



CPIPC

中国研究生
创新实践系列大赛



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 南京航空航天大学

参赛队号 19102870022

1. 杨凌辉

队员姓名 2. 鲍悦

3. 张嘉纹

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：

本文通过对经验传播模型与实测数据的分析，借助灰色关联分析、回归分析、假设检验工具进行特征工程中的特征设计；利用过滤法、基于 BP 神经网络的权值分析法等进行特征工程中的特征选择；根据上述步骤建立无线传播模型特征集以及训练数据集，数据清洗过程旨在为深度神经网络提供更加可靠的训练标签，随后建立基于 TensorFlow 框架的深度神经网络无线传播模型来对不同地理位置的平均信号接收功率(RSRP)进行预测。结果显示，训练模型在未参与训练的测试数据集上表现出较好的 RSRP 结果预测，相关 RMSE 指标为 11.64，说明结果具有一定的可靠性。

针对问题一：

1、通过分析赛题提供的训练数据集信息，总结数据集提供的各参数含义，并对训练集提供的全部实测数据借助灰色关联分析，判断各参数与标签数据的关联性。

2、研究分析自由空间传播模型、经验传播模型（COST231-Hata 模型），提取模型中涉及的各部分参数含义以及其对模型的影响。

3、基于实测数据，随机选取部分训练集文件数据，通过线性关系分析确定链路距离、链路三维距离、地物类型索引等特征。并基于经验传播模型设计可能与链路损耗有关的特征。最终设计 8 个可能的特征值：链路距离、链路三维距离、载波频率、发射天线有效高度、信号线相对栅格高度、用户天线高度、地物类型索引、小区发射机水平方向角。

针对问题二：

1、通过对数据做回归分析、假设检验等初步筛选特征。这一步主要分析了链路距离、信号线相对栅格高度、地物类型索引之间的相互关系及其与链路损耗之间的关系。

2、结合实测数据分析提出的 8 个特征对结果是否有明显影响。其原则是尽量不错过一个可能有用的特征，但是也不滥用太多的特征。可以选择过滤法，如方差筛选法、相关系数法，以及基于 BP 神经网络的权值分析法分别对特征值与结果的相关性进行排序。筛选法中，我们筛选除了链路三维距离这一特征，并结合 BP 神经网络的权值排除了小区发射机水平方向角特征。

3、通过对所有检验结果检验统计量对相关关系进行量化、排序，并通过理论与实际的结合，对量化结果进行评价。

针对问题三：

1、通过将问题一和问题二设计和选择的有效特征进行数据预处理后作为深度学习的数据集。并将已知的数据集拆分为训练集、测试集以及验证集。

2、利用 TensorFlow 框架搭建深层神经网络 DNN 模型，并初始化数据集和创建会话。利用训练数据集进行无线传播模型的训练。

3、利用建立好的无线智能传播模型对不同地理位置的 RSRP 进行预测并给出对比结果。通过结果分析得到模型预测数据与实测数据基本一致，从而验证了该模型的有效性、可靠性。

关键词：无线传播模型、灰色关联分析、回归分析、假设检验、BP 神经网络、DNN 网络

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。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。

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。

1.2 本文所需解决的问题

在赛题需对机器学习的工作方式有一定掌握并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（不限定只使用这种方法）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。

【问题一】特征工程中的特征设计：

高效的机器学习模型建立依赖于输入变量与问题目标的强相关性，因此输入变量也称为“特征”。充分利用现有的无线传播模型相关专业知识以及赛题提供的训练数据集信息设计合适的无线传播模型特征，并阐述原因。

【问题二】特征工程中的特征选择：

基于提供的各小区数据集，设计多个合适的无线传播模型特征，计算这些特征与目标的相关性，选择有意义的特征输入机器学习模型进行训练，并将结果量化、排序，形成表格，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：

根据问题二建立的无线传播模型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平均信号接收功率(RSRP)进行预测。系统将利用弱覆盖识别率(PCRR)及均方根误差(RMSE)对模型进行评估。

二、基本假设

- 假设 1：数据集及本文查询的其他相关数据是真实可靠的；
- 假设 2：数据集各参数适用于相对常见天气，不考虑极端恶劣天气带来的偶然性；
- 假设 3：数据集中的各小区工程参数及地图数据包含了大部分实际应用场景；
- 假设 4：数据集给出的参数之外的其他因素对无线传播模型无影响；
- 假设 5：各小区发射机之间的相互影响忽略不计；
- 假设 6：发射端与接收端的天线增益影响忽略不计；
- 假设 7：接收到的信号仅来自于数据集提供的发射机，不会来自于其他类型信号基站；
- 假设 8：不考虑移动台天线高度的差异；
- 假设 9：RSRP 的特征选取与建模即对链路损耗 P_{loss} 的特征选取与建模，特征选取不考虑以小区发射机发射功率 P_t 作备选项；

三、问题分析

3.1 问题一的分析

问题一要求充分利用提供的无线传播模型相关专业知识以及赛题提供的训练数据集信息设计合适的无线传播模型特征。

首先，应分析赛题提供的训练数据集信息，总结数据集提供的各参数含义，并对训练集提供的全部实测数据借助灰色关联分析，判断各参数与标签数据的关联性。

其次，研究分析经验传播模型：COST231-Hata 模型。由于自由空间传播模型是无线电波传播的研究中最简单的模型，也是所有经验传播模型的基石，其他的经验传播模型都是在此基础上通过参数修正而得到的，为了匹配特定的地理环境，每个传播模型都有各自的传播路径损耗经验公式。因此可以依据自由空间传播模型，结合 COST231-Hata 模型简化分析，模型中涉及的各部分参数含义以及其对模型的影响。

最后，分别基于实测数据与经验传播模型通过设计可能与链路损耗有关的特征。

3.2 问题二的分析

问题二要求基于提供的各小区数据集，设计多个合适的无线传播模型特征，并计算这些特征与目标的相关性，并将结果量化、排序，形成表格。

首先，通过对数据做回归分析、假设检验等初步筛选特征。

接着，结合实测数据分析前面提出的特征对结果是否有明显影响。其原则是尽量不错过一个可能有用的特征，但是也不滥用太多的特征。可以选择过滤法、基于 BP 神经网络的权值分析法分析自变量与目标变量之间的关联选择特征：

1、方差筛选法：方差衡量的是一个随机变量取值的分散程度。如果一个随机变量的方差非常小，那这个变量作为输入，是很难对输出有什么影响的。在进行特征选择时，可以丢弃那些方差特别小的特征。

2、相关系数法：相关系数表征的是两个随机变量之间的线性相关关系。特征与输出的相关系数的绝对值越大，说明对输出的影响越大，应该优先选择。

3、基于 BP 神经网络的权值分析法：BP 神经网络是一种具有三层或者三层以上的多层神经网络，不断修正误差逆传播训练，得到不同神经元的权值，权值越大，对输出的影响越大。

最后，通过对所有检验结果检验统计量对相关关系进行量化、排序，并结合理论与实际，对量化结果进行评价。

3.3 问题三的分析

问题三要求我们在设计和选择有效的特征之后，根据已知的小区数据以及地形数据建立基于 AI 的无线传播模型来对不同地理位置的 RSRP 进行预测。

我们应将 TensorFlow 框架与深度神经网络(Deep Neural Networks，简称 DNN)模型相结合来预测目标数值，评估通过模型所得到的预测值与真实值是否一样。

首先将已有的数据根据前述问题的特征选取进行数据预处理并导入 TensorFlow 框架，其次在 TensorFlow 框架中设置模型参数，通过加载的数据集建立神经网络层。经过训练后进行预测与实际 RSRP 的对比从而分析数据得到结论。

四、问题一的模型建立与求解

4.1 解题思路概述

高效的机器学习模型建立依赖于输入变量与问题目标的强相关性，因此输入变量也称为“特征”。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。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。

本题要求充分利用提供的无线传播模型相关专业知识以及赛题提供的训练数据集信息设计合适的无线传播模型特征。

首先，应分析赛题提供的训练数据集信息，总结数据集提供的各参数含义，并对训练集提供的全部实测数据借助灰色关联分析，判断各参数与标签数据的关联性。

其次，研究分析经验传播模型：COST231-Hata 模型。由于自由空间传播模型是无线电波传播的研究中最简单的模型，也是所有经验传播模型的基石，其他的经验传播模型都是在此基础上通过参数修正而得到的，为了匹配特定的地理环境，每个传播模型都有各自的传播路径损耗经验公式。因此可以依据自由空间传播模型，结合 COST231-Hata 模型简化分析，模型中涉及的各部分参数含义以及其对模型的影响。

最后，分别基于实测数据与经验传播模型理论设计可能与链路损耗有关的特征。

4.2 训练数据集信息

4.2.1 训练数据集字段含义解释

表 4.1 样本基本信息示例

| 工程参数数据 | | | | | | | | |
|---------------|----------------------|--------------------|--------|---------|---------------------|---------------------|----------------|----------|
| Cell Index | Cell X | Cell Y | Height | Azimuth | Electrical Downtilt | Mechanical Downtilt | Frequency Band | RS Power |
| 1003501 | 393621.9 | 3394449 | 35 | 300 | 6 | 4 | 2585 | 13.2 |
| 地图数据 | | | | | | | | |
| Cell Altitude | Cell Building Height | Cell Clutter Index | X | Y | Altitude | Building Height | Clutter Index | |
| 524 | 32 | 1 | 392800 | 3395210 | 524 | 0 | 5 | |
| RSRP 标签数据 | | | | | | | | |
| RSRP | | | | | | | | |
| -90.5 | | | | | | | | |

表 4.2 训练集数据的字段含义

| 字段名称 | 含义 | 单位 |
|------------|------------------|-----|
| 工程参数 | | |
| Cell Index | 小区唯一标识 | - |
| Cell X | 小区所属站点的栅格位置，X 坐标 | - |
| Cell Y | 小区所属站点的栅格位置，Y 坐标 | - |
| Height | 小区发射机相对地面的高度 | m |
| Azimuth | 小区发射机水平方向角 | Deg |

| | | |
|----------------------|---|-----|
| Electrical Downtilt | 小区发射机垂直电下倾角 | Deg |
| Mechanical Downtilt | 小区发射机垂直机械下倾角 | Deg |
| Frequency Band | 小区发射机中心频率 | MHz |
| RS Power | 小区发射机发射功率 | dBm |
| 地图数据 | | |
| Cell Building Height | 小区站点所在栅格(Cell X, Cell Y)的建筑物高度, 若该栅格没有建筑物, 则为 0 | m |
| Cell Altitude | 小区站点所在栅格(Cell X, Cell Y)的海拔高度 | m |
| Cell Clutter Index | 小区站点所在栅格(Cell X, Cell Y)的地物类型索引 | - |
| X | 栅格位置, X 坐标 | - |
| Y | 栅格位置, Y 坐标 | - |
| Building Height | 栅格(X,Y)上的建筑物高度, 若该栅格没有建筑物, 则为 0 | m |
| Altitude | 栅格(X,Y)上的海拔高度 | m |
| Clutter Index | 栅格(X,Y)上的地物类型索引 | - |
| RSRP 标签数据 | | |
| RSRP | 栅格(X, Y)的平均信号接收功率, 标签列 | dBm |

表 4.3 地物类型索引编号含义

| 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 |

4.2.2 训练数据集预处理

(a) 小区发射机中心频率大小统计

数据集提供 4000 组数据, 其中单一文件中的工程参数数据无任何区别, 因此我们统计了所有文件中的小区发射机中心频率大小。

表 4.4 所有文件中的小区发射机中心频率大小统计

| | | | |
|----------------------|----------|------|--------|
| 小区发射机中心频率大小 (MHz) | 2585 | 2604 | 2624 |
| 频数 | 11721977 | 7735 | 212505 |

由上表可知，数据集提供了 3 种不同中心频率，分别为 2585MHz、2694MHz、2624MHz，他们都属于 D 波段频率。其中数据主要集中 2585MHz，这是 D 波段中心频段。

(b) 地物类型索引号统计

表 4.5 所有文件中的地物类型索引号统计

| 地物类型索引号 | 频数 | 地物类型索引号 | 频数 |
|---------|---------|---------|--------|
| 1 | 0 | 11 | 151335 |
| 2 | 157233 | 12 | 718328 |
| 3 | 0 | 13 | 776184 |
| 4 | 0 | 14 | 368303 |
| 5 | 6159532 | 15 | 135082 |
| 6 | 2397951 | 16 | 12374 |
| 7 | 739713 | 17 | 13714 |
| 8 | 108659 | 18 | 4492 |
| 9 | 0 | 19 | 0 |
| 10 | 268933 | 20 | 0 |

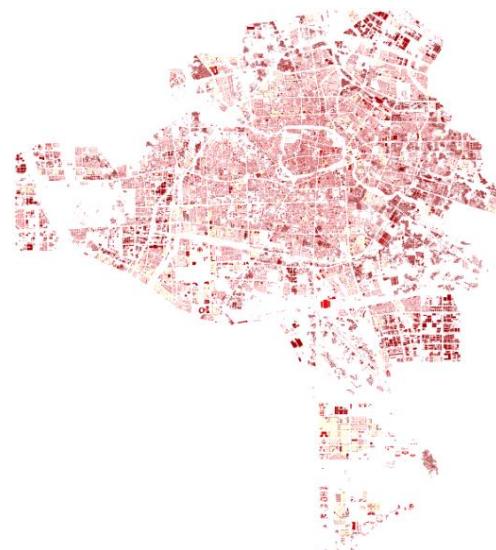
由上表可知，数据集提供的地物类型共 14 种，为了更直观了解地物类型，我们绘制了条形图加以分析。



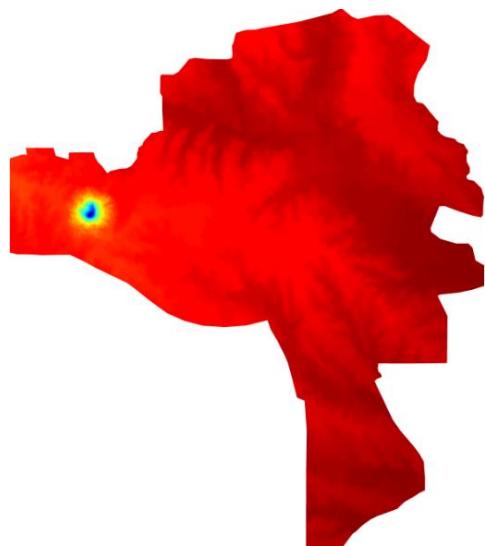
图 4.1 地物类型索引号频数条形图

从图中我们明显可以看出市区开阔区域的数据量最大，其次是道路开阔区域。城区高、中高、超高层建筑，植被覆盖区数据量也较大。这些地区也是人类活动常发生地区，所给数据明显站在设备供应商以及无线运营者的角度提供。

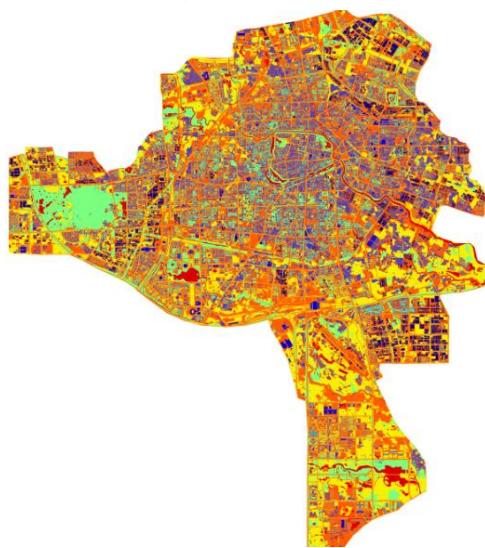
(c) 根据栅格坐标以及房屋高度、海拔高度和地物类型索引等参数做可视化处理。



(a) 建筑物高度



(b) 海拔高度



(c) 地物类型索引

图 4.2 训练集数据电子地图图像化示意

如图 4.2 所示，数据集给出的城市地形复杂，地势起伏较大，不同地区房屋疏密程度不同，市中心房屋较为密集，城区道路较多呈网格状分布。

4.2.3 灰色关联分析

灰色关联分析方法可以分析我们找到的各个特征对目标的影响程度，从而达到对各个特征定量描述与比较的目的。该算法的核心是一定规则确立随时间变化的母序列，把各个特征参数随时间的变化作为子序列，求各个子序列与母序列的相关程度，依照相关性大小得出结论。

灰色关联分析具体计算方法：

1、确定分析数列：将我们的目标链路损耗作为参考数列，将我们找到的 8 种特征参数组成的数据序列作为比较数列。

(1)参考数列(又称母序列) $Y=Y(k)|k=1,2,\cdots,n;$

(2)比较数列(又称子序列)为 $X_i=X_i(k)|k=1,2,\cdots,n,i=1,2,\cdots,m$

2、变量的无量纲化

由于系统中各因素列中的数据可能因量纲不同，不便于比较或在比较时难以得到正确的结论。因此在进行灰色关联度分析时，一般都要进行数据的无量纲化处理。主要有一下两种方法。

(1)初值化处理：

$$x_i(k) = \frac{x_i(k)}{x_i(1)}, k = 1, 2, \dots, n; i = 0, 1, 2, \dots, m \quad (4.1)$$

(2)均质化处理

$$x_i(k) = \frac{x_i(1)}{x_i}, k = 1, 2, \dots, n; i = 0, 1, 2, \dots, m \quad (4.2)$$

其中 k 对应时间段， i 对应比较数列中的一行（即一个特征）。

3、计算关系系数

$$\xi_i(k) = \frac{\min_i \min_k |y(k) - x_i(k)| + \rho \max_i \max_k |y(k) - x_i(k)|}{|y(k) - x_i(k)| + \rho \max_i \max_k |y(k) - x_i(k)|}, \rho \in (0, \infty) \quad (4.3)$$

ρ 称为分辨系数。 ρ 越小，分辨力越大，一般 ρ 的取值区间为 $(0, 1)$ 具体取值可视情况而定，当 $\rho \leq 0.5463$ 时，分辨力最好，通常取 $\rho=0.5$ 。

4、计算关联度

因为关联系数是比较数列与参考数列在各个时刻（即曲线中的各点）的关联程度值，所以它的数不止一个，而信息过于分散不便于进行整体性比较。因此有必要将各个时刻（即曲线中的各点）的关联系数集中为一个值，即求其平均值，作为比较数列与参考数列间关联程度的数量表示，关联度 r_i 公式如下：

$$r_i = \frac{1}{n} \sum_{k=1}^n \xi_i(k), k = 1, 2, \dots, n \quad (4.4)$$

5、关联度排序

关联度按大小排序，如果 $r_1 < r_2$ ，则参考数列 y 与比较数列 x_2 更相似。

在算出 $X_i(k)$ 序列与 $Y(k)$ 序列的关联系数后，计算各类关联系数的平均值，平均值 r_i 就称为 $Y(k)$ 与 $X_i(k)$ 的关联度。

我们对原始训练集数据与链路损耗（RS-RSRP）功率指标直接进行灰色关联分析，结

果如下表：

表 4.6 未处理的训练集数据灰色关联分析结果

| 字段名称 | 关联度 |
|----------------------|----------|
| Cell X | 0.998503 |
| Cell Y | 0.998526 |
| Height | 0.994304 |
| Azimuth | 0.991074 |
| Electrical Downtilt | 0.993435 |
| Mechanical Downtilt | 0.991302 |
| Frequency Band | 0.998525 |
| Cell Building Height | 0.998505 |
| Cell Altitude | 0.977106 |
| X | 0.998503 |
| Y | 0.998526 |
| Building Height | 0.998505 |
| Altitude | 0.973580 |

表中显示的关联度均接近于 1，我们认为训练集提供的各参数与链路损耗间都有明显相关性。RS-RSRP 作为参考序列，除地物类型索引和小区编号外，其余指标作为比较序列。

4.3 无线传播模型理论分析

4.3.1 自由空间传播模型

在自由空间（即无任何衰减、无任何阻挡、无任何多径的传播空间）中，无线电波的传播不受阻挡，不会出现反射、折射、散射或是绕射等现象，然而电波在路径传播中，其能量仍然是会衰减的，产生衰减的原因是辐射的能量会产生扩散。因此，电波在自由空间中传播时，其单位面积内的能量因为扩散而逐渐减少，将此称为自由空间传播损耗^[1]。自由空间传播的示意图如图 4.3 所示。

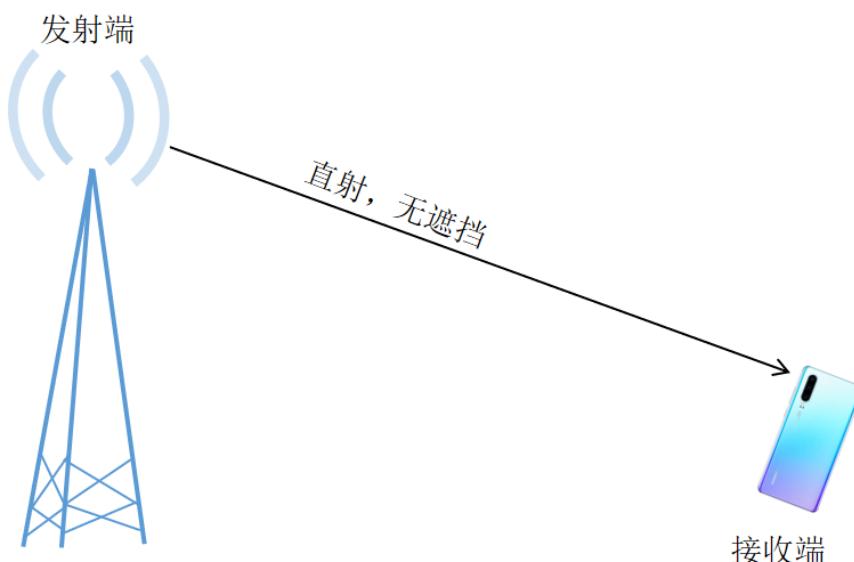


图 4.3 自由空间中电波传播示意图

自由空间中，电波传播损耗公式如下给出：

$$P_{loss}(dB) = 10 \lg \frac{P_t}{P_r} = -10 \lg \frac{G_t G_r c^2}{(4\pi)^2 d^2 f^2} \quad (4.5)$$

式中， P_t 为发射功率， P_r 为接收功率， d 为发射端与接收端的水平距离， c 为光在真空中的传播速度， f 为发射信号的频率， G_t 、 G_r 分别为发射端和接收端的天线增益。

在排除发射端和接收端的天线增益及设备损耗的情况下，自由空间传播损耗模型的公式可以简化为：

$$P_{loss}(dB) = 10 \log \frac{P_t}{P_r} = -10 \log \frac{G_t G_r c^2}{(4\pi)^2 d^2 f^2} = 32.44 + 20 \log f + 20 \log d \quad (4.6)$$

自由空间中，路径损耗与工作频率和传播距离的关系如图 4.4 所示。

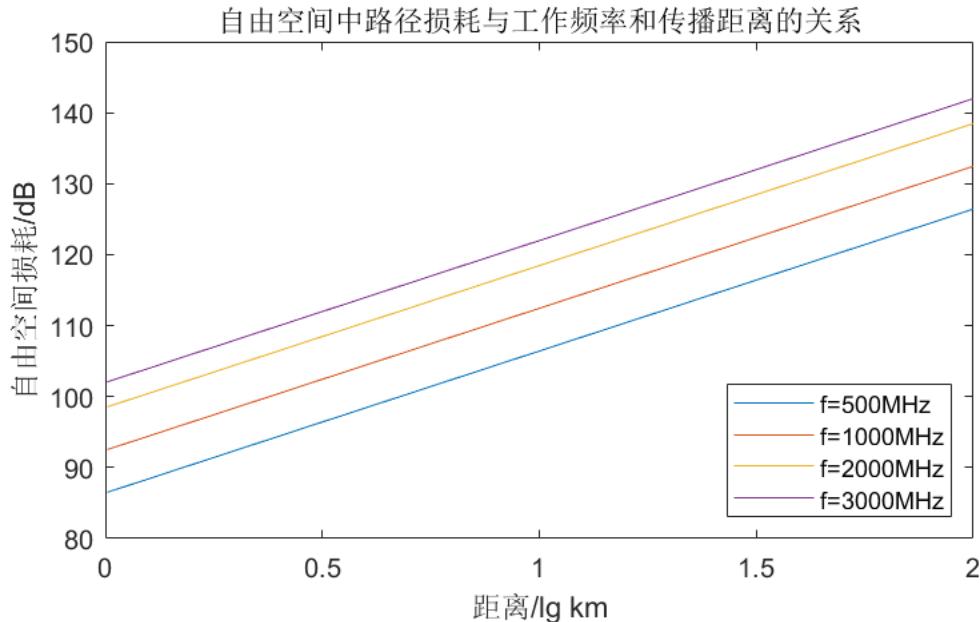


图 4.4 自由空间中路径损耗与工作频率和传播距离的关系

在自由空间中，电波传播的路径损耗只与传播距离和工作频率有关。当工作频率一定时，电波的传播距离越远，其路径损耗就越大；而传播距离一定时，工作频率越高，其路径损耗也就越大。

4.3.2 COST231-Hata 传播模型

在现实环境中，由于传播路径上存在着各种影响，如高空电离层影响，高山、湖泊、海洋、地面建筑、植被以及地球曲面的影响等，因而电磁波具有反射、绕射、散射和波导传播等比自由空间复杂得多的传播方式。

Cost 231-Hata 模型的路径损耗的计算经验公式为：

$$P_{loss}(dB) = 46.3 + 33.9 \lg f - 13.82 \lg h_b - \alpha + (44.9 - 6.55 \lg h_b) \lg d + C_{cell} + C_{landform} + C_B \quad (4.7)$$

其中 P_{loss} 定义为传播路径损耗(dB)、 f 为载波频率(MHz)、 h_b 为基站天线有效高度(m)、 α 为用户天线高度纠正项(dB)、 d 为链路距离(km)以及 C_m 为场景纠正常数(dB)。

如果我们采用校正过的传播模型，则可以提高覆盖预测的准确程度，并且可以针对特定的具体的传播环境，设计出相对准确的无线网络规划的建设方案。从而可以达到充分合理的利用系统的资源，提高网络的质量。

α 为移动接收台有效天线修正因子，为一个和覆盖区域相关的函数。 C_{cell} 根据小区覆盖场景校正因子，小区覆盖场景不一致取值也一致。

$$\alpha = \begin{cases} (1.1 \lg f - 0.7) h_{au} - (1.56 \lg f - 0.8) & \text{中小城市} \\ 8.29(\lg 1.54 h_{au})^2 - 1.1 & \text{大城市, } f \leq 300 MHz \\ 3.2(\lg 1.54 h_{au}) - 4.97 & \text{大城市, } f \leq 300 MHz \end{cases} \quad (4.8)$$

C_B 是大城市中心校正因子。

$$C_B = \begin{cases} 0(dB) & \text{中等城市和郊区} \\ 3(dB) & \text{大城市中心} \end{cases} \quad (4.9)$$

$C_{landform}$ 为地形校正因子，它主要是为了反映出一些特殊地形地貌对于传播损耗造成的结果。如果想取得适用的地形校正因子，能够凭借对现有传播模型进行试验和修正，也能够适当的取一些合适的参数^[2]。

$$C_{landform} = \begin{cases} 0 & \text{城市} \\ -2[\lg f / 28]^2 - 5.4 & \text{郊区} \\ -4.78(\lg f) - 18.33\lg f - 40.98 & \text{乡村} \end{cases} \quad (4.10)$$

对模型中涉及的基本参数分别对单一变量进行对照分析，定性分析理论中路径损耗与各参数的关系。

1、相同参数的基站天线在不同的环境类型中，Cost 231-Hata 模型路径损耗曲线如图 4.5 所示：

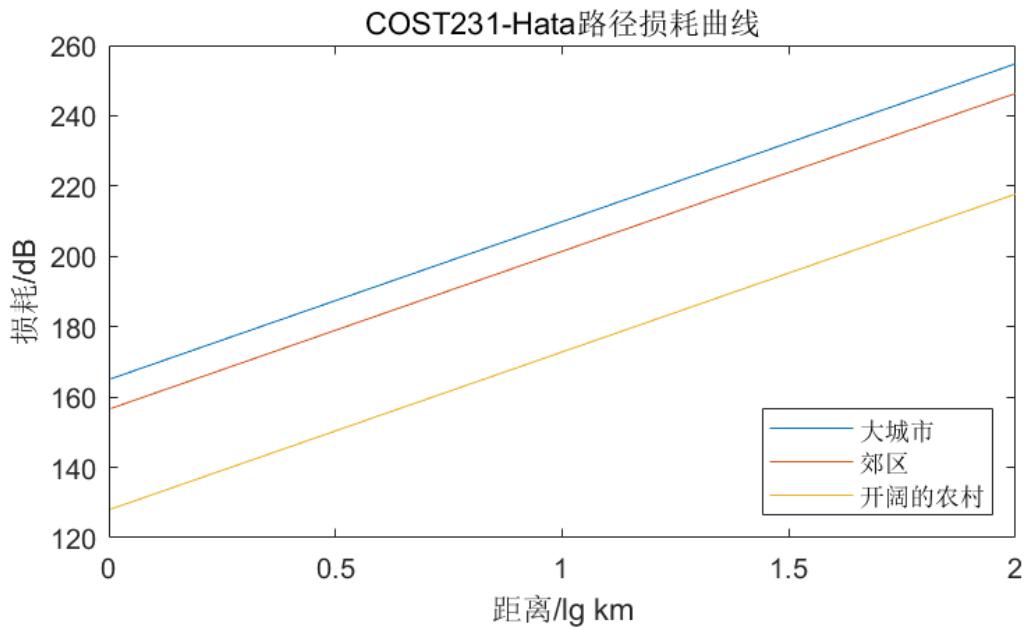


图 4.5 不同的环境类型 Cost 231-Hata 模型路径损耗曲线

在发射天线高度、接收天线高度以及距离不变时，传播路径损耗在不同环境下有明显差异。也就是说，电磁波在不同环境下的反射、绕射、散射和波导传播也有所差异。

2、大城市环境中，不同的基站天线有效高度条件下，Cost 231-Hata 模型路径损耗曲线如图 4.6 所示：

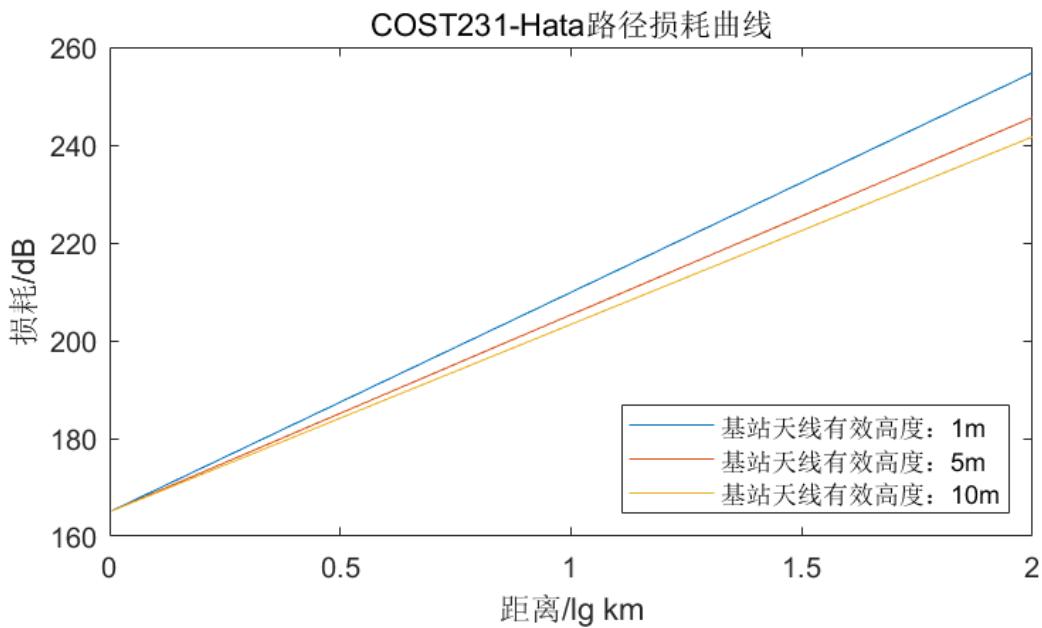


图 4.6 不同基站天线有效高度 Cost 231-Hata 模型路径损耗曲线

在应用环境、接收天线高度以及距离不变时，传播路径损耗随着基站天线有效高度的增加而降低；同时损耗与基站天线有效高度、链路距离应存在二维耦合关系。

3、大城市环境中，相同基站条件下，对不同移动平台天线高度的 Cost 231-Hata 模型路径损耗曲线如图 4.7 所示：

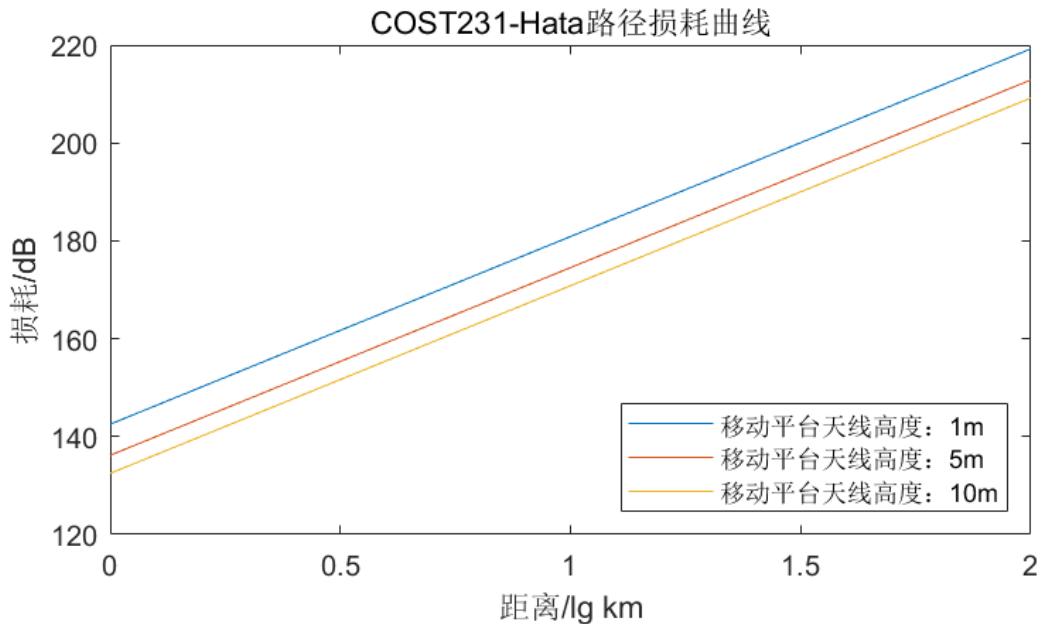


图 4.7 不同移动平台天线高度 Cost 231-Hata 模型路径损耗曲线

在应用环境、基站基本参数以及距离不变时，传播路径损耗随着移动平台高度的增加而降低。

为判断 COST321-Hata 模型的合理性，我们选取两个具有代表性的地形场景将理论模型与实际数据相对比。这里我们选取市区开阔场景和内陆湖泊场景，发现理论值与实测数据基本一致，RMSE 指标分别为 13.5 和 12.3。因此，我们可以借助 COST321-Hata 模型选取特征。

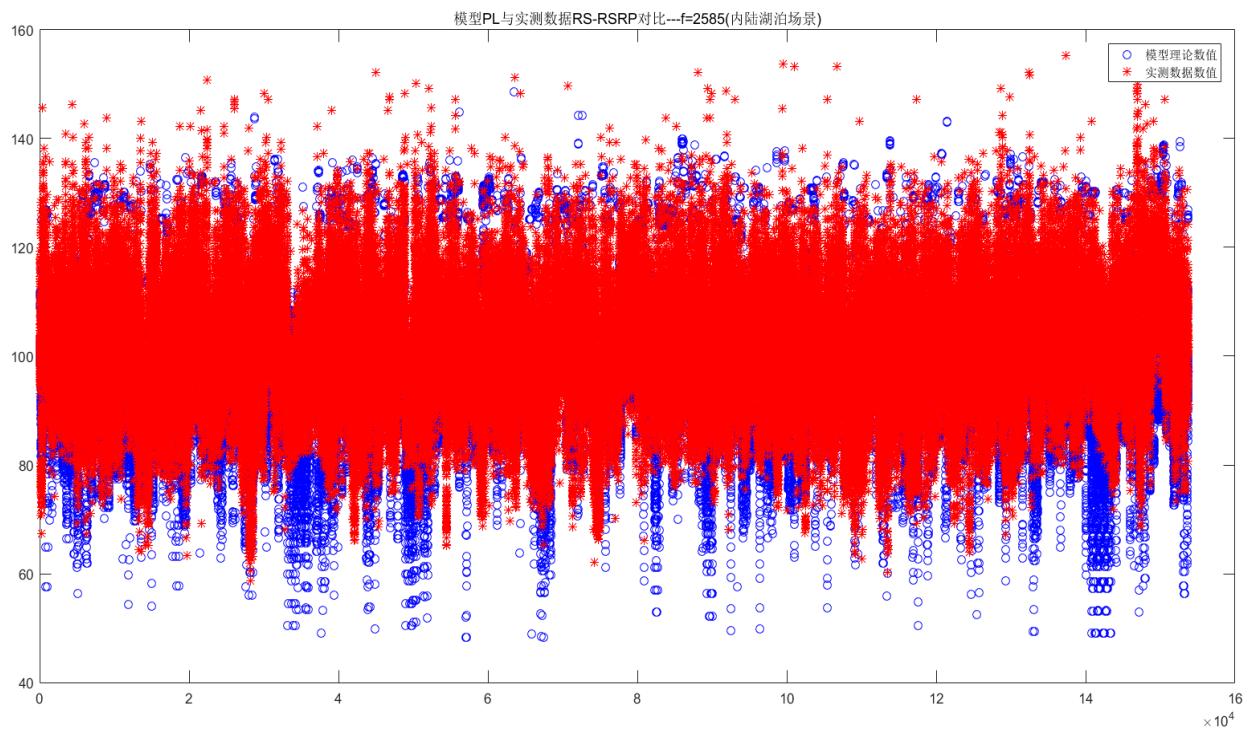
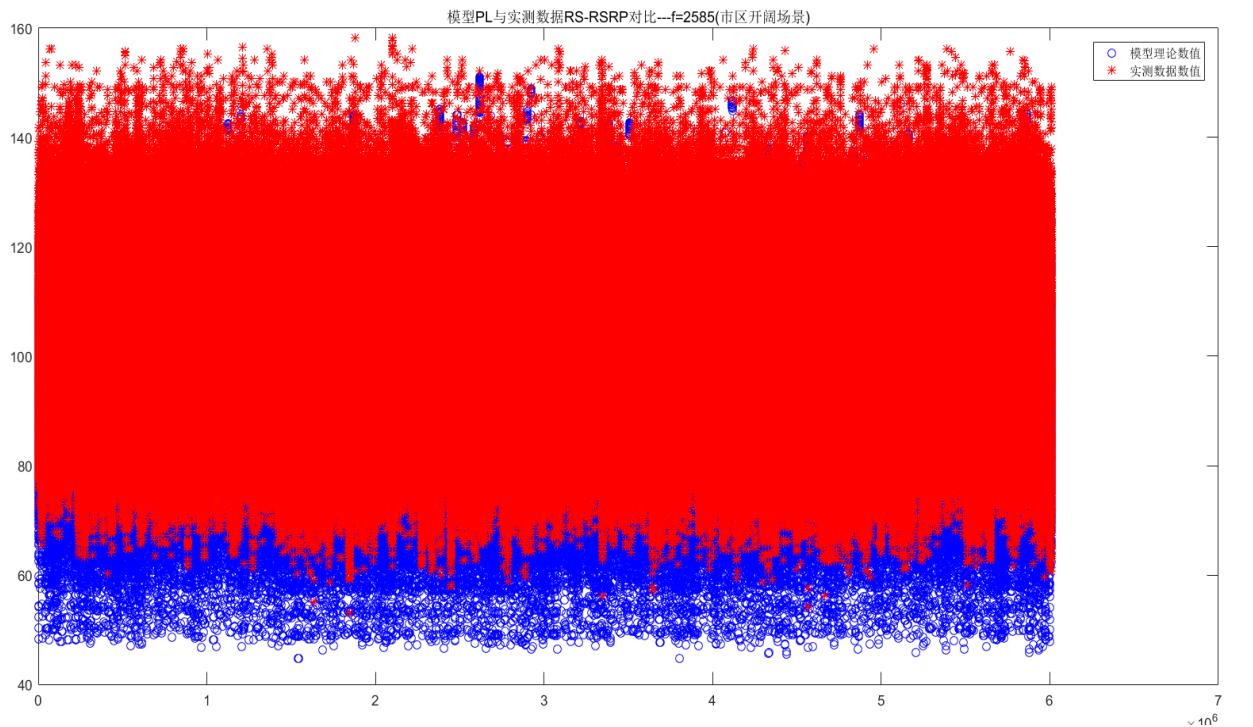


图 4.8 理论值与实测数据对比

由经验传播模型我们可以看到，各参数实际上是通过对链路损耗 P_{loss} 的影响间接影响 RSRP 值，因此以下讨论我们均分析特征与 P_{loss} 的关系，不再将小区发射机发射功率 RS Power 作为特征量分析。

4.4 特征设计

4.4.1 基于实测数据的特征设计

由 4.3 我们可以发现一个与传播路径损耗明显相关的参数 d , 即链路距离。如图所示:

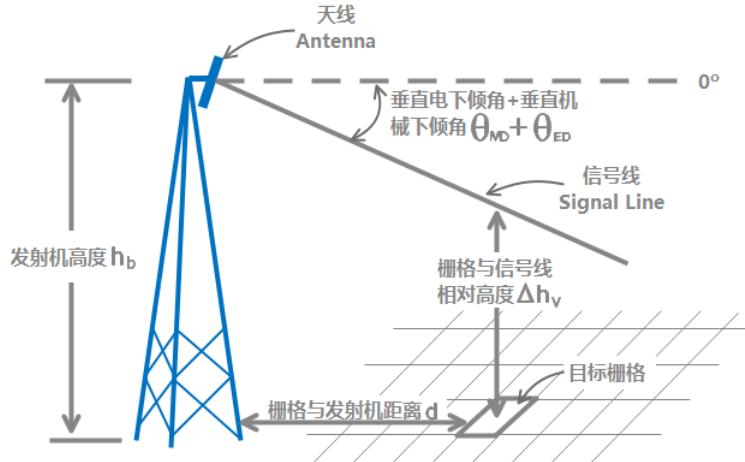


图 4.9 目标栅格与发射机的地理位置关系

从图中我们可以看出, 链路距离及栅格与发射机距离, 可以通过目标栅格坐标(X,Y)与发射机所在栅格坐标(Cell X,Cell Y)得到。其表达式为:

$$d = \sqrt{(X - CellX)^2 + (Y - CellY)^2} \quad (4.11)$$

为了验证理论关系的正确性, 不失一般性, 我们选取‘train_1001701.csv’文件中的数据进行分析。其工程参数数据如下表所示:

表 4.7 ‘train_1001701.csv’文件工程参数数据

| ‘train_1001701.csv’文件中的工程参数数据 | | | | | | | | |
|-------------------------------|--------|---------|--------|---------|---------------------|---------------------|----------------|----------|
| Cell Index | Cell X | Cell Y | Height | Azimuth | Electrical Downtilt | Mechanical Downtilt | Frequency Band | RS Power |
| 1001701 | 424515 | 3376325 | 24 | 300 | 6 | 3 | 2585 | 16.2 |

文件中的工程参数数据均一致, 即发射天线有效高度、载波频率是一致的, 便于我们对单一变量的讨论分析。

‘train_1001701.csv’文件中共有 2207 个数据, 我们首先分析其中地物类型索引号分布, 并绘制了不同地物类型索引号上的距离统计图:

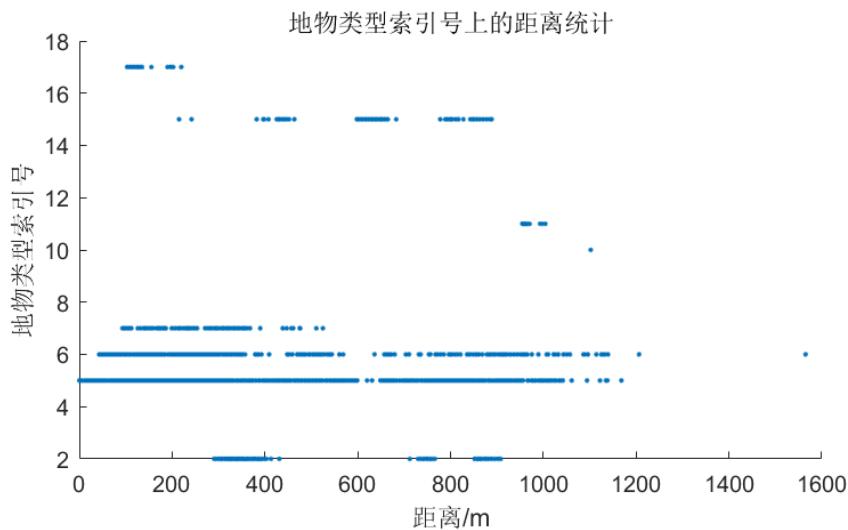


图 4.10 ‘train_1001701.csv’文件不同地物类型索引号上的距离统计

由图 4.10 可知，索引号 5 的数据出现次数最多为 1319 次，占总数据量的 59.76%，为市区开阔区域，所受干扰较小；链路距离范围较广；同时建筑物海拔高度相对集中仅有一个异常值，因此用户天线高度几乎无差异。因此该文件中索引号 5 的数据可认为仅有链路距离一个自变量，可以进行单一自变量分析。

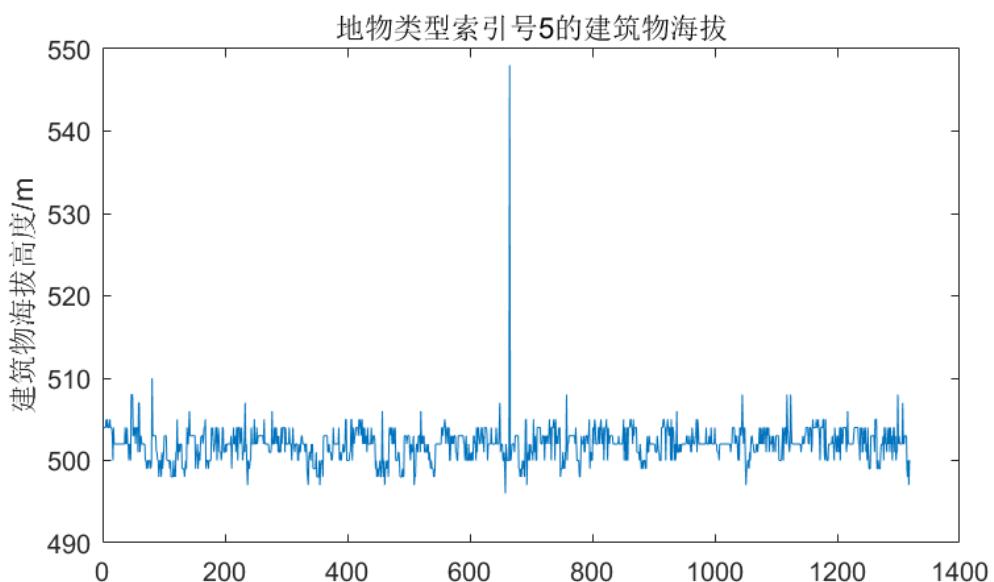


图 4.11 ‘train_1001701.csv’文件地物类型索引号 5 的建筑物海拔分析

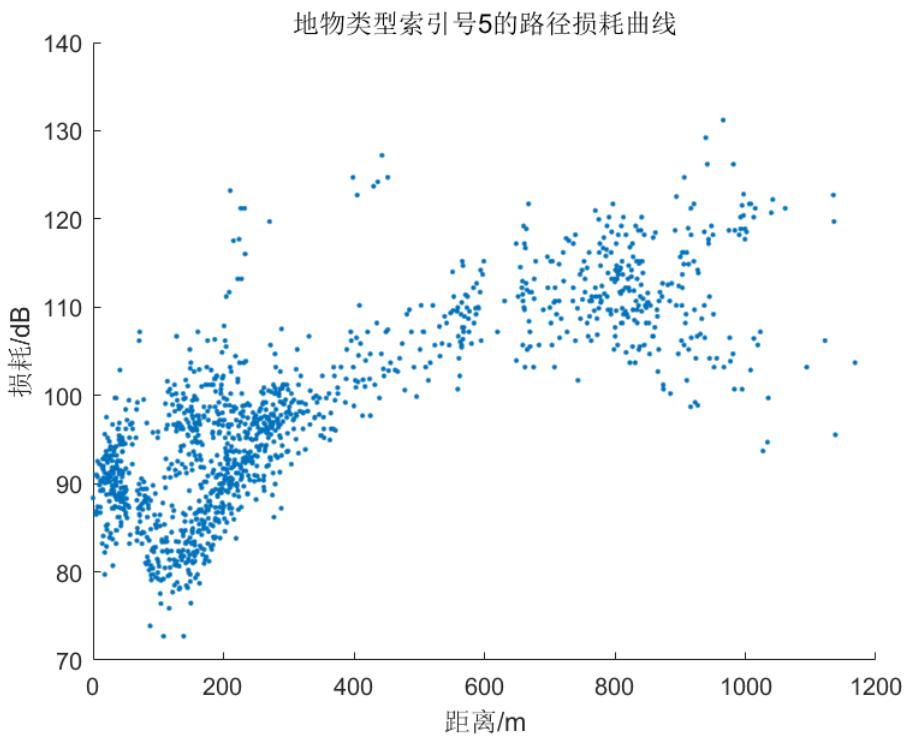
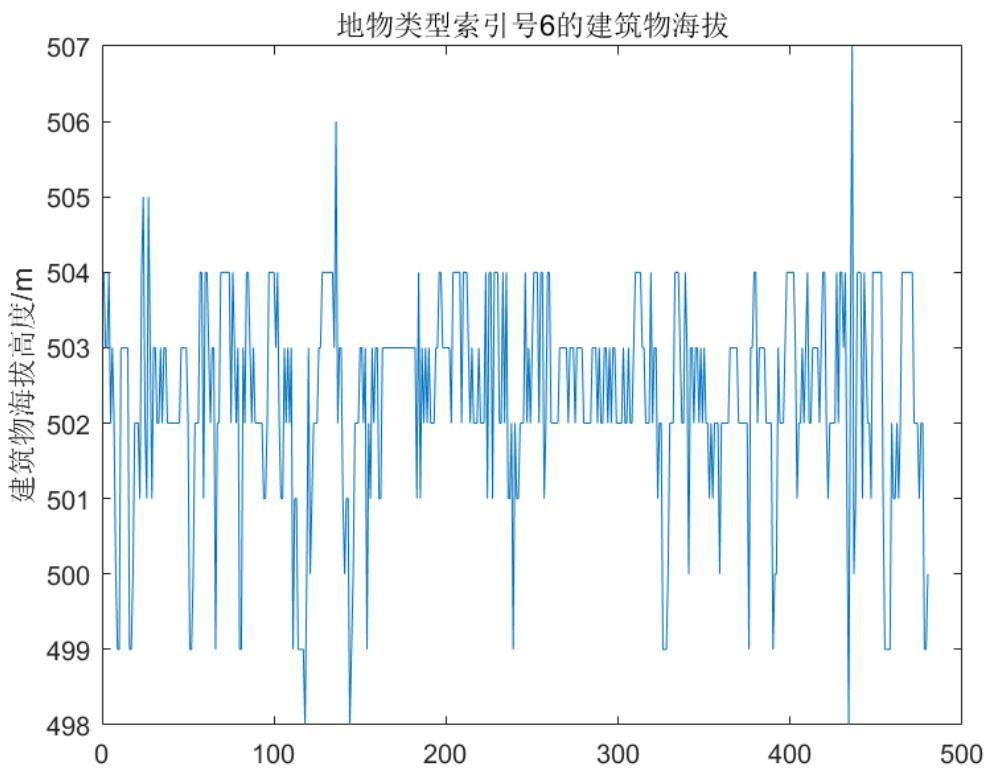


图 4.12 ‘train_1001701.csv’文件地物类型索引号 5 的路径损耗散点图

如图 4.12 所示，在基站工程参数数据一致的情况下，相同环境中的电波传播的路径损耗与传播距离有关，误差允许的范围内，电波的传播距离越远，其路径损耗就越大。

为排除偶然性，我们对索引号数据出现次数较多的 6 及其他文件数据（‘train_1095101.csv’）也进行了观察：



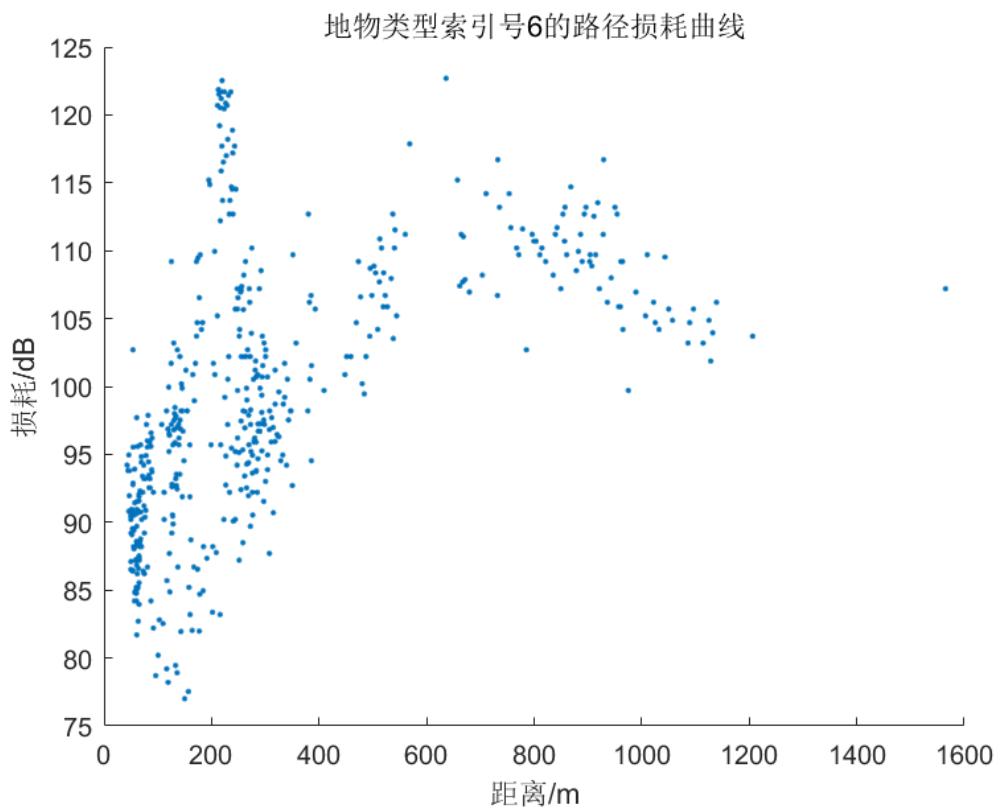
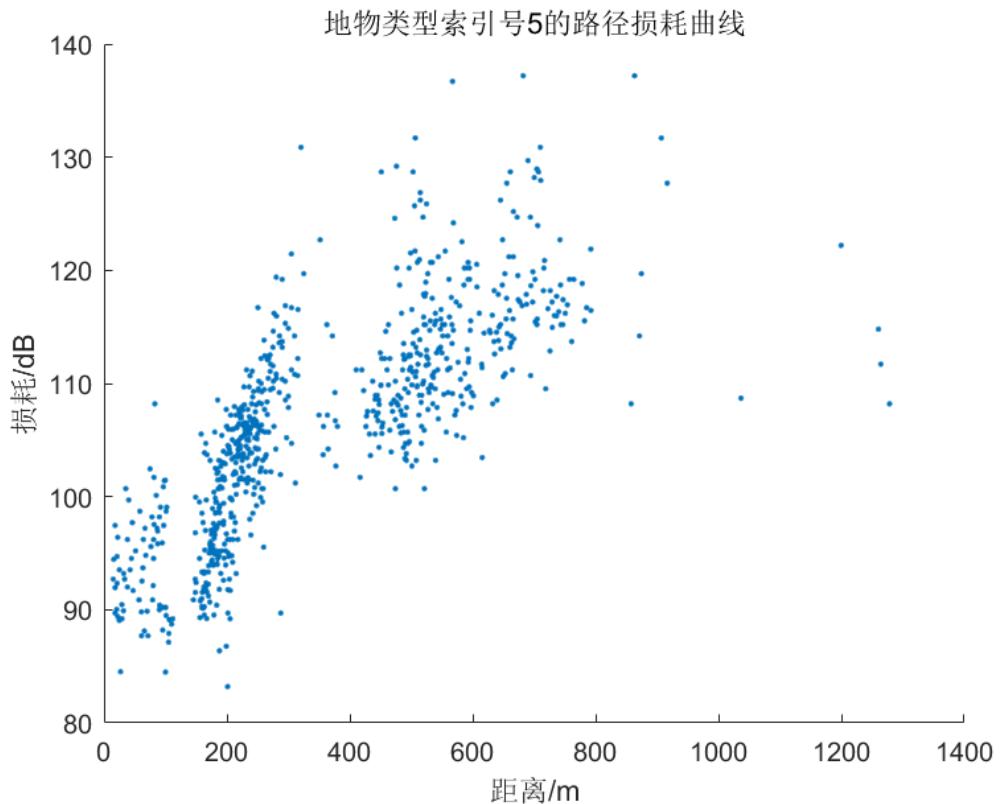


图 4.13 ‘train_1001701.csv’文件地物类型索引号 6 的数据分析

以下图形为‘train_1095101.csv’文件中不同地物索引类型及其对应路径损耗曲线。不同文件作为不同基站工程参数对比分析。



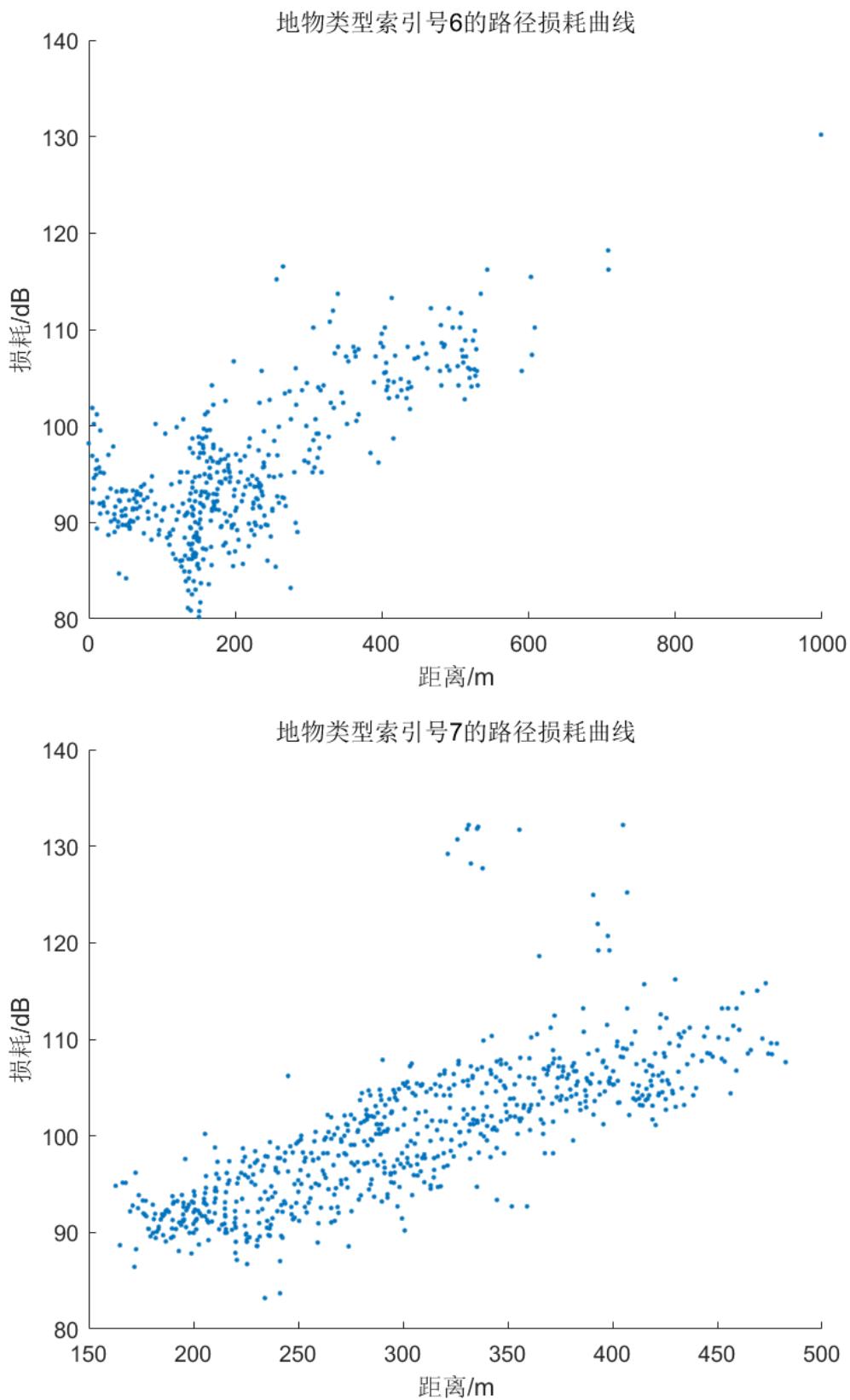


图 4.14 ‘train_1095101.csv’文件的数据分析

综上所述，基站工程参数一致、环境相同时的电波传播的路径损耗与链路距离 d 有关， d 越大，其路径损耗就越大。因此，我们认为链路距离 d 是一个合适的工程特征。其相关关系在问题二中通过引入其他变量进行详细分析。

4.4.2 基于经验传播模型的特征设计

由 4.3 可知，链路距离 d 外，载波频率 f 、基站天线有效高度 h_b 、用户天线高度 h_{ue} 、场景类型（即地物类型）Clutter Index 均与损耗 P_{loss} 有关。而 4.2 节工程数据中提供了三个角度：小区发射机水平方向角（记为： θ_{AZ} ）、小区发射机垂直电下倾角（记为： θ_{ED} ）、小区发射机垂直机械下倾角（记为： θ_{MD} ）。如图 4.15 所示，这三个角度可以确定天线的方向。

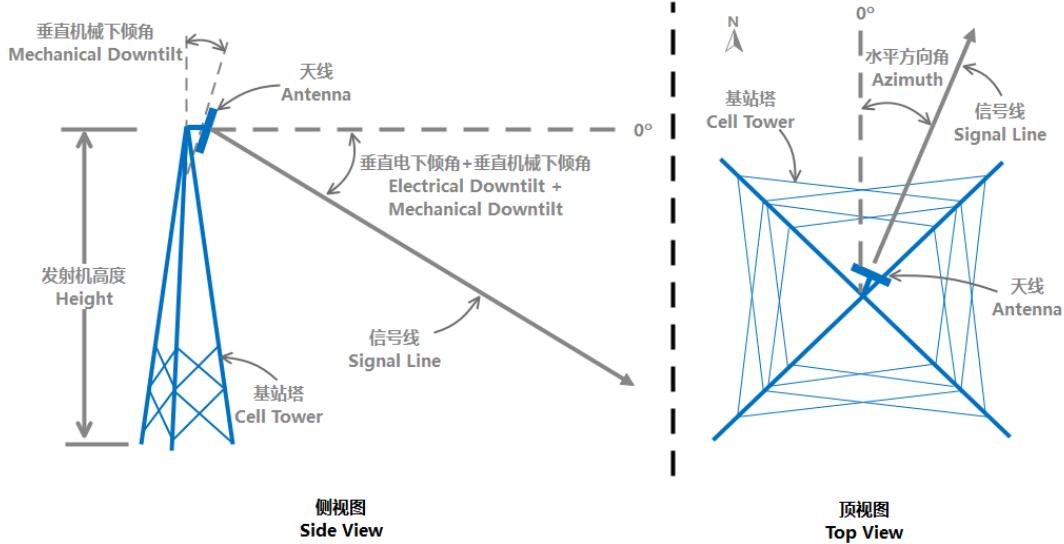


图 4.15 天线角度参数示意图

小区发射机垂直电下倾角 θ_{ED} 、垂直机械下倾角 θ_{MD} 直接影响了信号线相对栅格高度。我们已知链路距离对信号损耗会造成影响，这是在水平距离上产生的影响，然而发射天线有效高度 h_b 、信号线相对栅格高度 Δh_v 以及用户天线高度 h_{ue} 将在垂直方向上影响信号的传播距离。其中：

$$h_b = \text{Height} + \text{Cell Altitude} \quad (4.12)$$

$$\Delta h_v = h_b - d * \tan(\theta_{MD} + \theta_{ED}) \quad (4.13)$$

$$h_{ue} = \text{Building Height} + \text{Altitude} \quad (4.14)$$

此外，垂直方向上若小区站点所在栅格的建筑物高度高于该小区发射机相对地面的高度，电磁波将无法穿透该建筑，也会对链路损耗产生影响。

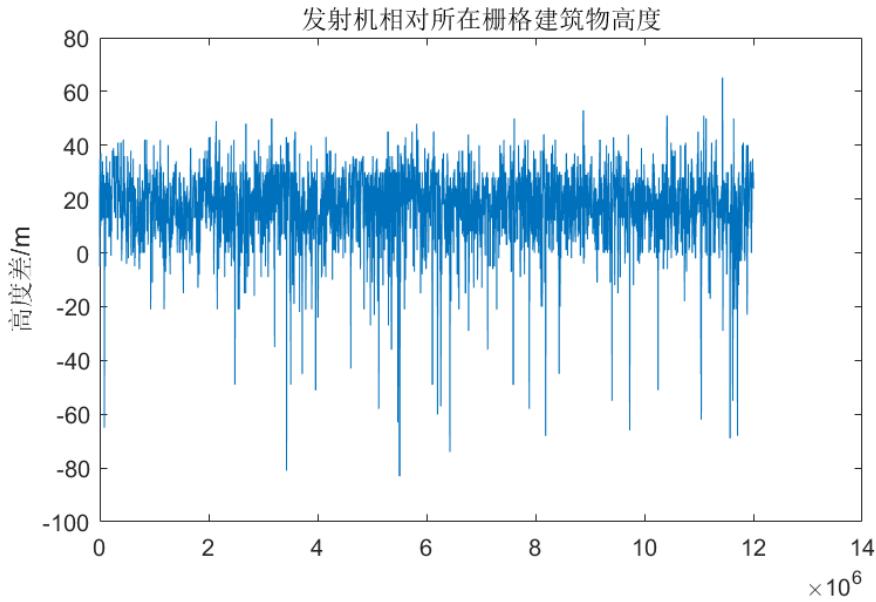


图 4.16 发射机相对所在栅格建筑物高度示意图

图 4.16 所示发射机与其所在栅格建筑物相对高度，由于发射机高度低于建筑物高度的样本量较少，我们可以忽略其影响；且每台所覆盖的区域均集中在其所在栅格附近，我们可将其影响在地物类型中作分析。

训练数据集提供了小区站点所在栅格的三维坐标(Cell X, Cell Y, CellAltitude)与栅格位置的三维坐标(X, Y, Altitude)，因此我们还可以设计链路三维距离 d_3 作为该模型的特征。

$$d_3 = \sqrt{(CellX - X)^2 + (CellY - Y)^2 + (CellAltitude - Altitude)^2} \quad (4.15)$$

由于题目并未提供发射天线的方向性图，因此我们应通过训练数据集提供的信息判断该天线为全向天线或定向天线，以确定 θ_{AZ} 是否可作为模型特征。

综上所述，我们初步设计了以下特征：

表 4.8 特征的初步设计

| 特征名称 | 特征符号 | 与特征有关的字段名称 | 单位 |
|------------|---------------|---|------------|
| 链路距离 | d | <i>Cell X, CellY, X, Y</i> | <i>m</i> |
| 链路三维距离 | d_3 | <i>Cell X, Cell Y, CellAltitude, X, Y, Altitude</i> | <i>m</i> |
| 载波频率 | f | <i>Frequency Band</i> | <i>MHz</i> |
| 发射天线有效高度 | h_b | <i>Height, Cell Altitude</i> | <i>m</i> |
| 信号线相对栅格高度 | Δh_v | <i>Electrical Downtilt, Mechanical Downtilt, Height</i> | <i>m</i> |
| 用户天线高度 | h_{ue} | <i>Building Height, Altitude</i> | <i>m</i> |
| 地物类型索引 | Index | <i>Clutter Index</i> | - |
| 小区发射机水平方向角 | θ_{AZ} | <i>Azimuth</i> | <i>Deg</i> |

我们对上述特征进行灰色相关分析，关联度如下表所示：

表 4.9 基于经验传播模型提取特征数据灰色关联分析结果

| 特征名称 | 关联度 |
|------------|----------|
| 链路距离 | 0.826418 |
| 链路三维距离 | 0.826422 |
| 载波频率 | 0.979514 |
| 发射天线有效高度 | 0.978381 |
| 信号线相对栅格高度 | 0.943946 |
| 用户天线高度 | 0.978841 |
| 地物类型索引 | 0.924284 |
| 小区发射机水平方向角 | 0.889602 |

根据经验模型，我们利用原始数据中各字段进行特征的进一步提取，例如：根据地理位置 x, y 坐标进行二维距离设计，在此基础上进一步引入海拔参数，给出了三维距离的统计特性，总结出定向天线信号强度与接收端相对高度特征等。通过关联度分析出，这些量化指标均具有相对于 RSRP 的强相关性。第二问会具体将数据根据地物类型索引以及发射机中心频率带宽等参数进行数据分割与细化分析。

五、问题二的模型建立与求解

问题二要求基于提供的各小区数据集，设计多个合适的无线传播模型特征，并计算这些特征与目标的相关性，将结果量化、排序，形成表格。还要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。

5.1 解题思路概述

首先，通过对数据做回归分析、假设检验等初步筛选特征。

接着，结合实测数据分析前面提出的特征对结果是否有明显影响。其原则是尽量不错过一个可能有用的特征，但是也不滥用太多的特征。可以选择过滤法、基于 BP 神经网络的权值分析法分析自变量与目标变量之间的关联选择特征：

a): 方差筛选法：方差衡量的是一个随机变量取值的分散程度。如果一个随机变量的方差非常小，那这个变量作为输入，是很难对输出有什么影响的。在进行特征选择时，可以丢弃那些方差特别小的特征。

b): 相关系数法：相关系数表征的是两个随机变量之间的线性相关关系。特征与输出的相关系数的绝对值越大，说明对输出的影响越大，应该优先选择。

c): 基于 BP 神经网络的权值分析法：BP 神经网络是一种具有三层或者三层以上的多层神经网络，不断修正误差逆传播训练，得到不同神经元的权值，权值越大，对输出的影响越大。

最后，通过对所有检验结果检验统计量对相关关系进行量化、排序，并结合理论与实际，对量化结果进行评价。

5.2 特征关系的初步判断

特征关系的初步判断中，我们应选取适当方法推测目标与一个或者几个特征之间的相关关系以验证特征选取的合理性，本文采用以下两种方法。

a): 回归分析法：根据已知的若干有关变量的数据通过回归分析进行曲线拟合，分别观察特征与目标的拟合程度，可以推测目标与一个或者几个特征之间的相关关系。

b): 假设检验法：假设检验是一种统计推断方法，假设某个特征和输出是有显著相关性的，如果假设成立，即选择该特征；反之，丢弃该特征。

(a) 回归分析法

通过经验传播公式的分析我们可得知：链路距离 d 、栅格与信号线相对高度 Δh_v 与路径损耗存在明显的对应关系。因此，在分析链路距离 d 与链路损耗之间的关系及栅格与信号线相对高度与链路损耗之间的关系时，我们采用一元线性回归模型。

一元线性回归的模型为：

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (5.1)$$

其中， β_0 , β_1 为回归系数， ε 是随机误差项，我们总是假设 $\varepsilon \sim N(0, \sigma^2)$ 。若对 y 和 x 分别进行 n 次独立观测，就可以得到 n 对观测值，这 n 对观测值之间的关系符合模型：

$$y_i = \beta_0 + \beta_1 x + \varepsilon_i, i = 1, 2, \dots, n \quad (5.2)$$

x_i 是自变量在第 i 次观测时的取值，它是一个非随机变量，并且没有测量误差。

对于 x_i 来说， y_i 是一个随机变量，它的随机性是由 ε_i 造成的， $\varepsilon \sim N(0, \sigma^2)$ ，对于不同的观测样本，当 $i \neq j$ 时， ε_i 与 ε_j 是相互独立的。在本题中我们采用一个文件的所有样本进行测试。分别将链路距离 d 和栅格与信号线相对高度作为自变量，将链路损耗作为因变量进行一元线性回归建模^[3]。

其次我们先采用用最小二乘法求线性回归系数^[4], 用最小二乘法估计 β_0 , β_1 的值, 使 y_i 与 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x$ 的误差平方和达到最小, 若记

$$Q(\hat{\beta}_0, \hat{\beta}_1) = \min_{\beta_0, \beta_1} Q(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \quad (5.3)$$

则显然 $Q(\beta_0, \beta_1) \geq 0$, 且关于 β_0 , β_1 可微, 则由多元函数存在极值的必要条件得:

$$\begin{aligned} \frac{\partial Q}{\partial \beta_0} &= -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 * x_i) = 0 \\ \frac{\partial Q}{\partial \beta_1} &= -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 * x_i) = 0 \end{aligned} \quad (5.4)$$

整理后, 得到下面的正规方程组,

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \\ \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \end{cases} \quad (5.5)$$

求解可以得到:

$$\begin{cases} \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases} \quad (5.6)$$

其中, $\hat{\beta}_0$, $\hat{\beta}_1$ 是 β_0 , β_1 的最小二乘估计, \bar{x} , \bar{y} 分别是 x_i , y_i 的样本均值。

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i \\ \bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \end{aligned} \quad (5.7)$$

得到残差:

$$e_i = y_i - \hat{y}_i, i = 1, 2, \dots, n \quad (5.8)$$

残差的样本均值为:

$$\bar{e} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i) = 0 \quad (5.9)$$

残差的样本方差为:

$$MSE = \frac{1}{n-2} \sum_{i=1}^n (\bar{e}_i - \bar{e})^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (5.10)$$

记:

$$S_e = \sqrt{MSE} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n e_i^2} \quad (5.11)$$

经过理论分析我们可以知道如果自变量与目标拟合较好，则其残差总和应越小越好。残差越小，拟合值与观测值越接近，各观测点在拟合直线周围聚集的紧密程度越高。当 S_e 越小时，还说明残差值的变异程度越小。由于残差的样本均值为零，所以，其离散范围越小，拟合的模型就越为精确。

记:

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (5.12)$$

这是原始数据 y_i 的总变异平方和， $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ 这是用拟合直线 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 可解

释的变异平方和。由于 $SST = SSR + SSE$ ，所以，可解释变异 SSR 越大，则必然有残差 SSE 越小。这个分解式可同时从两个方面说明拟合方程的优良程度:

(1) SSR 越大，用回归方程来解释 y_i 变异的部分越大，回归方程对原数据解释得越好；

(2) SSE 越小，观测值 y_i 绕回归直线越紧密，回归方程对原数据的拟合效果越好。

定义一个测量标准来说明回归方程对原始数据的拟合程度，这就是所谓的判定系数。判定系数是指可解释的变异占总变异的百分比，用 R^2 表示:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (5.13)$$

从定义可以看出：当 $R^2=1$ 时， $SR=SST$ ，此时原数据的总变异完全可以由拟合值的变异来解释，并且残差为零。当 $R^2=0$ 时，回归方程完全不能解释原数据的总变异， y 的变异完全由与 x 无关的因素引起。一方面它可以从数据变异的角度指出可解释的变异占总变异的百分比，从而说明回归直线拟合的优良程度；另一方面，它还可以从相关性的角度，说明原因变量与拟合变量的相关程度。

链路距离 d 与路径损耗的一元线性回归仿真结果为下图:

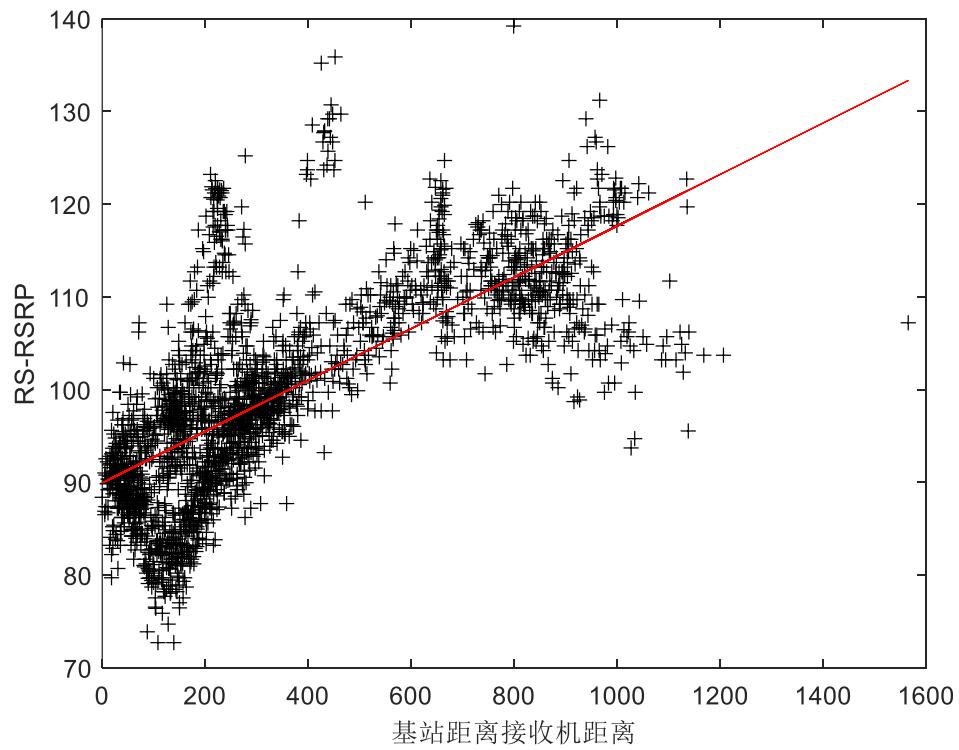


图 5.1 链路距离 d 与路径损耗一元线性回归示意图

从图中我们可以看到，随着基站距离与接收机距离的增大，链路的损耗是呈增加趋势的。所以二者大体呈线性相关关系，仿真后得到 $R^2=0.5221$ 。

栅格与信号线相对高度与链路损耗的一元线性回归仿真结果为下图：

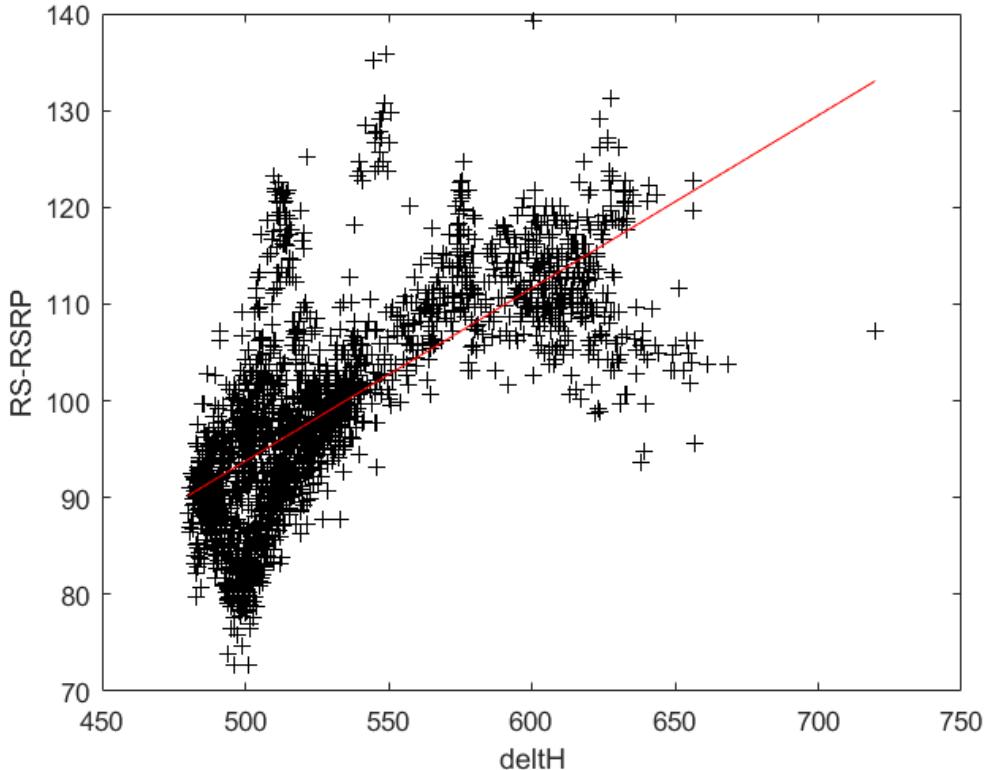


图 5.2 栅格与信号线相对高度 Δh 与链路损耗的一元线性回归示意图

从图中我们可以看到，随着栅格与信号线相对高度的增大，链路的损耗是呈增加趋势的。所以二者大体呈线性相关关系，仿真后得到 $R^2=0.5122$ 。

其余特征我们采用二元线性回归分析。二元线性回归分析的模型为：

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \\ \varepsilon \sim N(0, \sigma^2) \end{cases} \quad (5.14)$$

其中 $\beta_0, \beta_1, \beta_2$ 为回归系数，仍用最小二乘法估计，即应选取估计值 $\hat{\beta}_j$ 使当 $\hat{\beta}_j = \hat{\beta}_j, j=1,2$ 时，误差平方和为：

$$Q = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2 \quad (5.15)$$

达到最小，则令：

$$\frac{\partial Q}{\partial \beta_j} = 0, j = 0, 1, 2, \dots, n \quad (5.16)$$

得：

$$\begin{cases} \frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}) = 0 \\ \frac{\partial Q}{\partial \beta_j} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}) x_{ij} = 0, \quad j = 1, 2 \end{cases} \quad (5.17)$$

正规方程组的矩阵形式为：

$$X^T X \beta = X^T Y \quad (5.18)$$

当矩阵 X 列满秩时，解为：

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (5.19)$$

代回原模型得到 y 的估计值。而这组数据的拟合值为：

$$\hat{Y} = X \hat{\beta} \quad (5.20)$$

拟合误差为 $e = Y - \hat{Y}$ 称为残差，可作为随机误差 ε 的估计，而：

$$Q = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (5.21)$$

为残差平方和（或剩余平方和）。

$$EQ = (n-3)\sigma^2, \quad \frac{Q}{\sigma^2} \sim \chi^2(n-3) \quad (5.22)$$

对总平方和 $SST = \sum_{i=1}^n (y_i - \bar{y})^2$ 进行分解，有：

$$SST = Q + U, U = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (5.23)$$

其中 Q 是残差平方和, 反映随机误差对 y 的影响, U 称为回归平方和, 反映自变量对 y 的影响。

将链路距离 d 及栅格与信号线相对高度同时作为特征, 链路损耗作为目标进行二元线性回归建模, 仿真如下:

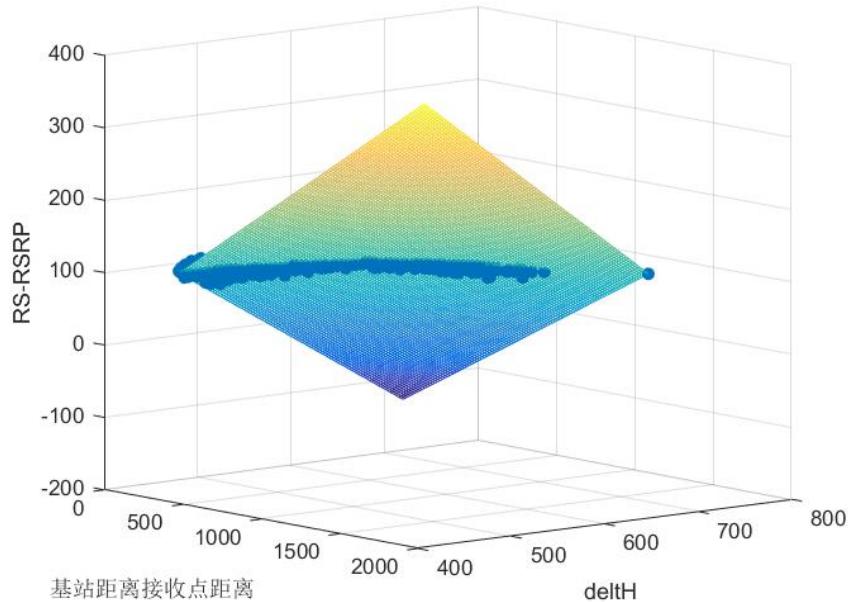
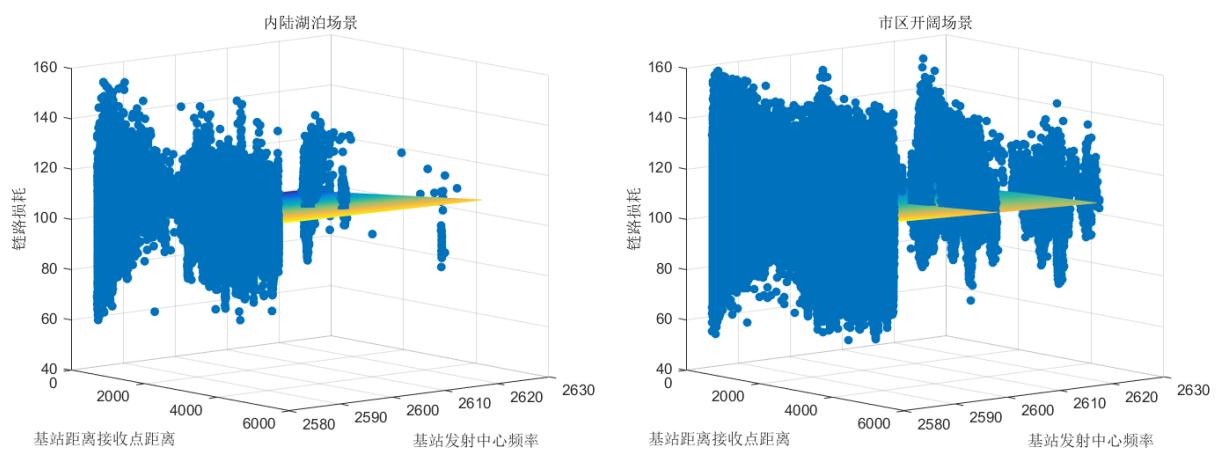


图 5.3 链路距离 d 及 Δh_v 与链路损耗的二元线性回归示意图

从图中拟合曲线我们可以看到, 所以三者呈线性相关关系。仿真后得到 $R^2=0.8242$, $S^2=98.4521$ 。

由于我们的数据样本点只包括三种频率值, 所以频率与链路损耗的关系不进行单独分析, 考虑在不同的地形环境下, 将基站频率与链路距离及链路损耗进行二元线性回归建模, 仿真得到:



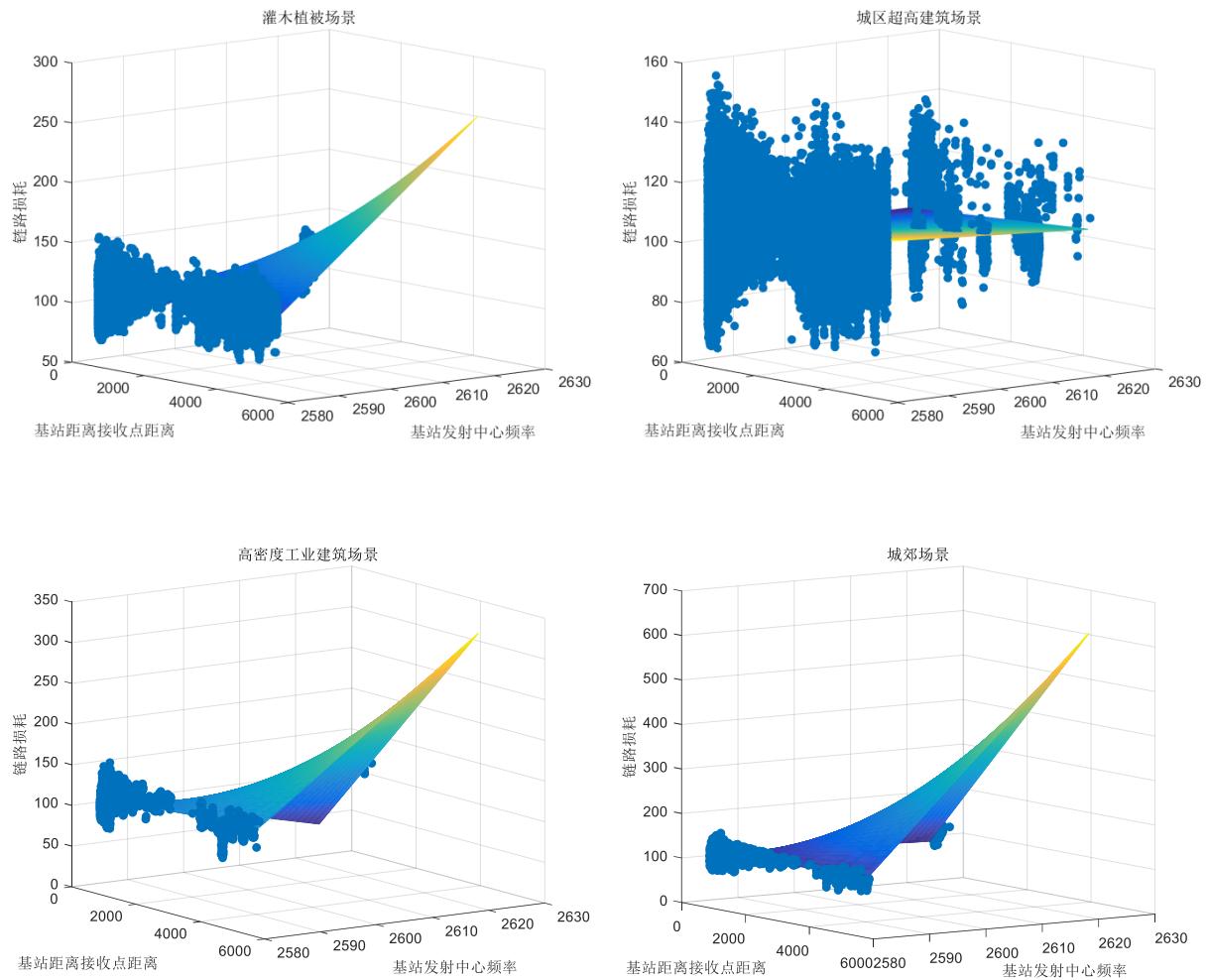


图 5.4 不同环境基站频率、链路距离及链路损耗二元线性回归结果

(b) 假设检验法

在拟合回归方程之前，我们曾假设数据总体是符合线性正态误差模型的，也就是说， y 与 x 之间的关系是线性关系，这种假设是否真实，还需进行检验。现给出假设：

$$H_0: \beta_1 = 0 \quad (5.24)$$

假设检验使用的统计量为：

$$\begin{aligned} F &= \frac{SSR / 1}{SSE(N-2)} = \frac{MSR}{MSE} \\ F &= \frac{MSR}{MSE} \sim F(1, n-2) \end{aligned} \quad (5.25)$$

若 $F \leq F_\alpha(1, n-2)$ 就接收假设，反之则拒绝。

链路距离 d 与路径损耗做 F 检验得： $F=2.4093$, $P=0$, 满足 P 小于 0.05, 证明链路距离 d 与链路损耗是线性关系。

栅格与信号线相对高度 Δh_v 与路径损耗做 F 检验得： $F=2.3151$, $P=0$, 满足 P 小于 0.05, 证明链路距离 Δh_v 与链路损耗是线性关系。

5.3 特征选择

完成特征设计后，需要选择有意义的特征输入机器学习模型进行训练。我们将通过方差筛选法、相关系数法、假设检验法对特征进行量化、排序。

5.3.1 方差筛选法

方差可以衡量一个随机变量取值的分散程度。若特征的方差非常小，那么它作为输入时，对输出的影响极小，可以丢弃。我们对 5.2 节中选出的 7 个特征分别求取了方差，并设置阈值为 1：

表 5.1 特征的方差

| 排序 | 特征名称 | 特征符号 | 方差（绝对值） |
|----|------------|---------------|-------------|
| 1 | 链路距离 | d | 1188795.482 |
| 1 | 链路三维距离 | d_3 | 1188795.482 |
| 3 | 信号线相对栅格高度 | Δh_v | 35173 |
| 4 | 用户天线高度 | h_{ue} | 333.4989 |
| 5 | 发射天线有效高度 | h_b | 216.6888 |
| 6 | 载波频率 | f | 29.5816519 |
| 7 | 地物类型索引 | Index | 9.2546 |
| 8 | 小区发射机水平方向角 | θ_{AZ} | 3.2497 |

由表 5.1 我们得知，这七个特征的方差均大于阈值 1，因此方差筛选中，七个特征值均不能被舍弃。

5.3.2 相关系数法

相关系数可以表征两个随机变量之间的线性相关关系，特征与输出的相关系数的绝对值越大，对输出的影响越大，应该优先选择。我们将 5.1 中选取的特征值分别与链路损耗求取互相关系数，如表 5.2 所示。

表 5.2 特征与链路损耗的互相关系数

| 排序 | 特征名称 | 特征符号 | 互相关系数（绝对值） |
|----|------------|---------------|------------|
| 1 | 链路距离 | d | 0.1775 |
| 1 | 链路三维距离 | d_3 | 0.1775 |
| 3 | 信号线相对栅格高度 | Δh_v | 0.1499 |
| 4 | 用户天线高度 | h_{ue} | 0.0493 |
| 5 | 地物类型索引 | Index | 0.0257 |
| 6 | 载波频率 | f | 0.0104 |
| 7 | 小区发射机水平方向角 | θ_{AZ} | 0.0066 |
| 8 | 发射天线有效高度 | h_b | 0.002 |

由表 5.2 我们得知，小区发射机水平方向角与链路损耗基本不相关，可初步认为小区发射机所用天线为全向天线。

表 5.1 与 5.2 数据中，关于链路距离 d 和链路三维距离 d_3 的方差及其余链路损耗互相关系数相同，且由(4.7)、(4.11)我们可以发现， d_3 的获取是建立在 d 基础之上的，认为二者可能存在较强相关性；而信号线相对栅格高度 Δh_v 表达式(4.9)也与 d 有关。因此对部分特征求取互相关系数用于判断其关系：

表 5.3 部分特征间的互相关系数

| 特征名称 1 | 特征名称 2 | 互相关系数（绝对值） |
|--------|-----------|------------|
| 链路距离 | 链路三维距离 | 1 |
| 链路距离 | 信号线相对栅格高度 | 0.8782 |
| 用户天线高度 | 发射天线有效高度 | 0.0603 |

由表 5.3 可知，链路距离 d 和链路三维距离 d_3 的互相关系数为 1，因此二者选其一做特征即可。

5.3.3 基于 BP 神经网络的权值分析法

我们还采用 BP 神经网络在 Matlab 中运行来对所选特征进行数据分析，首先需要对数据进行归一化处理，然后建立神经网络，接着进行网络训练，并根据训练结果，输入测试数据进行仿真，最后将数据进行反归一化处理。这里我们将 5.3.2 挑选的 7 个特征设置为 7 个神经元，经过 BP 网络训练，得到 7 个神经元的权值。

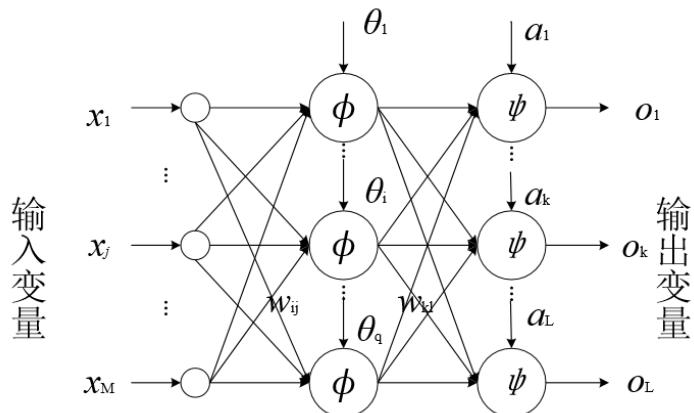


图 5.5 BP 神经网络示意图

BP(Back Propagation)神经网络是一种具有三层或者三层以上的多层神经网络，每一层都由若干个神经元组成，它的左、右各层之间各个神经元实现全连接，即左层的每一个神经元与右层的每个神经元都由连接，而上下各神经元之间无连接。BP 神经网络按有导师学习方式进行训练，当一对学习模式提供给神经网络后，其神经元的激活值将从输入层经各隐含层向输出层传播，在输出层的各神经元输出对应于输入模式的网络响应。然后，按减少希望输出与实际输出误差的原则，从输出层经各隐含层，最后回到输入层（从右到左）逐层修正各连接权。由于这种修正过程是从输出到输入逐层进行的，所以称它为“误差逆传播算法”。随着这种误差逆传播训练的不断修正，网络对输入模式响应的正确率也将不断提高^[5]。

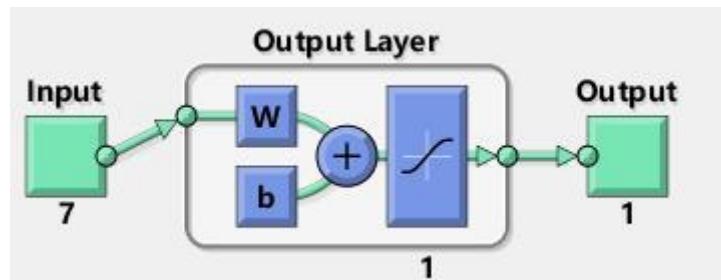


图 5.6 特征神经网络训练示意图

基本 BP 算法包括两个方面：信号的前向传播和误差的反向传播。即计算实际输出时按从输入到输出的方向进行，而权值和阈值的修正从输出到输入的方向进行。

经过仿真分析，我们得到了 7 个神经元的权值信息，如表 5.5：

表 5.4 特征的 BP 神经网络权重

| 排序 | 特征名称 | 特征符号 | BP 网络权重系数（绝对值） |
|----|------------|---------------|----------------|
| 1 | 链路距离 | d | 0.34875 |
| 2 | 用户天线高度 | h_{ue} | 0.093839 |
| 3 | 信号线相对栅格高度 | Δh_v | 0.035021 |
| 4 | 地物类型索引 | $index$ | 0.013501 |
| 5 | 发射天线有效高度 | h_b | 0.011241 |
| 6 | 载波频率 | f | 0.0042783 |
| 7 | 小区发射机水平方向角 | θ_{AZ} | 0.0023087 |

综合上述量化分析表数据以及经验模型理论，我们给出如下相关性对比表。

表 5.5 特征与目标的相关性排序表

| 排序 | 特征名称 | 该特征与目标的相关性 |
|----|------------|------------|
| 1 | 链路距离 | 强相关 |
| 2 | 用户天线高度 | 强相关 |
| 3 | 信号线相对栅格高度 | 较强相关 |
| 4 | 地物类型索引 | 较强相关 |
| 5 | 载波频率 | 弱相关 |
| 6 | 发射天线有效高度 | 弱相关 |
| 7 | 小区发射机水平方向角 | 弱相关 |

六、问题三的模型建立与求解

问题三要求我们在设计和选择有效的特征之后，根据已知的小区数据以及地形数据建立基于 AI 的无线传播模型来对不同地理位置的 RSRP 进行预测。该部分代码见论文最后附件部分。

6.1 解题思路概述

将 TensorFlow 框架与深度神经网络(Deep Neural Networks，简称 DNN)模型相结合来预测目标数值，评估通过模型所得到的预测值与真实值是否一样。首先将已有的数据根据前述问题的特征选取进行数据预处理并送入 DNN 模型训练，通过改变模型参数，例如：网络层数、网络每层单元数以及权重、激活函数等参数进行训练。经过训练后进行预测与实际 RSRP 的对比，从而分析数据得到结论。

6.2 TensorFlow 框架概述

TensorFlow 的分层系统架构为：从底层依次为设备管理层和网络层、数据操作层、图形计算层、API 接口层和应用层。前端系统提供编程模型，构造计算图；后端系统提供运行环境，执行计算图。

- (1) 底层设备管理层主要用于实现 TensorFlow 设备的异构特性。
- (2) 第二层是 TensorFlow 的核(OpKernels)实现。Tensor 为处理对象，通过网络通信和设备内存分配，实现了各种 Tensor 操作或计算。
- (3) 第三层是图计算层(Graph 层)，用于实现本地计算流图和分布式计算流图。
- (4) 第四层是 API 接口层。TensorC API 是对 TensorFlow 功能模块的接口封装，便于其他语言调用。
- (5) 第五层是应用层。不同编程语言在应用层通过 API 接口层调用 TensorFlow 核心功能进行相关实验和应用^[6]。

TensorFlow 程序开发流程如图所示：

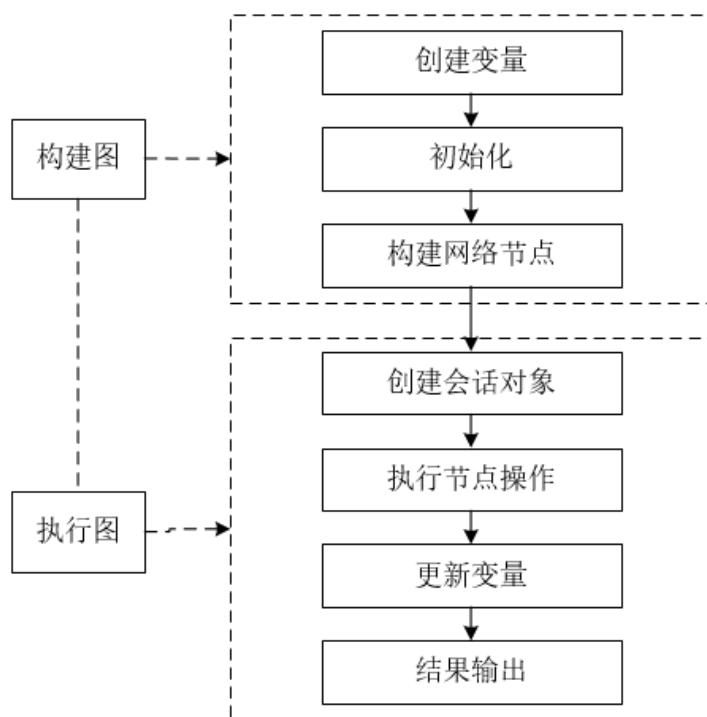


图 6.1 TensorFlow 程序开发过程

6.3 数据清洗与预处理

我们首先将全部 4000 个小区的数据合并，一共 12011833 条数据，通过第一二问分析，结合提取出的特征以及原始数据有用指标，总结出 13 条特征数据列，对模型进行评估训练，具体指标如下表所示。将数据进行切割，其中训练集：测试机为 4:1，其次在测试集中又以 1:1 比例切割为测试机和交叉验证集。具体如下图所示。

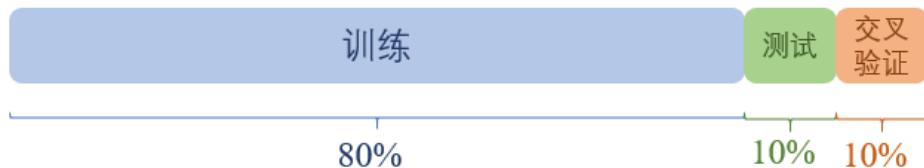


图 6.2 数据预处理过程示意图

训练过程中，我们使用第二问表 5.6 总结的特征，同时将原始参数中的部分变量辅助分析，训练使用的特征如图所示。

表 6.1 模型训练使用特征

| 特征名称 | 特征符号 |
|----------------|---------------|
| 链路距离 | d |
| 链路三维距离 | d_3 |
| 信号线相对栅格高度 | Δh_v |
| 用户天线高度 | h_{ue} |
| 地物类型索引 | $index$ |
| 载波频率 | f |
| 小区发射机水平方向角 | θ_{AZ} |
| 发射天线有效高度 | h_b |
| 小区发射机相对地面的高度 | h |
| 小区站点所在栅格建筑物高度 | h_{bc} |
| 小区站点所在栅格海拔高度 | h_{ac} |
| 小区站点所在栅格地物类型索引 | $CellIndex$ |
| 小区发射机发射功率 | P_{RS} |

6.4 深度神经网络(Deep Neural Networks)模型

6.4.1 DNN 模型概述

我们所用到的深度学习的实质，是通过构建具有很多隐层的机器学习模型和海量的训练数据，来学习更有用的特征，从而最终提升分类或预测的准确性。

本题目设计网络各层参数如下表结构所示：

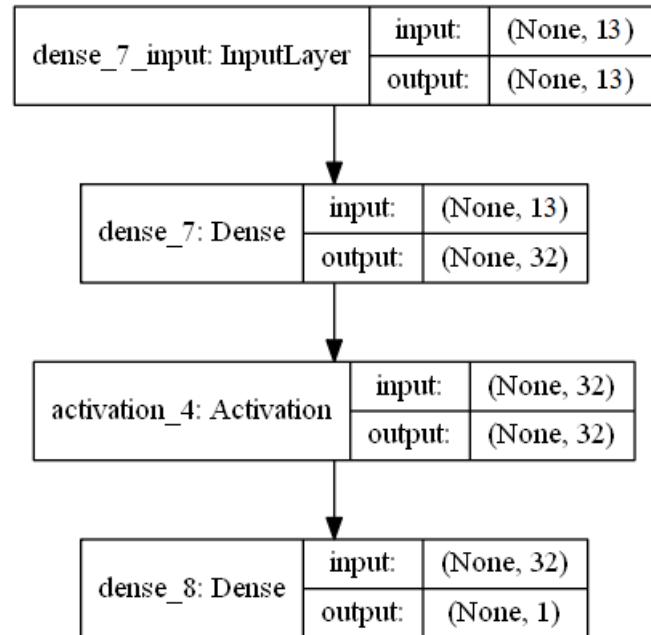


图 6.3 网络层参数设置

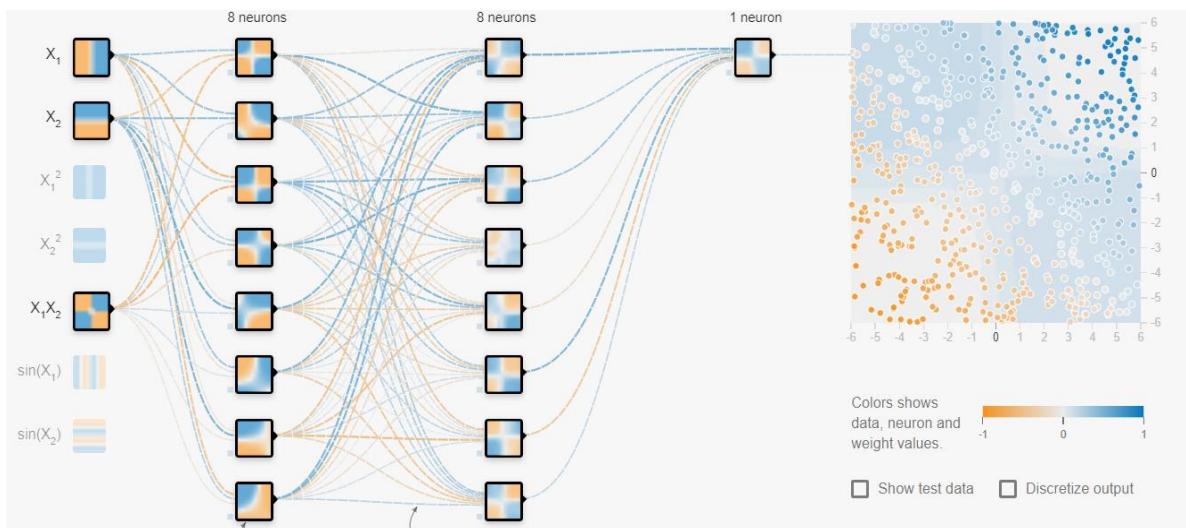


图 6.4 DNN 网络在回归预测中的数据流演示图例

(以三个训练特征，两个隐藏层各层八个单元，一个输出层为例，**不代表本题目真实网络结构**)

在 DNN 中，损失函数优化极值求解的过程最常见的一般是通过梯度下降法来一步步迭代完成的，本题目设置 loss 为实际与模型预测数据的均方差为指标，以梯度下降作为迭代优化器，沿着减小 loss 的方向迭代优化模型权重，偏置等模型参数。以下是模型训练结果：

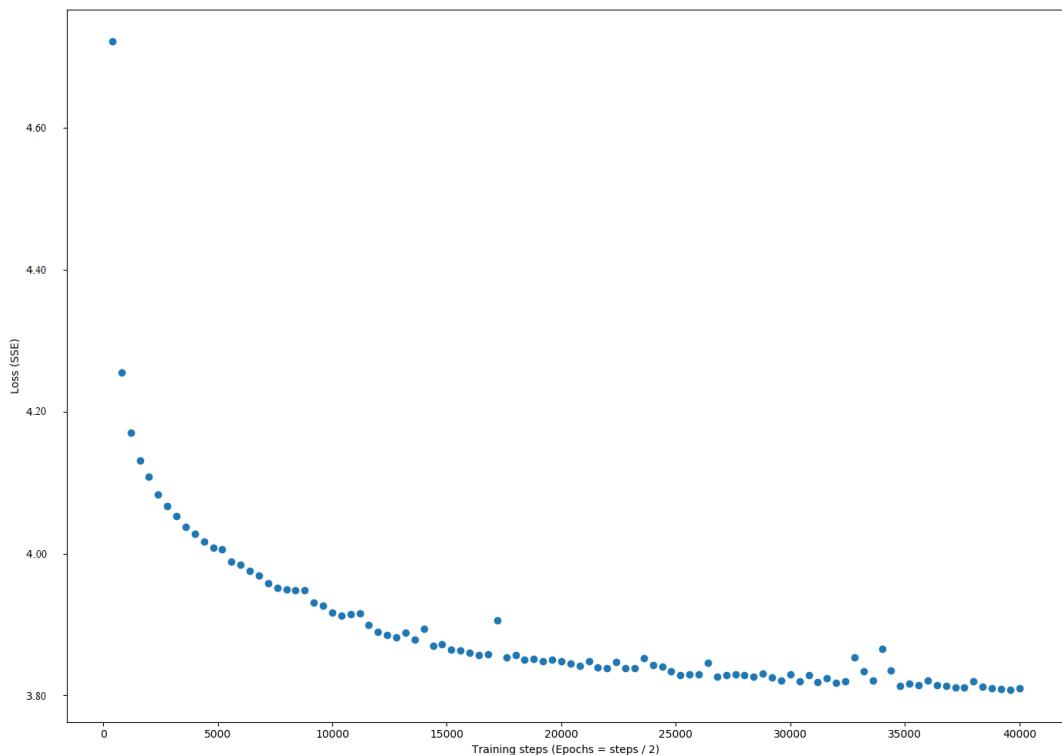


图 6.5 DNN 网络训练 loss 趋势图

表 6.2 部分预测 RSRP 数据与实际 RSRP 数据对比

| 预测 RSRP | 实际 RSRP |
|---------|---------|
| -71.83 | -97.74 |
| -94 | -91.10 |
| -97.75 | -88.25 |
| -105.45 | -96.93 |
| -94 | -93.69 |
| -76.89 | -89.86 |
| -91 | -90.10 |
| -92.04 | -82.961 |
| -106 | -85.29 |
| -91 | -87.74 |
| -100 | -96.07 |
| -71.95 | -88.83 |
| -93.33 | -84.03 |
| -82.5 | -86.39 |
| -82.5 | -98.31 |
| -88.33 | -83.41 |
| -100.1 | -88.89 |
| -95.5 | -86.88 |
| -82.11 | -95.24 |
| -103 | -88.14 |
| -103.33 | -83.14 |
| -97.67 | -89.58 |
| -79.5 | -92.53 |
| -90.23 | -90.95 |
| -77.78 | -94.68 |
| -87.33 | -93.42 |
| -98.5 | -93.76 |
| -83.33 | -85.14 |
| -79.62 | -88.51 |
| -74.8 | -92.64 |
| -95 | -97.08 |
| -99.33 | -86.98 |
| -102.43 | -87.65 |
| -90.5 | -95.85 |
| -98.67 | -98.46 |
| -65 | -88.81 |
| -107.8 | -95.98 |
| -101.88 | -89.09 |
| -83.1 | -98.13 |

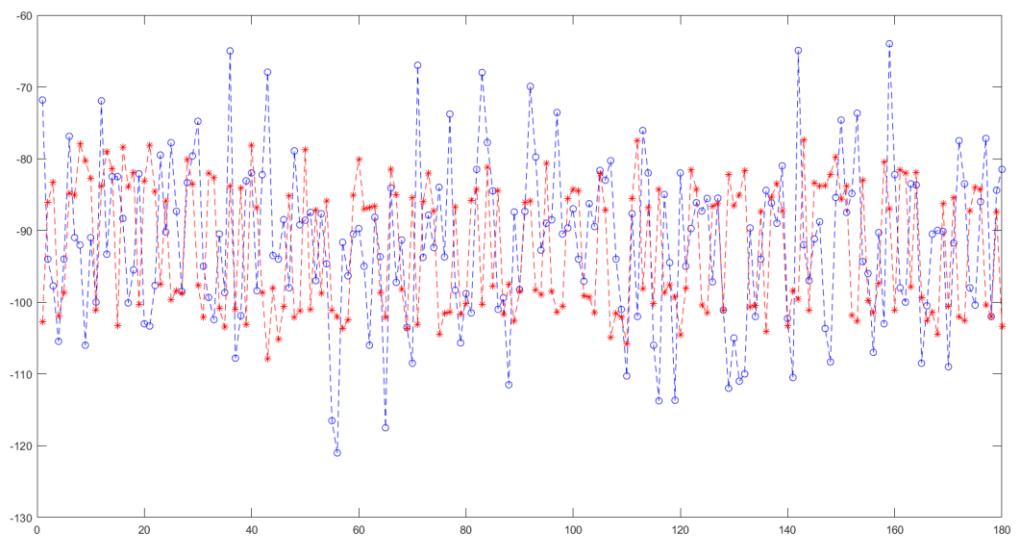


图 6.6 部分预测 RSRP 数据与实际 RSRP 数据对比示意图

```
D:\anconda\python.exe C:/Users/Young/PycharmProjects/Centrelines/loadData.py
The MSE: 135.59 degrees Celcius
The RMSE: 11.64 degrees Celcius
```

我们的模型均方误差 **RMSE** 为 **11.64**, 体现模型的有效性、可靠性。

七、总结与展望

7.1 模型评价

本文研究的是根据提供的训练数据集利用机器学习建立适当的无线智能传播模型，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，使得小区通信速率等指标的估算成为可能。为解决该问题，本文通过对经验传播模型与实测数据的分析，利用回归分析、假设检验等方法进行特征工程中的特征设计，再通过对所选特征与目标的拟合，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。根据所建立的无线传播模型特征集以及训练数据集，建立基于 TensorFlow 框架的深度神经网络模型来对不同地理位置的平均信号接收功率(RSRP)进行预测并与实际数据进行对比分析。

针对问题一：

1、总结已知数据集各参数的含义并对数据集中的中心频率、地物类型、发射机所在海拔高度等的数据进行分类统计。

2、对自由空间传播模型、经验传播模型（COST231-Hata 模型）进行理论分析，并利用 MATLAB 仿真出了模型参数与链路损耗相关的曲线图。并基于实测数据对传播模型进行了校正。提取出对模型有影响的 8 个特征，分别为：发射机中心频率、地物类型、链路距离、信号线相对栅格高度、用户天线高度、发射机水平方向角以及发射机有效天线高度。

针对问题二：

1、将问题一找到的特征量通过对数据回归分析、假设检验等初步筛选，通过一元线性相关模型与二元线性相关模型研究这些特征量与目标链路损耗具有的相关性的程度。结果表明上述特征均与目标有着线性相关的特性。

2、结合实测数据使用方差筛选法、相关系数法，以及基于 BP 神经网络的权值分析法分别对特征值与目标的相关性进行量化、排序，我们发现链路距离以及用户天线高度与目标链路损耗具有较强的相关性。

针对问题三：

1、对 TensorFlow 框架以及深层神经网络模型进行搭建，介绍了模型的结构内容与运算方法等。

2、通过将选择的有效特征利用 TensorFlow 框架搭建的深层神经网络 DNN 模型以及训练数据集进行无线传播模型的训练。

3、利用建立好的无线智能传播模型通过数据集中的测试集对不同地理位置的 RSRP 进行预测并与预测集给出对比结果。通过结果分析得到模型预测数据与实测数据基本一致，波动不大，得到数据的方差为 -0.17，平均绝对误差为 9.72Deg，绝对误差中位数为 7.85Deg，其中 RMSE 指标为 **11.64**，从而验证了该模型的有效性。

7.2 模型改进

在问题分析的过程中我们假设建模过程中的客观环境良好，假设条件较多，后期需要加强模型的适用性。今后的研究可以尽可能的对邻近多个基站的信号数据进行分析，考虑到发射机之间的相互干扰。我们找到的特征均与目标是线性相关关系，可在后期引入更加全面的影响因素对模型进行进一步完善。无线智能传播模型也并非一味地制定最高最优的性能标准，而是在不断提升用户感受的基础上，力争降低成本，提高经济效益与资源利用率。由于比赛时间和精力限制，我们在华为云 ModelArts 平台提交了基础分，论文中 **RMSE: 11.64** 为在本地训练测试结果，特此说明。

参考文献

- [1] 于璐国. 无线电监测网络规划中网格化覆盖的研究与实现[D]. 2016.
- [2] 王亚辉. eNB 蜂窝小区盲点分析及传播模型研究[D]. 2017.
- [3] 冯守平, 石泽, 邹瑾. 一元线性回归模型中参数估计的几种方法比较[J]. 统计与决策, 2008, 2008(24):152-153.
- [4] 刘严. 多元线性回归的数学模型[J]. 沈阳工程学院学报(自然科学版), 2005, 1(2):128-129.
- [5] 张景华, 张军. 基于 MATLAB 神经网络工具箱 BP 神经网络仿真[C]. 中国系统工程学会过程系统工程年会. 2001.
- [6] 李晶. 基于 TensorFlow 的交通标识智能识别系统设计[D]. 2018.

附录：模型训练代码（python—tensorflow 框架）

```
import pandas as pd
import numpy as np
import tensorflow as tf
from sklearn.metrics import explained_variance_score, \
    mean_absolute_error, \
    median_absolute_error
from sklearn.model_selection import train_test_split
from makeData import dataRead
import matplotlib.pyplot as plt
import pickle
X = []
y = []
df = pd.read_csv("./train.csv")
print(df.columns)
X = df[[col for col in df.columns if col != 'allRSRP']]
y = df['allRSRP']

X_train, X_tmp, y_train, y_tmp = train_test_split(X, y, test_size=0.2, random_state=23)
X_test, X_val, y_test, y_val = train_test_split(X_tmp, y_tmp, test_size=0.5, random_state=23)

# X_train.shape, X_test.shape, X_val.shape
print("Training instances {} , Training features {}".format(X_train.shape[0], X_train.shape[1]))
print("Validation instances {} , Validation features {}".format(X_val.shape[0], X_val.shape[1]))
print("Testing instances {} , Testing features {}".format(X_test.shape[0], X_test.shape[1]))

feature_cols = [tf.feature_column.numeric_column(col) for col in X.columns]
regressor = tf.estimator.DNNRegressor(feature_columns=feature_cols,
                                      hidden_units=[32,32],
                                      model_dir='tf_wx_model')

def wx_input_fn(X, y=None, num_epochs=None, shuffle=True, batch_size=400):
    return tf.estimator.inputs.pandas_input_fn(x=X,
                                                y=y,
                                                num_epochs=num_epochs,
                                                shuffle=shuffle,
                                                batch_size=batch_size)

evaluations = []
STEPS = 400
for i in range(100):
    regressor.train(input_fn=wx_input_fn(X_train, y=y_train), steps=STEPS)
    evaluation = regressor.evaluate(input_fn=wx_input_fn(X_val, y_val,
                                                         num_epochs=1,
                                                         shuffle=True),steps=1)
```

```

evaluations.append(regressor.evaluate(input_fn=WX_input_fn(X_val,
                                         y_val,
                                         num_epochs=1,
                                         shuffle=True)))

evaluations[0]
plt.rcParams['figure.figsize'] = [14, 10]
loss_values = [ev['loss'] for ev in evaluations]
training_steps = [ev['global_step'] for ev in evaluations]

plt.scatter(x=training_steps, y=loss_values)
plt.xlabel('Training steps (Epochs = steps / 2)')
plt.ylabel('Loss (SSE)')
plt.show()

pred = regressor.predict(input_fn=WX_input_fn(X_test,
                                              num_epochs=1,
                                              shuffle=True))

predictions = np.array([p['predictions'][0] for p in pred])

pickle.dump(y_test, open('./y_test.txt', 'wb'))
pickle.dump(predictions, open('./predictions.txt', 'wb'))
print(y_test)
print(predictions)
print("The Explained Variance: %.2f" % explained_variance_score(
    y_test, predictions))
print("The Mean Absolute Error: %.2f degrees Celcius" % mean_absolute_error(
    y_test, predictions))

print("The Median Absolute Error: %.2f degrees Celcius" % median_absolute_error(
    y_test, predictions))
MSE = np.sum(np.power(y_test - predictions, 2)) / len(y_test)
print("The MSE: %.2f degrees Celcius" % np.sum(np.power(y_test - predictions, 2)) / len(y_test))
print("The RMSE: %.2f degrees Celcius" % np.sqrt(MSE))

```