特征选择

问题一：请根据 Cost231-Hata 模型以及下述数据集信息设计合适的特征，并阐述原因。

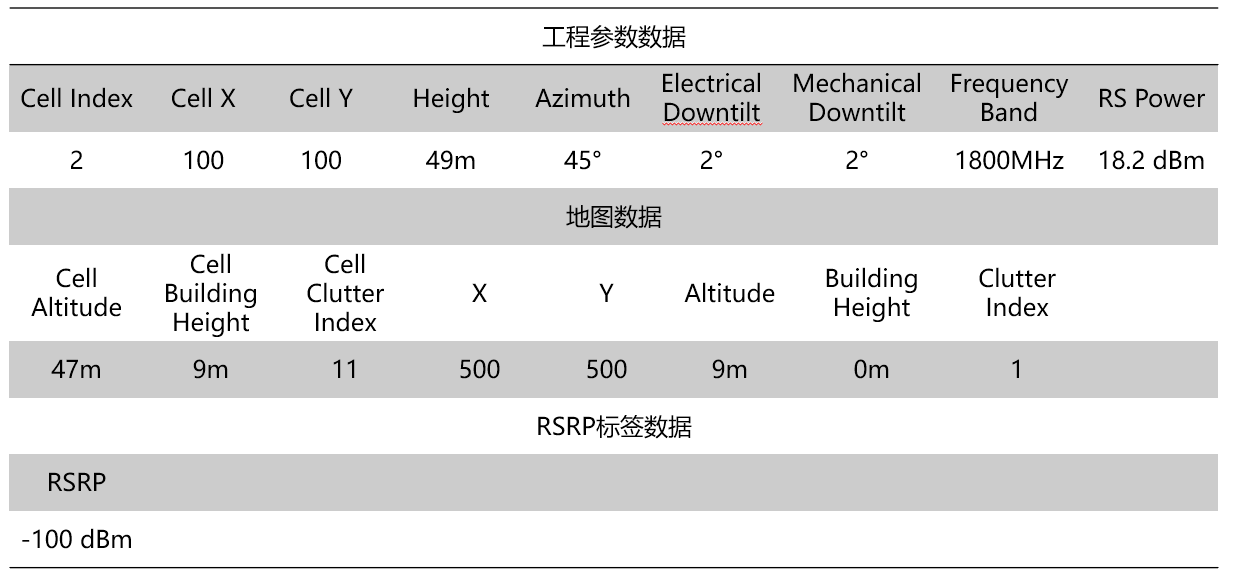


图 数据表信息

根据Cost231-Hata[]模型：

可知，影响接收点信号强度的关键因素为“发射信号频率”，“基站高度”，“接收点高度”，“接收点与基站的距离”以及“地物类型”。

所以对原始数据表做以下处理：

* 对原始数据表提供的发射机频率特征进行取对数处理；
* 构造基站与接受点之间的距离，如下图所示

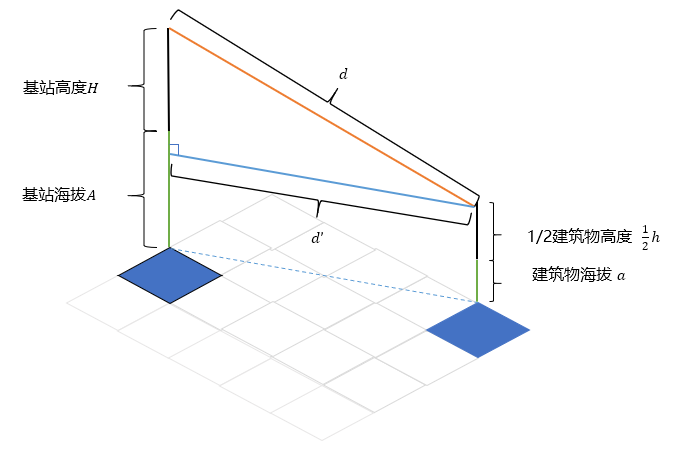


图 基站与接受点之间距离示意图

如上图所示，基于基站所处栅格以及接收点所处栅格的坐标，结合两处的海拔，以及基站和接收点建筑物高度,进行计算.因为最终的结果是平均信号强度，所以在这一步的时候，对于建筑高度进行了平均处理。计算过程如公式所示。

* 计算接收处栅格相对于基站发出信号线的相对高度。如下图所示

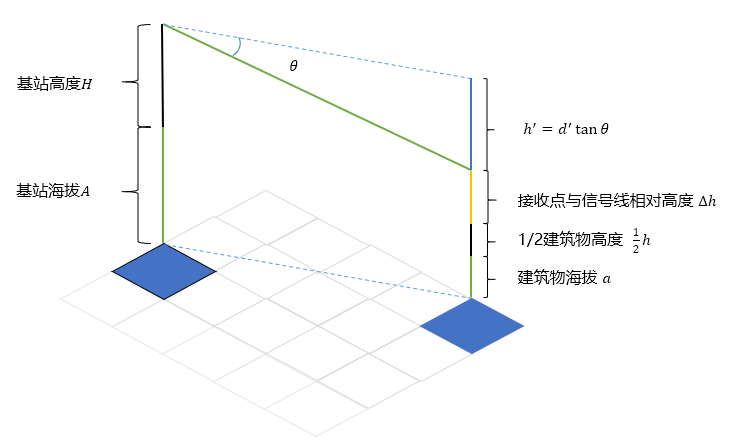


图 接收处栅格与信号线的相对高度

上图所示,根据上一步提取的距离特征，结合信号发射角度进一步计算接收处栅格与信号线的相对高度，计算过程如下式

其中，角度是原始数据表提供的垂直电下倾角与垂直机械下倾角之和。

* 划分地物类型,将连续数据离散化

原始数据表给定的地物类型是有固定含义的名称编号，这显然不适合直接作为特征进行相关计算,本文采用分类的方法，将原有的编号(1-20)离散为20类不同的类型,本栅格所属类型置为1,其余为零。例如，某一栅格的地物类型索引为5,表示是“市区开阔区域”，那么离散后的数据就将“市区开阔区域”对应的数据置为1，其余19类全部置为0。这样就将原有的地物类型索引映射到具体的离散数据上了。

但是考虑到基站到接收处沿途经过的所有栅格的地物类型都将会对最终信号强度产生影响，所以，对于沿途所有栅格的地物类型累计相加，这样，最终的得到的离散数据表示的是从基站到接收点沿途经过的所有栅格的地物类型的总和。例如，某基站到某一点沿途经过的栅格地貌类型有3种，其中，沿途栅格地物类型为“市区开阔区域”的有10个，其余两种都是1个，那么就将这一点对应的离散特征中对应“市区开阔区域”的值设置为10，另外两个置为1，其余17种置为0。