所示，是每个基站的观测点数

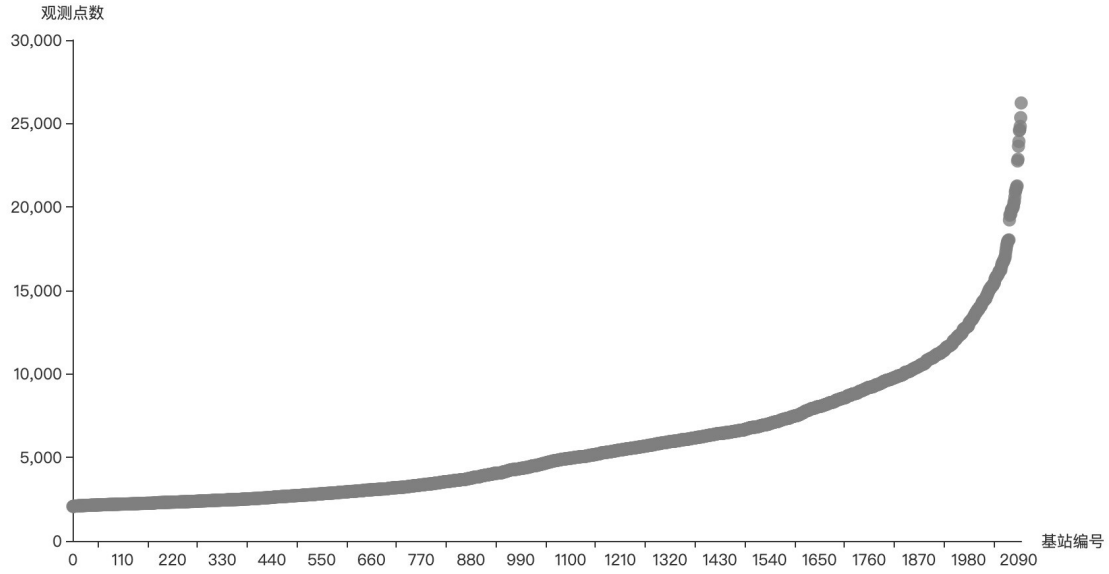

图 11　每个基站的观测点数量分布图

量的散点分布图，可以看出 2097 个基站，其观测点分布相当不均匀，基本服从幂律分布，95% 基站的观测点数量在 15000 以下，仅有很少的基站具有特别多的观测点数。

3. 建筑物密度损耗 L_B

在 CCIR 模型中，研究者通过在在 Okumura-Hata 模型的基础上，增加地形对信号传播损耗的影响来改进模型，其引入的地形覆盖物校正因子 B 的经验性公式为：

$$B = 30 - 25\lg\left(\text{地面建筑物覆盖率}\right) \tag{18}$$

受 CCIR 模型启发，本文将属于同一个基站的所有观测点中，建筑物高度不为 0 的观测点的比例定义为建筑物密度损耗 L_B，其计算公式为：

$$L_B = -\lg\left(M_i/N_i\right) \tag{19}$$

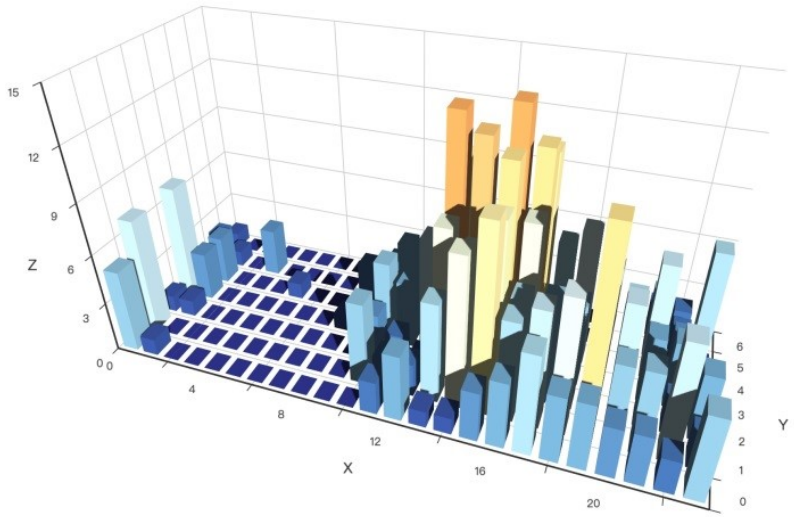

图 12　建筑物密度图

其中，M_i 为建筑物高度不为 0 的观测点数量，为便于对数计算，M_i 的最小取值为 1。

4. 平均建筑物高度影响因子 W_H

与建筑物密度损耗的思想相似，本文引入属于同一个基站的所有观测点的的平均建筑物高度影响因子，其计算公式为：

$$\overline{W_H} = \frac{1}{N_i}\sum_{j=1}^{n_i} H_j \tag{20}$$

其中，H_i 为观测点的建筑物高度。这个公式中，可以看出建筑物平均高度越大的区域，传播损耗也会也大。

5. 平均地形影响因子 W_{CI}

同理，可以得到属于同一个基站的所有观测点的的平均地形影响因子，其计算公式为：

$$\overline{W_{CI}} = \frac{1}{N_i}\sum_{j=1}^{n_i} (W_{CI})_j \tag{21}$$

其中，W_{CI} 为地物类型索引的地形矫正因子。

（三）发射功率的传播路径修正项 Pr^*

无线电波的信号强度除了受环境特征的影响外，很大一部分源自于基站发射机发射功率的影响，但由于观测点与基站的相对位置，以及基站所处的区域不同，需要对发射功率进行修正。

1. 相对位置的判定

首先，需要判定观测点与基站的相对位置。在水平方向上，设发射机张角大小为 γ，即信号偏离信号发射机中心线的最大角度。在查阅大量文献与资料后，取 γ 值为 30 度。为判断观测点与发射机的相对位置，即观测点在基站发射机方向的正面、侧面、反面，本文

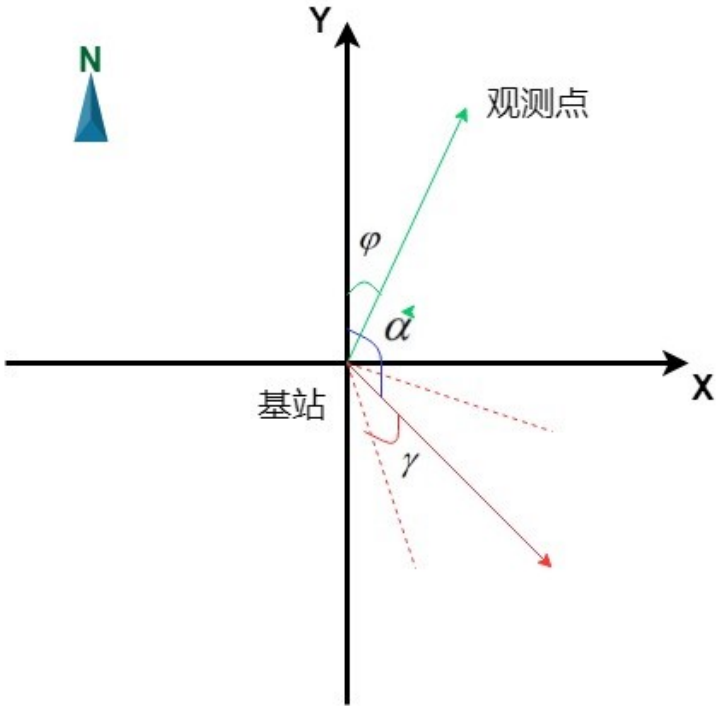

图 13　相对位置示意图

计算了以基站为中心，在水平方向上，基站和观测点连线与正北方向的夹角 φ，其计算公式为：

$$\varphi == -\arctan\left(\frac{|X-X_C|}{|Y-Y_C|}\right) \tag{22}$$

其判断条件为：

$$|\alpha-\varphi| \in \begin{cases} [0,\gamma] & \text{正向} \\ (\gamma,180-\gamma)\cup(180+\gamma,360) & \text{侧向} \\ [180-\gamma,180+\gamma] & \text{反向} \end{cases} \tag{23}$$

2. 电波传播方式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，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，所以在相对位置的判定之后，需要对不同的电波传播方式进行量化研究。

（1）直射因子

在直射因子中，本文将直射因子分为完全直射因子和相对直射因子，完全直射因子应用于无线电波能直线传播到的观测点，其公式为：

$$\lambda_{11} = \mathrm{P_r} \tag{24}$$

如图14所示，为判断无线电波的直射范围，首先，需要定义在垂直平面上，发射机的张角为 β，在垂直方向观测点基站连线与水平方向的夹角为 ϕ。在查阅大量文献与资料后，

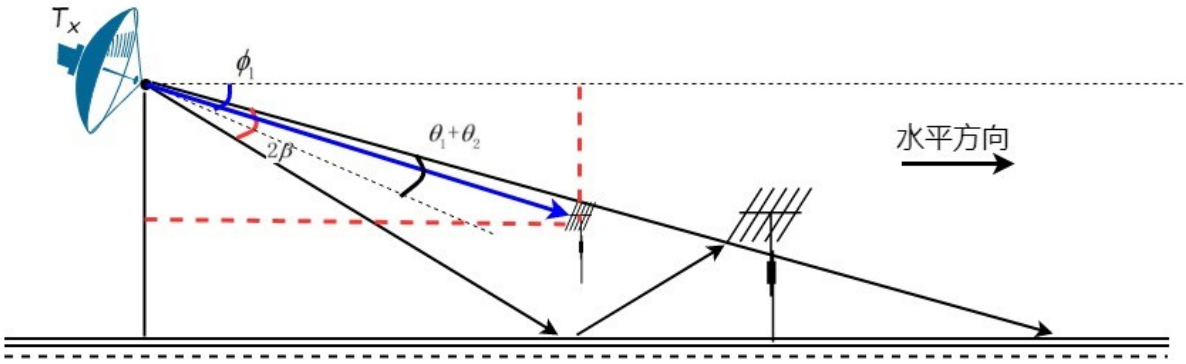

图 14　完全直射示意图

取 β 值为 30 度。

$$\phi_1 = \arctan\left(\frac{|H_\mu + A_c - A|}{\sqrt{(X_c - X)^2 + (Y_c - Y^2)^2}}\right) \tag{25}$$

当 $\phi_1 \in (\theta_1 + \theta_2 - \beta, \theta_1 + \theta_2 + \beta)$ 时，本文认为，存在完全直射，否则，该观测点不存在完全直射，而存在绕射现象。

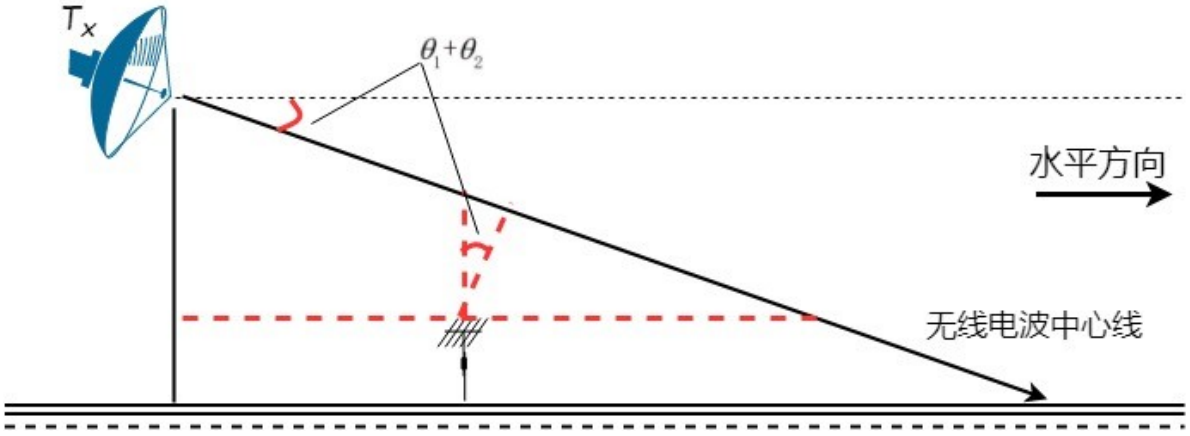

图 15　相对直射示意图

如图图15所示，为引入相对直射因子，用于无线电波距离中心线的远近，来间接量化直射作用，其公式为：

$$\lambda_{12} = (1/\mathrm{D}) * \mathrm{P_r} \tag{26}$$

其中，D 为观测点坐标距离中心直射线的最短距离，为了使 $\frac{1}{D}$ 有意义，默认当 $\mathrm{D} < 1$，则取 $\mathrm{D} = 1$。其计算公式为:

$$D = (H_M + A_C - A) \times \sin(\theta_1 + \theta_2) \tag{27}$$

（2）地面反射因子

研究发现，当无线电波传播时遇到尺寸比自身波长大得多的障碍物时，会反生发射现象，这种现象一般发生于物体表面不同介质交接处，为量化这种影响，引入地面反射因子 λ_2 ，其公式为：

$$\lambda_2 = \mathrm{W_{CI}} * \mathrm{P_r} \tag{28}$$

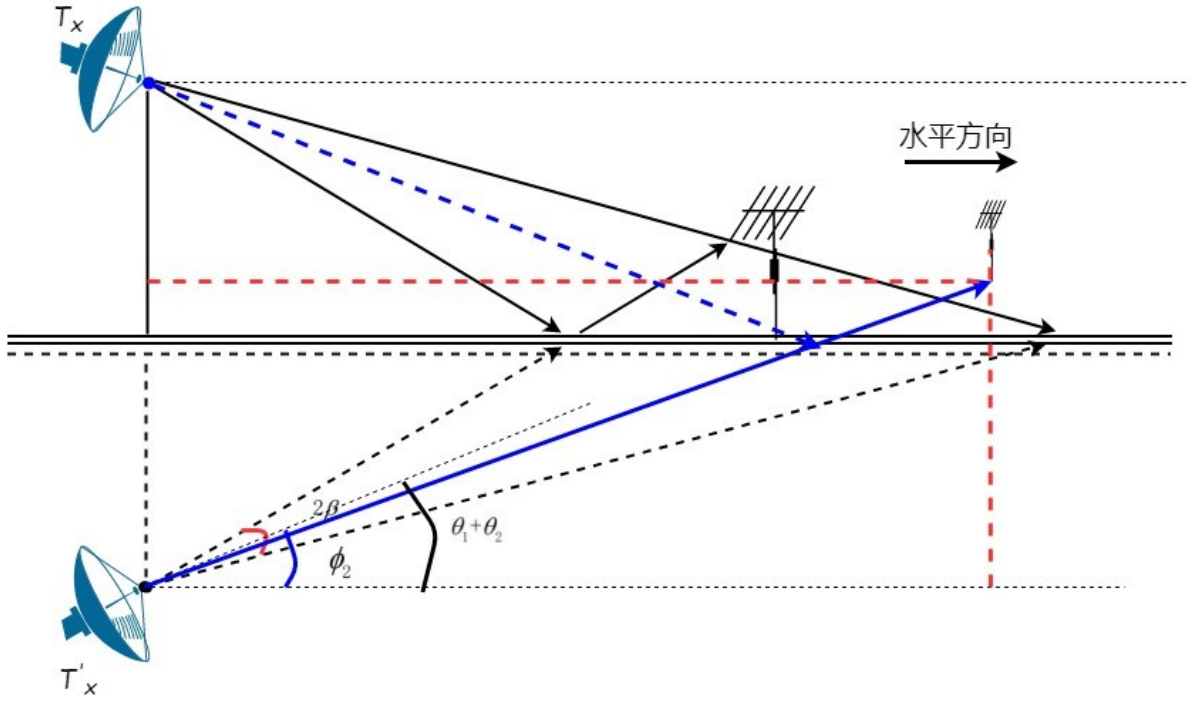

图 16　反射示意图

如图16所示，为判断无线电波的直射范围，首先，需要定义在垂直平面上，发射机的张角为 β ，在垂直方向观测点基站连线与水平方向的夹角为 ϕ 。在查阅大量文献与资料后，取 β 值为 30 度。

$$\phi_2 = \arctan\left(\frac{(H_M + A_C + A)}{\sqrt{(X_C - X)^2 + (Y_C - Y)^2}}\right) \tag{29}$$

当 $\phi_2 \in (\theta_1 + \theta_2 - \beta, \theta_1 + \theta_2 + \beta)$ 时，认为存在地面反射，否则，该观测点不存在地面反射，而存在绕射现象。

（3）散射因子

当无线电波传播时存在小于波长的物体且单位体积内该物体数量巨大，会产生散射现象。比如粗糙表面、小物体和不规则物体（植被、广告灯等）。地表面散射波示意图如图17。

根据大量文献及资料，本文定义了散射因子 λ_3 ，其公式为：

$$\lambda_3 = (\mathrm{W_{CI}^{*}LB}) * \mathrm{P_r} \tag{30}$$

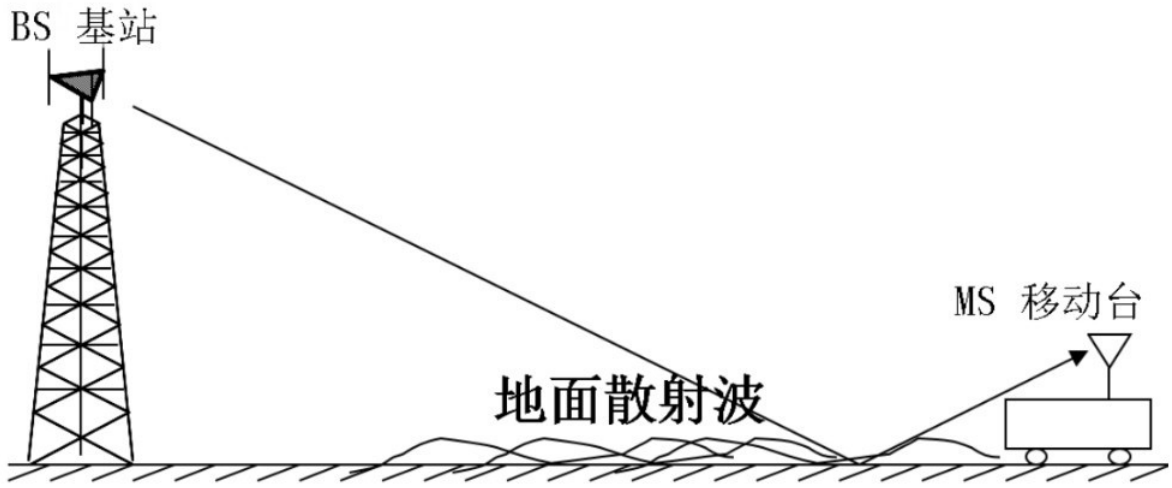

图 17　地表面散射示意图

（4）绕射因子 λ_4

当无线电波传播时遇到尖利的边缘阻挡时发生绕射现象，波长越长，绕射能力越强；波长越短，则绕射能力越弱。无线电波绕射传播示意图如图18所示.

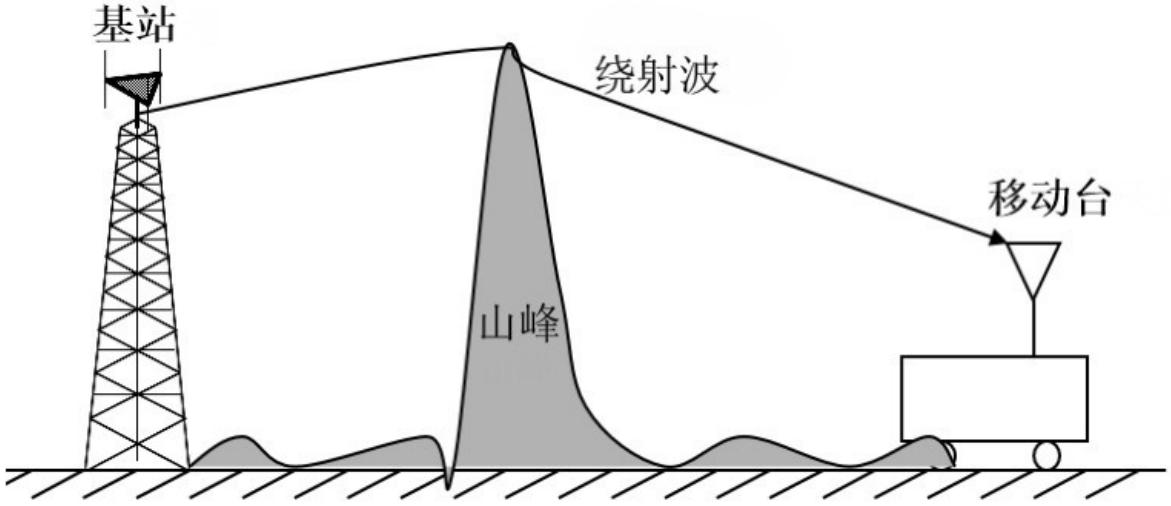

图 18　绕射示意图

根据大量文献及资料，本文定义了散射因子 λ_4 ，其公式为：

$$\lambda_4 = (W_{CI}^{*}(Ni/1000) + W_{CI} * (Mi/000)) * Pr \tag{31}$$

3. 基于相对位置的传播方式的判定

在确定了基站和观测点相对位置和量化传播方式影响的基础上，基于大量文献资料，本文对各个区域的发射功率进行修正。首先，对正面区域的发射功率 Pr 进行修正：若符合

完全直射与地面反射的判断条件，即存在无线电波的完全直射、相对直射、地面反射、散射、绕射现象，则 $\mathrm{Pr}^{*}_{\text{正}} = \lambda_{11} + \lambda_{12} + \lambda_2 + \lambda_3 + \lambda_4$

若仅符合完全直射的判断条件，即存在无线电波的完全直射、相对直射、散射、绕射现象，则 $\mathrm{Pr}^{*}_{\text{正}} = \lambda_{11} + \lambda_{12} + \lambda_3 + \lambda_4$

若仅符合地面反射的判断条件，即存在无线电波的相对直射、地面反射、散射、绕射现象，则 $\mathrm{Pr}^{*}_{\text{正}} = \lambda_{12} + \lambda_2 + \lambda_3 + \lambda_4$

其次，对侧面区域的发射功率 Pr 进行修正：

在侧面区域，基于大量文献与资料研究，认为存在无线电波的散射、绕射现象，则 $\mathrm{Pr}^{*}_{\text{正}} = \lambda_3 + \lambda_4$

最后，对反面区域的发射功率 Pr 进行修正：在反面区域，基于大量文献与资料研究，认为存在无线电波的散射、绕射现象，则 $\mathrm{Pr}^{*}_{\text{反}} = \mathrm{W_{Cl}} * (\lambda_3 + \lambda_4)$

（四）其他模型校正项

1. Okumura-Hata 模型损耗校正项 $\mathrm{L}_o k$

Okumura-Hata 模型为：

$$\mathrm{Lb} = 69.55 + 26.16 \lg \mathrm{f} - 13.82 \lg \mathrm{h}_b - \alpha(\mathrm{h}_m) + (44.9 - 6.55 \lg \mathrm{h}_b) \lg d \tag{32}$$

其中，f：工作频率（MHz）；h_b：基站天线有效高度（m）；h_m：移动台天线有效高度（m）；d：移动台与基站之间的距离（km）；$\alpha(\mathrm{h}_m)$：移动台天线高度因子。在不同的地形地区，可以通过调整 α、K 使公式的预测效果更好。

由于本赛题的大多数数据点楼高较高，因此，认为该数据处于大城市，且数据的无线电波频率大多为 2585Mhz>300Mhz，因此，$\alpha(\mathrm{h}_m) = 3.2 \lg(11.75\mathrm{h}_m)]2 - 4.97dB$，K 为地区矫正因子,$K = -2[\lg(f/28)]\text{^}2 - 5.4$。

因此，得到 Okumura-Hata 模型损耗校正项：

$$\begin{aligned} L_{\mathrm{dx}} &= 69.55 + 26.16 \lg(f) - 13.82 \lg(H_{\mathrm{M}} + A_c) - 3.2\left((\lg(11.75 * A)\right)^2 \\ &+ (44.9 - 6.55 \lg(H_{\mathrm{M}} + A_c)) * L_D - 2(\lg \tfrac{f}{28})^2 - 10.37 \end{aligned} \tag{33}$$

2. Cost 231-Hata 模型损耗校正项 $\mathrm{L}_c ost$

Cost 231-Hata 模型为：

$$\mathrm{PL} = 46.3 + 33.9 \lg f - 13.82 \lg(h_b) - \alpha(h_m) + (44.9 - 6.55 \lg(h_u e)) \lg d + \mathrm{C}_m \tag{34}$$

其中 $\alpha(h_m)$ 取值同上，$\alpha(h_m) = 3.2[\lg(11.75\mathrm{hm})]2 - 4.97\mathrm{dB}$，$\mathrm{C}_m$ 置为 0 因此，得到 Cost 231-Hata 模型损耗校正项：

$$\begin{aligned} L_{\text{cost}} &= 46.3 + 33.91g(f) - 13.821g(H_M + A_c) - 3.2\left((\lg(11.75 * A)\right)^2 \\ &-4.97 + (44.9 - 6.551\mathrm{g}(H_\mu + A_c)) * L_D \end{aligned} \tag{35}$$

3. SPM 模型损耗校正项 LSPM SPM 模型为：

$$\begin{aligned} L_1 = K_1 + K_2 \lg d + K_3 \lg H_{\text{Txeff}} + K_4 \text{Diff\_}loss \\ + K_5 \lg H_{\text{Txeff}} \lg d + K_6 H_{\text{Rxeff}} + K_{\text{clutter}} f(\text{clutter}) \end{aligned} \tag{36}$$

其中参数取值如表10所示：

表 10　SPM 模型参数

变量名	中文含义	取值
K1	与频率相关因子	23.5
K2	距离衰减因子	44.9
K3	移动台天线高度相关因子	5.83
K4	与衍射计算相关因子	1(>0)
K5	与发射天线有效高度和距离相关的因子	-6.55
K6	移动台高度相关因子	0
K7	地貌相关因子	1

因此，得到 SPM 模型损耗校正项：

$$L_{SPM} = 23.5 + 44.9 L_D + 5.83 \lg (H_M + A_C) - 6.55 \lg (H_M + A_C) * L_D + 10 \tag{37}$$

4.2.6 特征相关性测定

综上所述，本文得到了 16 个原始变量特征、3 个统计模型矫正项特征以及 7 个新的组合的经验特征，对综上所述的 26 个特征进行相关性测定并排序，得到结果如表11所示。相关分析的数据是具有生成特征的 600000 随机抽样数据。

表 11　特征相关性测定表

** 在 0.01 级别（双尾），相关性显著。

* 在 0.05 级别（双尾），相关性显著。

如表11所示，对 26 个变量与 RSRP 值计算 Pearson 相关系数、Kendall 相关系数、Spearman 相关系数，其中对三个 600000 数据的样本三次计算求平均值，进一步，求三个相关系数的平均值，并剔除存在不显著相关系数的特征，最终得到 19 个有效特征，并依据其相关系数大小进行排序。**从表中，可以看出除了经典统计模型矫正项，空间距离即区域环境特征对于 RSRP 值有较大的影响。**

排序	特征名称	该特征与目标的相关性						变量是否有效	相关系数均值
		Pearson 相关系数	sig.	Kendall 相关系数	sig.	Spearman 相关系数	sig.		
1	SPM 损耗矫正项	-0.337**	0.000	-0.331**	0.000	-0.344**	0.000	是	-0.337
2	Cost 231-Hata 损耗矫正项	-0.336**	0.000	-0.330**	0.000	-0.344**	0.000	是	-0.337
3	Okumura-Hata 模型矫正项	-0.333**	0.000	-0.327**	0.000	-0.341**	0.000	是	-0.334
4	自由空间的距离损耗	-0.185**	0.000	-0.180**	0.000	-0.192**	0.000	是	-0.186
5	观测点所在格点的建筑物高度	-0.046**	0.000	-0.040**	0.000	-0.044**	0.000	是	-0.043
6	平均地形影响因子	-0.036**	0.000	-0.029**	0.000	-0.039**	0.000	是	-0.035
7	基站地物类型索引	-0.018**	0.000	-0.015**	0.009	-0.021**	0.000	是	-0.018
8	观测点 Y 坐标	0.010**	0.000	0.029**	0.000	0.013**	0.002	是	0.017
9	基站发射机发射功率	-0.015**	0.000	-0.022**	0.000	-0.014**	0.000	是	-0.017
10	基站 Y 坐标	0.008**	0.000	0.027**	0.000	0.010*	0.018	是	0.015
11	平均建筑物高度影响因子	-0.014**	0.000	-0.010**	0.000	-0.018**	0.000	是	-0.014
12	波频影响因子	-0.015**	0.000	-0.010**	0.000	-0.008**	0.000	是	-0.011
13	基站海拔高度	0.008**	0.000	0.010**	0.000	0.013**	0.001	是	0.010
14	观测点海拔高度	0.006**	0.000	0.010**	0.000	0.013**	0.002	是	0.010
15	观测点 X 坐标	0.010**	0.000	0.005**	0.000	0.012**	0.004	是	0.009
16	基站 X 坐标	0.010**	0.000	0.004**	0.000	0.011**	0.008	是	0.008
17	观测点地物类型索引	-0.029**	0.000	-0.026**	0.000	0.030**	0.000	是	-0.008
18	发射功率修正项	0.006*	0.012	0.007**	0.009	0.007**	0.003	是	0.007
19	基站发射机中心频率	-0.001**	0.000	-0.002**	0.000	-0.001**	0.000	是	-0.001
	基站发射机垂直电下倾角	-0.003*	0.013	-0.004	0.819	-0.009*	0.028	否	-0.005
	基站发射机相对地面的高度	-0.002	0.225	-0.002	0.712	-0.007*	0.020	否	-0.004
	基站发射机垂直机械下倾角	-0.004	0.013	-0.001	0.256	-0.005	0.541	否	-0.003
	基站所在格点的建筑物高度	0.001	0.923	0.001	0.831	0.008	0.145	否	0.003
	建筑物密度损耗	-0.002	0.128	-0.004	0.527	-0.003	0.214	否	-0.003
	基站发射机水平方向角	0.002	0.098	0.002	0.794	0.004	0.094	否	0.003
	观测点密度损耗	-0.001	0.297	0.000	0.969	-0.001	0.790	否	-0.001

4.3 问题三建模与求解

问题三属于建模预测与工程实践问题。

针对问题三，首先，基于问题一和问题二中设计和选择了有效特征，本文通过官方给出的训练集，建立了 2 种无线信号传播预测模型。其次，本文将对这 2 种模型进行依次的说明与训练求解。训练过程均使用的是一台 48 核 CPU、128G 内存、4*1080Ti 显卡的服务器。代码是基于 Python3.6、TensorFlow1.8.0、Lightgbm2.2.3 进行编写的，详见上传的附件。最后，在华为云的 ModelArts 平台进行了部署预测，且获得了优异的成绩。

4.3.1 单 GBDT 模型

GBDT 是由决策树模型衍生而来的一种集成学习的方法，它采用前向分布算法和加模型的方法，来实现学习的优化过程。GBDT 使用负梯度作为划分的指标（信息增益），在每次迭代训练中，它都是通过负梯度来拟合目标 Loss（残差），从而得到一颗决策树。不足之处是，计算信息增益时会扫描所有样本，从而找到最优划分点，这样就会导致寻找最优解变成一个特别耗时的过程。一种不错的优化方式是，在已经排好序的特征值上枚举所有可能的特征点，但这也是一种枚举的方法。

本文使用的则是另一种 GBDT 的衍生模型 LightGBM (Light Gradient Boosting Machine)，它在 2017 年由微软提出[9] 。它采用的是直方图算法来提升训练速度，这种方法会把连续特征值划分到多个桶中，划分点则在桶中选取。由于分桶后的桶数量是远小于特征点数量的，所以在训练速度和内存上都优于原有 GBDT 模型。且在大量数据集与实际应用场景中也得到了证明，这种分桶方式影响并不大，甚至会好一些。原因在于决策树本身就是一个弱学习器，采用柱状图式分桶算法会起到正则化的效果，也有助于防止模型过拟合。

通过前文的数据清洗与特征提取后，本文对整体数据集进行了 10 折划分，以 RSRP 为标签，均方根误差（RMSE）为优化目标，然后训练。LightGBM 的模型结构比较简单，且训练速度迅速，有利于特征的快速尝试与探究，在本文的多次尝试后，得到了以下的特征重要性表12。其中两个地物类型索引会被当做类别特征（Categorical Feature）输入模型，以区别于其他的数值特征。

同时，在多次训练与尝试的过程中，本文从模型重要性低、特征重复度高、特征冗余等方面，剔除了如下表14的特征。

最终，由于华为云的限制，此模型只用于离线特征快速分析，以及被当做后文中与 TensorFlow 融合的一个模型。**此模型的的线下均方根误差（RMSE）为 9.1736，PCRR 值（Poor Coverage Recognition Rate）为 12.7329%。**

4.3.2 融合模型与多目标学习

在有了前面的单 GBDT 模型的基础之上，本文对其模型的输出与原有特征进行了融合，然后将整体融合特征，作为了基于 TensorFlow 的多层神经网络模型的输入数据。模型

表 12　Top15 个特征重要性说明（数值越大越重要）

中文含义	代码中的变量名	重要度
观测点 Y 轴坐标值	Y	78
观测点 X 轴坐标值	X	69
X 轴相对距离	delta_X	68
按发射站聚合所有观测点平均建筑高度	meanH	68
发射机水平方向角	Azimuth	67
Y 轴相对距离	delta_Y	65
发射机的地物类型索引（类别特征）	Cell_Clutter_Index	56
按发射站聚合所有观测点高度取对数后平均	div_size	53
按发射站聚合观测点数量	grouped_size	52
观测点的 X 坐标值	Cell_X	49
发射机离地面的高度	Height	45
按发射机聚合后的平均损失因子	mean_index_delta	45
平均功率损失因子	X2	44
发射机的 Y 轴坐标值	Cell_Y	44
发射机离地面高度与海拔高度的和	hte	40

结构参考网上的开源代码[10] 并进行了模型结构调整，以适应本问题数据，整体流程见下图19。

图中，单 GBDT 模型是通过前文训练得到的，它的预测标签为 RSRP，优化目标是均方根误差（RSME），最终它的超参数如下表。

然后将此模型和原有特征拼接在一起，输入到一个多层神经网络模型中即可实现模型融合的效果。

对于此多层神经网络的输出节点，由于最终优化目标为 PCRR 值和 RMSE。所以本文将其当做了一个多目标学习问题来处理。一个输出节点优化 PCRR 值，是一个二分类问题，损失函数为交叉熵。另一个节点直接输出 RSRP 值，需要优化损失函数 RMSE。

表 13　无用特征剔除表

中文含义	代码中的变量名	剔除理由
发射机中心频率	Frequency_Band	特征重复度高
地物类型索引的损耗因子	index_delta	特征重要性低
水平距离	log_2d_len	特征重要性低
发射机垂直电下倾角	Electrical_Downtilt	被统一合成为了下倾角
发射机垂直机械下倾角	Mechanical_Downtilt	被统一合成为了下倾角

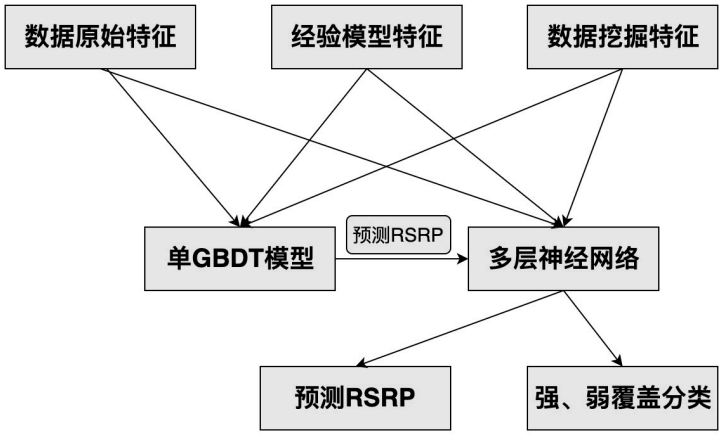

图 19　融合模型流程图

综上所述，本文用到的所有模型的线下与线上榜单得分如下表15。

表 14　无用特征剔除表

超参数名	参数值	解释
objective	regression	学习目标是回归问题
num_round	100	学习总步长
learning_rate	0.2	学习率
num_threads	40	并发数目
max_depth	4	最大树深度
bagging_fraction	0.85	数据采样（防止过拟合）
bagging_freq	5	bagging 算法频率
feature_fraction	0.9	特征采样（防止过拟合）
metric	rmse	观测指标

表 15　模型评价指标

模型信息	线下 PCRR 值	线下 RMSE	线上 RMSE 得分
模型融合 + 剔除无用特征	22.5834%	8.5807	9.11321
模型融合 + 全量特征	26.2%	8.3459	9.22589
模型融合 + 数据清洗	17.6231%	8.8729	未提交
模型融合 + 原始数据	10.3564%	9.1963	9.25693
单 GBDT 模型	12.7329%	9.1276	未提交

5. 模型评价

5.1 优点

1. 本文在进行特征提取过程中充分结合变量的实际意义，综合使用模型分析与数据分析的方法，将定性分析和定量分析充分结合，使模型更加合理。
2. 本文在进行建模的过程中，充分考虑了模型的不同使用场景，使得模型对于不同区域类型的数据均能给出较好的预测结果
3. 本文通过深刻分析讨论问题，剔除部分目标相关性较低的特征，使得模型更加简洁高效，试图用最简单的模型和算法解决了复杂的问题。

5.2 缺点

1. 本文只对模型结构进行了改进，使得传统神经网络模型能够很好地预测本题数据，由于比赛时间限制，本文仅在部分模型细节上进行创新与优化处理，没有创新性的提出新的模型与架构。
2. 本文提出的模型，尽管本文已经找到了相对最优的模型超参数以防止过拟合现象的发生，但在学习步数过多的情况下，仍存在过拟合现象。

5.3 创新点

1. 采用多目标学习的方式来训练模型，同时优化 PCRR 和 RMSE 两个目标。
2. 将原始特征与 GBDT 模型输出特征做融合，使得 TensorFlow 的模型能学到更多维度的特征。

参考文献

[1] 段宗林, WCDMA 无线信号的传播预测与研究, 2011.

[2] 移动通信系统中无线电波传播特性及高频段传播模型研究, 云南师范大学, 2018.

[3] 任佳敏, 城市环境无线传播损耗预测研究, 南京航空航天大学, 2012.

[4] 基于专用基站 BCCH 信道的传播模型校正方法研究, 2016.

[5] 无线电波传播损耗预测方法与应用研究, 南京航空航天大学, 2014.

[6] 赵育才, 无线电波传播预测与干扰分析研究及实现, 国防科学技术大学, 2009.

[7] 贾坤, 移动通信中电波传播预测模型的改进研究, 四川大学, 2003.

[8] 移动通信系统中无线信号传播损耗预测算法研究与系统实测, 电子科技大学, 2013.

[9] Ke G, Meng Q, Finley T, et al., Lightgbm: A highly efficient gradient boosting decision tree, Advances in Neural Information Processing Systems, 3146-3154, 2017.

[10] Damien A, TensorFlow-Examples, https://github.com/aymericdamien/TensorFlow-Examples, 2019.09.22.