****  **** 

**中国研究生创新实践系列大赛**

**“华为杯”第二十一届中国研究生**

**数学建模竞赛**

|  |  |
| --- | --- |
| **学 校** |  |
| **参赛队号** |  |
| **队员姓名** | **1.** |
| **2.** |
| **3.** |

**中国研究生创新实践系列大赛**

**“华为杯”第二十一届中国研究生**

**数学建模竞赛**

题 目： ­­­­\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

摘 要：

对于问题一，首先进行数据清洗，剔除发送流量类型和标签集存在缺失值的部分。然后，对输入集进行Z-score标准化、0-1归一化，剔除异常数据。对清洗后的数据，采用xxxx分析影响重要性。将得到的3个重要变量，作为模型输入。按照同频AP个数分类建模，将训练集进行8：2随机划分，前者用于算法训练，后者用于验证算法准确率。全面评估了线性回归、XGBoost回归、随机森林回归、全连接神经网络，预测发送数据帧序列的总时长（seq\_time）的表现，并通过4 种评估标准自适应地选择了最优的 XGBoost模型，多次实验表明，收敛后的模型在随机划分的测试上的均方根误差低于 0.07，同时 R2 Score 超过了 0.72。

对于问题二，为了同时满足预测空间流数和调制编码方案的要求，本文提出了联合编码构建标签集方法。将训练集进行8：2随机划分，全面评估了线性回归、XGBoost回归、随机森林回归、全连接神经网络，分类预测空间流数和调制编码方案的表现，并通过4 种评估标准自适应地选择了最优的 XGBoost模型，多次实验表明，收敛后的模型在随机划分的测试上的均方根误差低于 0.07，同时 R2 Score 超过了 0.72。

对于问题三，为了预测系统吞吐量，在问题一、二的基础上，输入集添加实测中统计的数据帧真实空间流数和调制编码方案。按照同频AP个数分类建模，将训练集进行8：2随机划分，前者用于算法训练，后者用于验证算法准确率。全面评估了线性回归、XGBoost回归、随机森林回归、全连接神经网络，预测发送数据帧序列的总时长（seq\_time）的表现，并通过4 种评估标准自适应地选择了最优的 XGBoost模型，多次实验表明，收敛后的模型在随机划分的测试上的均方根误差低于 0.07，同时 R2 Score 超过了 0.72。

对于问题一，首先进行数据清洗，依次剔除了方差为0的变量，过滤了部分存在缺失值的变量。接着，对清洗后的数据与 pIC50 分别进行 Z-score 标准化和0-1 归一化。分别采用灰色关联分析、最大互信息系数、随机森林与平均精度损失计算分子描述符变量对ERa 活性影响重要度，并将结果相融合，以此提出了基于加权集成重要度分析的变量筛选法。 同时，考虑了变量间可能存在的线性相关和韭线性相关性，对高重要度变量进行独立性分析并利用自适应迭代变量选择法地进行筛选，从而获得最终的20个具有代表性、独立性的变量。

对于问题三，为了同时满足分子描述符变量对生物活性与 ADMET 具有高重要度的要求，因此重新构建了特征选择方法，构建了两阶段 XGBoost-MLP 特征选择与并行预测方造。在阶段一中借助 XGBoost 选择同时面向 ADMET 与生物活性的最优变量，在阶段二中，在此基础上构建了五个分别预测 ADMET 的基于 MLP模型。在随机划分的测试集上， 本文方法对五种特性的预测精度均超过了 80%，AUC 均超过了 0.9。

对于问题四，提出了基于 DE-MILP 的全局优化搜索算法。首先构建并训练了六个独立的 MLP，分别预测 ADMET 以及 ERa水平。在此基础上，将其作DE（差分进化）的目标函数，通过不断的交叉编译，搜索能够使 ADMET 总体较优，且 ERa 水平较好的化合物分子变量取值范围，药物研发提供了一定的帮助。

关键词：

# 问题重述

## 问题背景

无线局域网（Wireless Local Area Network，WLAN）是一种无线计算机网络，使用无线信道作为传输介质连接两个或多个设备。WLAN基于IEEE 802.11标准不断演进，从语音/邮件到网页和视频以及更加复杂的虚拟现实/增强现实，WLAN的部署场景也从家庭场景迅速覆盖到无线化办公、教育、医疗、工业制造、仓储等场景。下一代Wi-Fi 7标准支持峰值速率30Gbps，但在高密部署场景下，节点密集度增加，相邻小区覆盖范围重叠使得干扰、碰撞等问题突出，实际部署带宽和数据传输速率大幅下降，因此可支持系统吞吐量仍然有限，需要对WLAN系统进一步优化。

WLAN优化问题的基础问题是吞吐量预测，精准和快速的吞吐量预测能够为设计鲁棒性和高性能WLAN系统提供很大的技术提升空间。最近一些研究利用机器学习的方法，通过特征提取，如信道、带宽、流量、发送功率、信道接入机制等基本信息、各个节点之间的接收信号强度（Received Signal Strength Indication, RSSI）、信干噪比（Signal to Interference and Noise Ratio, SINR）、传输时间等架构信息、以及节点动态位置、动态干扰等临时信息，进行训练和建模，从而预测系统的吞吐量。这些研究大多通过与仿真结果对比，来验证模型的预测精度。然而实际部署中，WLAN系统的通信行为受到信道环境、干扰的快速变化，业务流量复杂多样的影响，仿真结果并不精确，因此依靠仿真所训练的模型无法真正商用。

综上，基于WLAN实测数据，分析WLAN网络拓扑、节点间RSSI、信道接入机制、干扰等因素对WLAN数据发送、速率的影响，并进一步对WLAN系统吞吐量进行精确预测和优化，具有十分重要的价值和意义。

## 问题描述

在本次题目的研究中，我们根据赛题所给出的三个子题目对WLAN的优化问题进行拆解，具体描述如下：

**问题一：基于测试基本信息的AP发送机会影响分析及预测**

针对WLAN网络实测数据集中的测试基本信息与AP发送机会的相关性进行分析，根据各个因素的影响性强弱进行排序。同时，根据不同的AP测试场景进行分类讨论，建立AP发送机会的预测模型，并讨论不同情况下建立通用预测模型的可行性。最终，对测试集中的模型输入信息进行预测。

**问题二：基于RSSI与门限信息的（MCS，NSS）参数预测**

根据问题一中AP发送机会分析的结果，筛选最多次数的（MCS，NSS）参数。基于实测数据中的基本测试信息，主要为RSSI与门限信息，建立（MCS，NSS）参数的预测模型，最终，对测试集中的模型输入信息进行预测。

**问题三：基于测试基本信息与（MCS，NSS）参数的网络吞吐量预测**

根据问题一与问题二的分析结果，讨论网络吞吐量受基本测试信息以及（MCS，NSS）参数的影响。基于以上的关键参数建立网络吞吐量预测模型，同时，考虑同频AP个数对模型的影响情况。最终，结合测试集中的模型输入与统计的数据帧真实（MCS，NSS）预测网络吞吐量。

# 基本假设与符号说明

## 基本假设

本次题目对以下关键词进行定义：

**WLAN吞吐量**：节点单位时间内成功发送的比特数。

## 符号说明

# 问题一的模型建立与求解

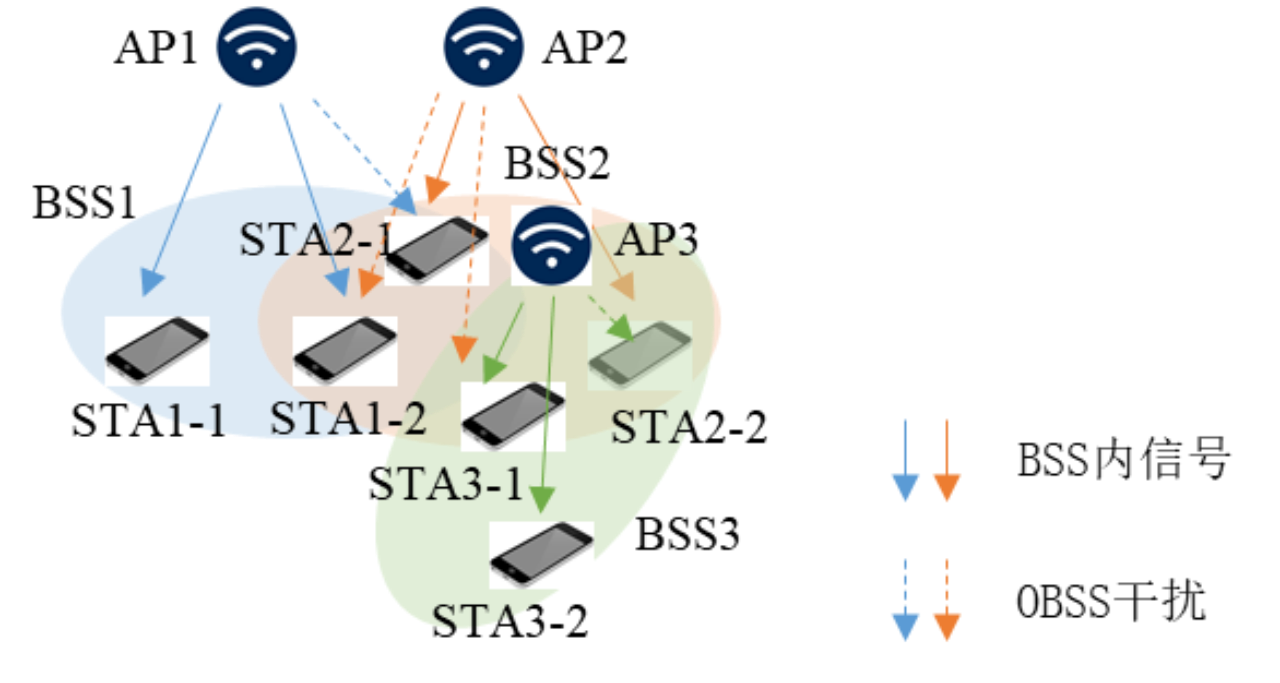
## 问题分析

本赛题提供的WLAN组网场景的拓扑图如下图X所示，同时，数据集采集的同频AP数量为2或3。根据题目描述，数据集字段的定义如下：

**测试基本信息（模型输入）**：网络拓扑、业务流量、门限、节点间RSSI

**数据帧统计信息（模型输出）**：各节点发送时长、（MCS，NSS）、PER、吞吐量

针对问题一，我们将其分为两个阶段进行处理：第一阶段，解决影响性强弱的排序问题，对实测数据集进行观察和分析，对数据集中参数的影响方式进行分类，并讨论其影响的模式，从而建立影响性分析模型，实现影响性强弱的排序；第二阶段，根据提取出的关键影响参数，针对2个AP的场景和3个AP的场景分别建立预测模型，实现AP发送机会的预测。



1. WLAN组网场景的测试拓扑

## 影响因子相关性分析

我们首先对实测数据集中的测试基本信息进行分析，包括网络拓扑信息，业务流量信息，门限信息以及节点间RSSI信息。通过观察数据的趋势和形态，我们初步得到了如表X的影响因子与AP发送机会的相关性分析结果，由于对原始数据集的观察较为丰富，详细的分类依据见附件1，后文仅对本文选取的关键影响因子进行详细讨论。

1. 影响因子与AP发送机会的相关性分析结果

|  |  |  |  |
| --- | --- | --- | --- |
| **影响因子**  （测试基本信息） | **相关性** | **名称** | **相关性分析** |
| eirp | 高 | 发送功率 | AP发送功率影响AP间发生干扰的阈值 |
| loc\_id | 高 | 部署位置 | 部署位置对AP发送机会影响较高 |
| nav | 高 | NAV门限 | NAV门限会影响AP间发生干扰的阈值 |
| protocol | 中 | 流量类型 | UDP或TCP的AP发送机会有明显差异 |
| RSSI\_ap\_from\_ap | 低 | AP间RSSI | AP互相监听的RSSI，AP位置相关 |
| RSSI\_sta\_to\_ap | 低 | AP监听STA | AP监听STA的RSSI，AP位置相关 |
| RSSI\_sta\_from\_ap | 低 | STA监听AP | STA监听AP的RSSI，AP位置相关 |
| RSSI\_sta\_from\_sta | 低 | STA监听STA | STA监听STA的RSSI，STA位置相关 |
| test\_id | 无 | 测试编号 | 用于标识测试 |
| bss\_id | 无 | BSS编号 | 用于标识BSS场景，与AP绑定 |
| ap\_name, ap\_id | 无 | AP标识 | 用于标识AP |
| sta\_mac, sta\_id | 无 | STA标识 | 用于标识STA，接收设备不影响发送 |
| test\_dur | 无 | 测试时间 | 固定常数 |
| pkt\_len | 无 | 数据长度 | 固定常数 |
| pd | 无 | 包门限 | 固定常数 |
| ed | 无 | 能量门限 | 固定常数 |

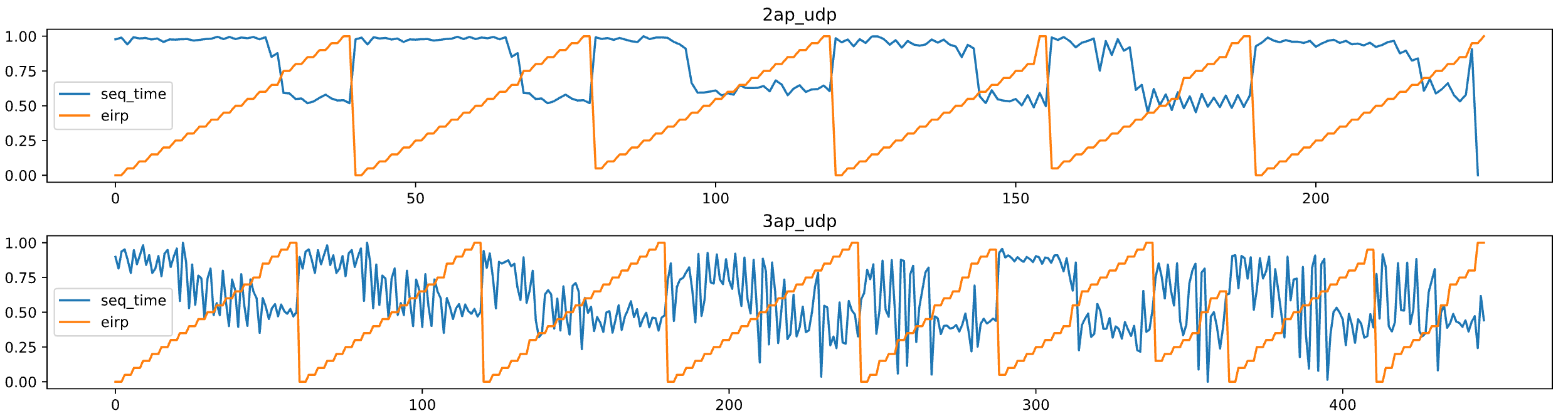
在上表的分析结果中，用于标识数据唯一性与固定常数的影响因子在实测数据集中无法提供足够的相关性信息，因此我们在影响性排序的过程中将会进行排除。其次，部署位置（loc\_id）、NAV门限、发送功率（eirp）、流量类型（protocol）与AP的发送机会具有较强的相关性，同时，经过我们的分析，RSSI可以作为AP与AP，AP与STA间信号强度或相对位置的判断依据。下面我们将会对这些影响因子进行详细的影响类型和程度的分析。

**发送功率（eirp）**：通过观察实测数据集，我们发现每个测试记录文件中发送功率随着实验次数递增，于是我们将实验按照测试顺序进行组合，同时，固定流量类型和两个AP与三个AP交互的场景，对发送功率与AP发送机会进行可视化分析，分析结果如下图X和图X所示。横坐标为随实验次数递增的序号，纵坐标分别表示AP发送机会与发送功率，这里我们将两个数值进行了归一化便于分析。结果表明：

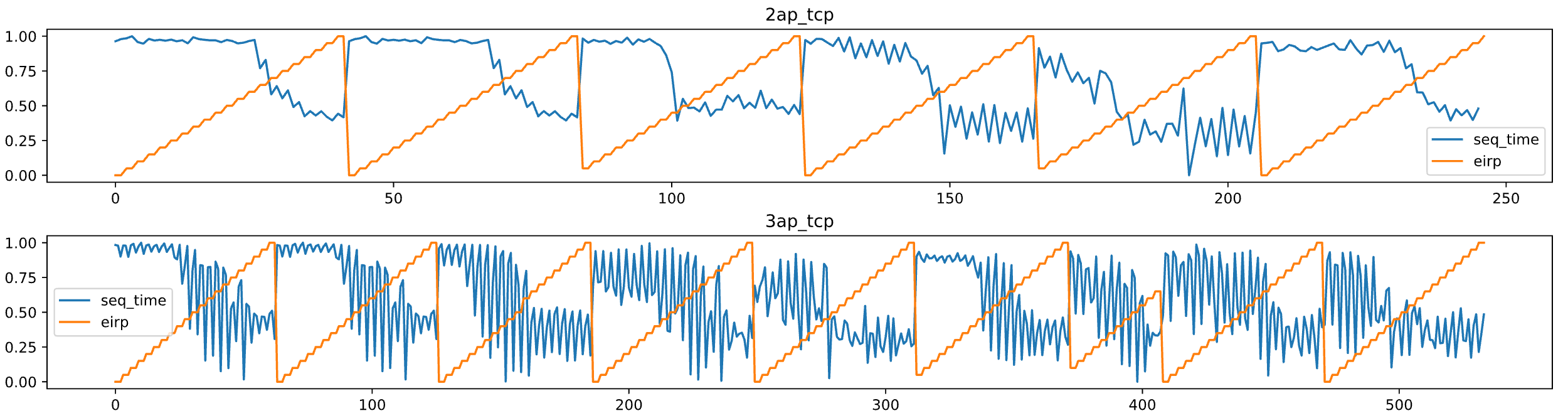
（1）随着发送功率的增大，AP发送机会在达到NAV门限的阈值后，将会出现AP间互相干扰的情况，因此在发送功率达到一定的阈值时，AP的发送机会出现明显下降。

（2）TCP模式相比于UDP模式其发送机会的波动性更大。

因此，发送功率与AP发送机会的相关性较高，作为我们主要关注的影响因子。



1. UPD模式下发送功率与AP发送机会的可视化结果



1. TCP模式下发送功率与AP发送机会的可视化结果

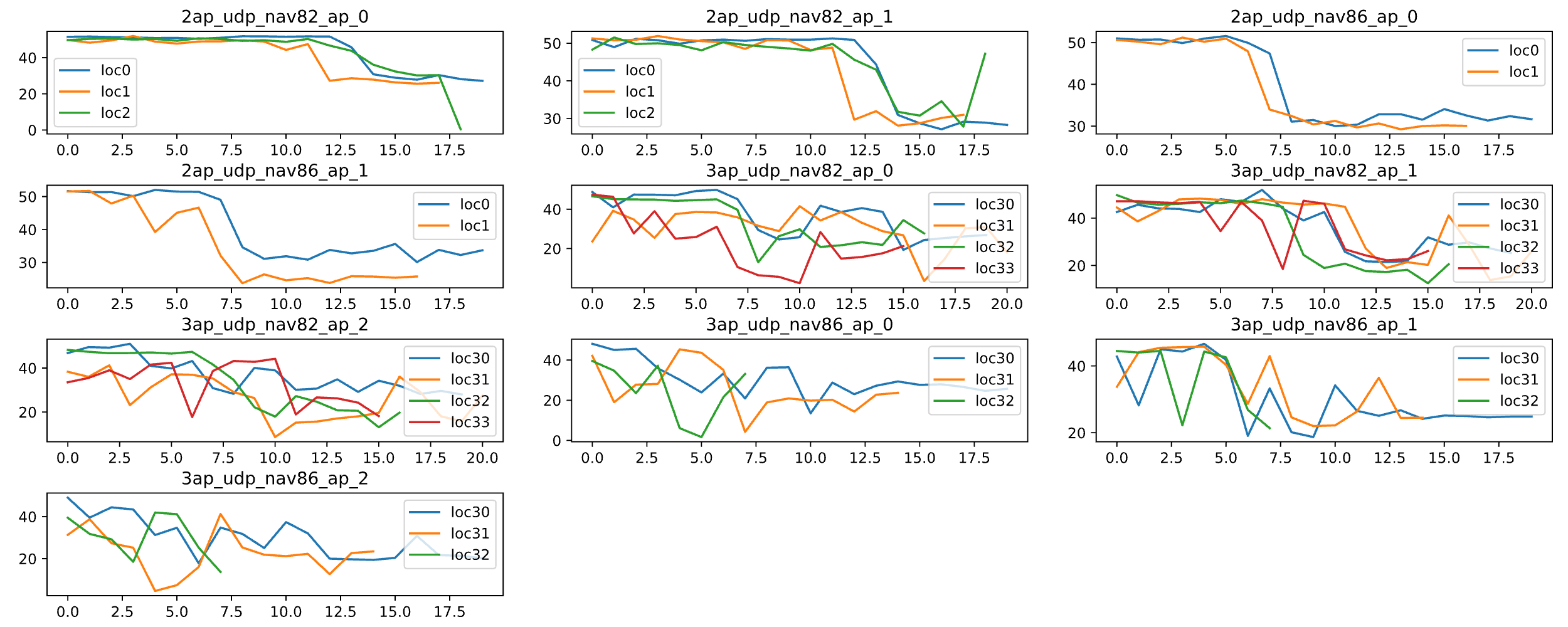
**部署位置（loc\_id）**：针对不同的部署位置，我们固定了NAV门限、流量类型、以及AP个数，对每个AP在固定场景下受到部署位置的影响进行了可视化分析，分析结果如下图X和图X所示。横坐标为逐渐递增的发送功率，纵坐标为AP发送机会。结果表明：

（1）在两个AP交互的场景下，部署位置对AP发送机会的影响较小，并且可以观察到loc1的位置发送机会明显较loc0，loc2偏低。

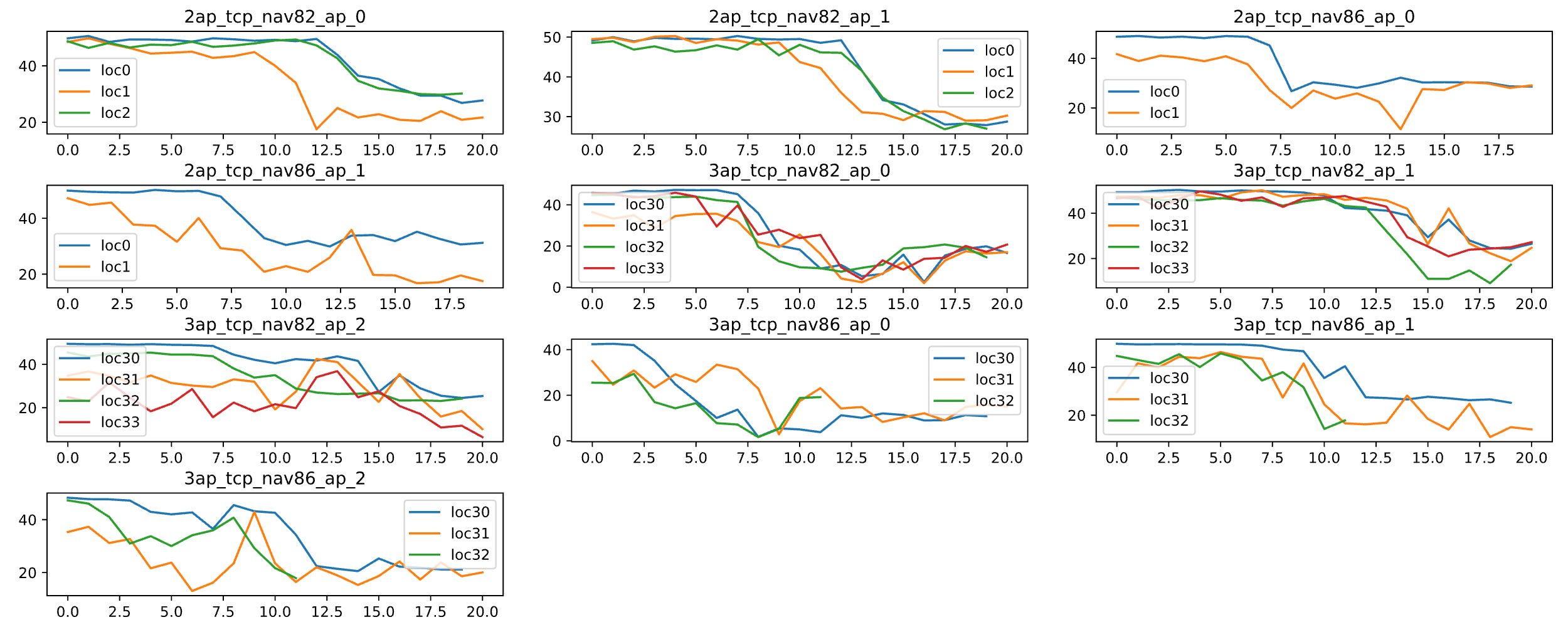
（2）在三个AP交互的场景下，部署位置对AP发送机会的影响较大，使其产生了更复杂的波动情况。这在loc33，loc31表现的更加明显。

（3）TCP模式相比UDP模式受到的影响更小，具有更强的稳定性。

因此，部署位置与AP发送机会的相关性较高，作为我们主要关注的影响因子。



1. UDP模式下不同部署位置对AP发送机会的影响



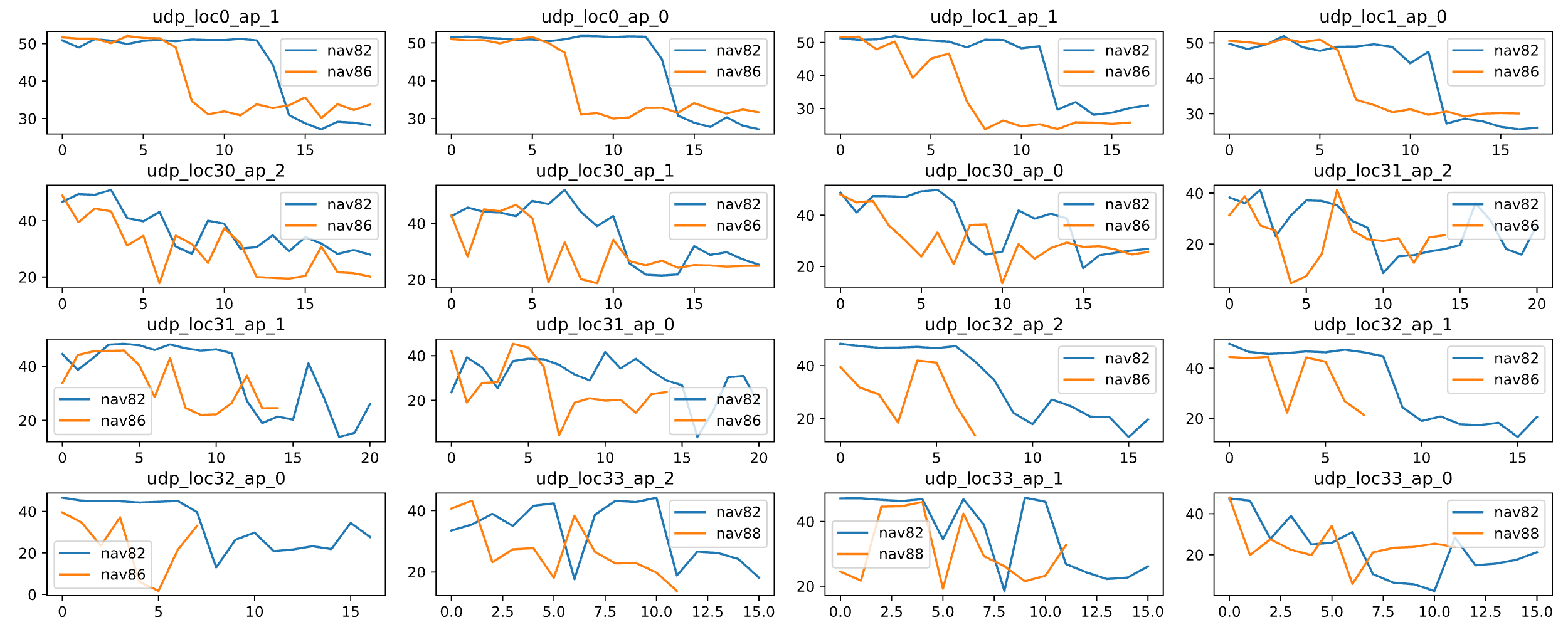
1. TCP模式下不同部署位置对AP发送机会的影响

**NAV门限（nav）**：针对不同的NAV门限，我们固定了部署位置与流量类型，对每个AP在固定场景下受到NAV门限的影响进行了可视化分析，分析结果如下图X和图X所示，横坐标为逐渐递增的发送功率，纵坐标为AP发送机会。结果表明：

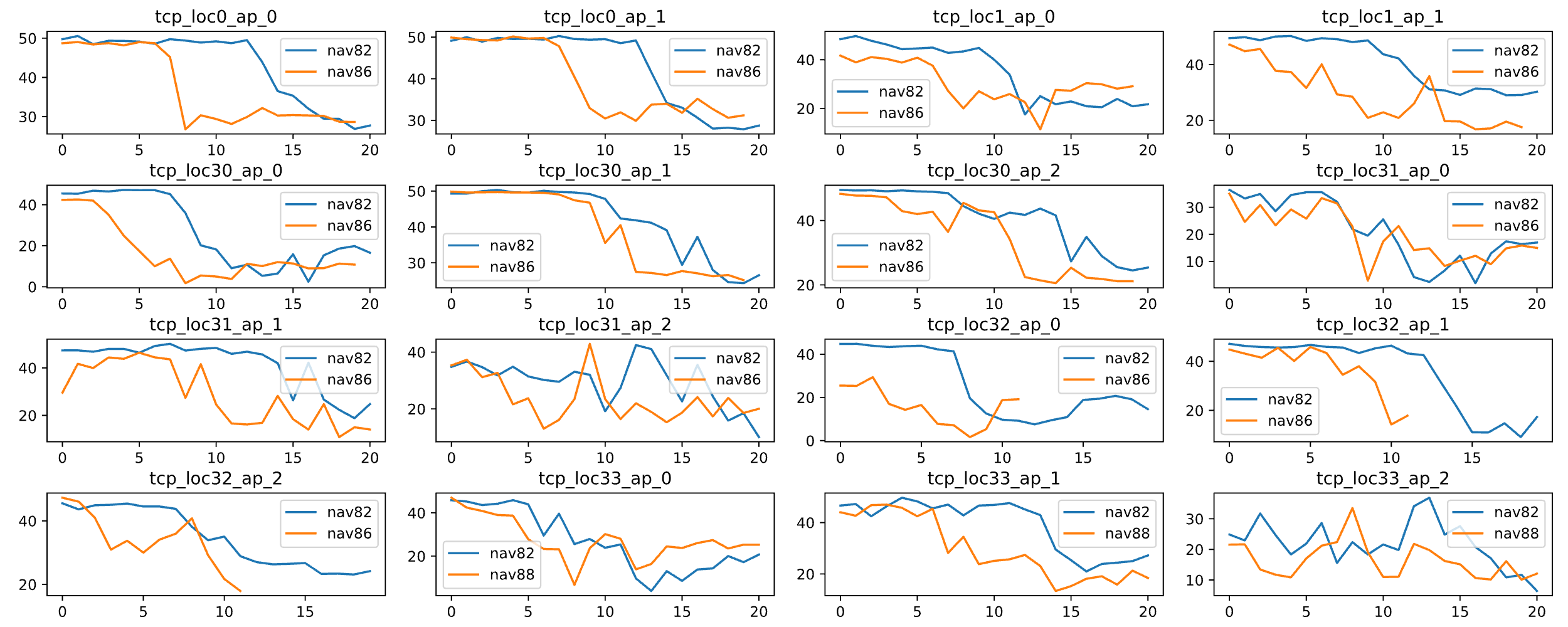
（1）NAV门限是NAV机制触发的关键因素，当AP接收到相邻BSS的数据帧时，将会使用RSSI与NAV门限判断是否接收数据。在实测数据集中，随着发送功率的升高，AP间发生数据传输和干扰的概率将会变大。而较低的NAV门限（数据集中为-82dBm与-86dBm），即nav86所标注的AP发送机会曲线会在更低的功率时触发AP间的NAV机制，从而对AP发送机会产生影响。

（2）TCP模式相比UDP模式呈现出更加稳定的趋势。

因此，NAV门限与AP发送机会的相关性较高，作为我们主要关注的影响因子。



1. UDP模式下不同NAV门限对AP发送机会的影响



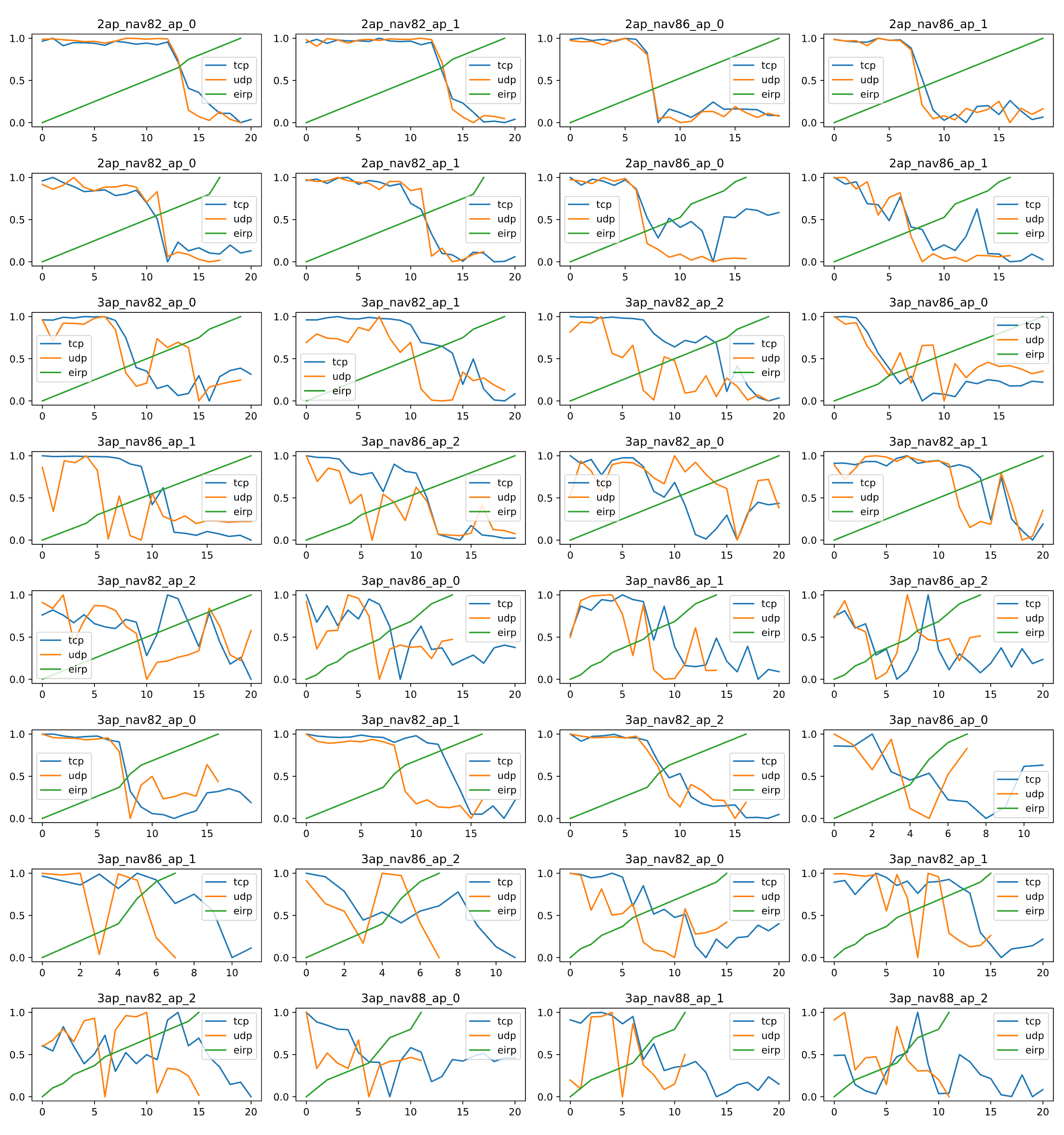
1. TCP模式下不同NAV门限对AP发送机会的影响

**流量类型（protocol）**：通过以上的分析，可以明显的观测到TCP模式与UDP模式在AP发送机会的稳定性上具有明显的差异。针对不同的流量类型，我们固定了部署位置与NAV门限，对不同流量类型的AP发送机会随发送功率的变化进行了可视化分析，结果如下图X所示。结果表明：

（1）UDP模式相比于TCP模式更容易受到其他影响因素的影响，呈现出更明显的波动性。

（2）UDP和TCP在发送功率达到一定阈值时，会出现明显的下降，这说明功率达到AP的NAV门限后，AP间出现明显的互相干扰。

因此，流量类型与AP发送机会的相关性较高，作为我们主要关注的影响因子。



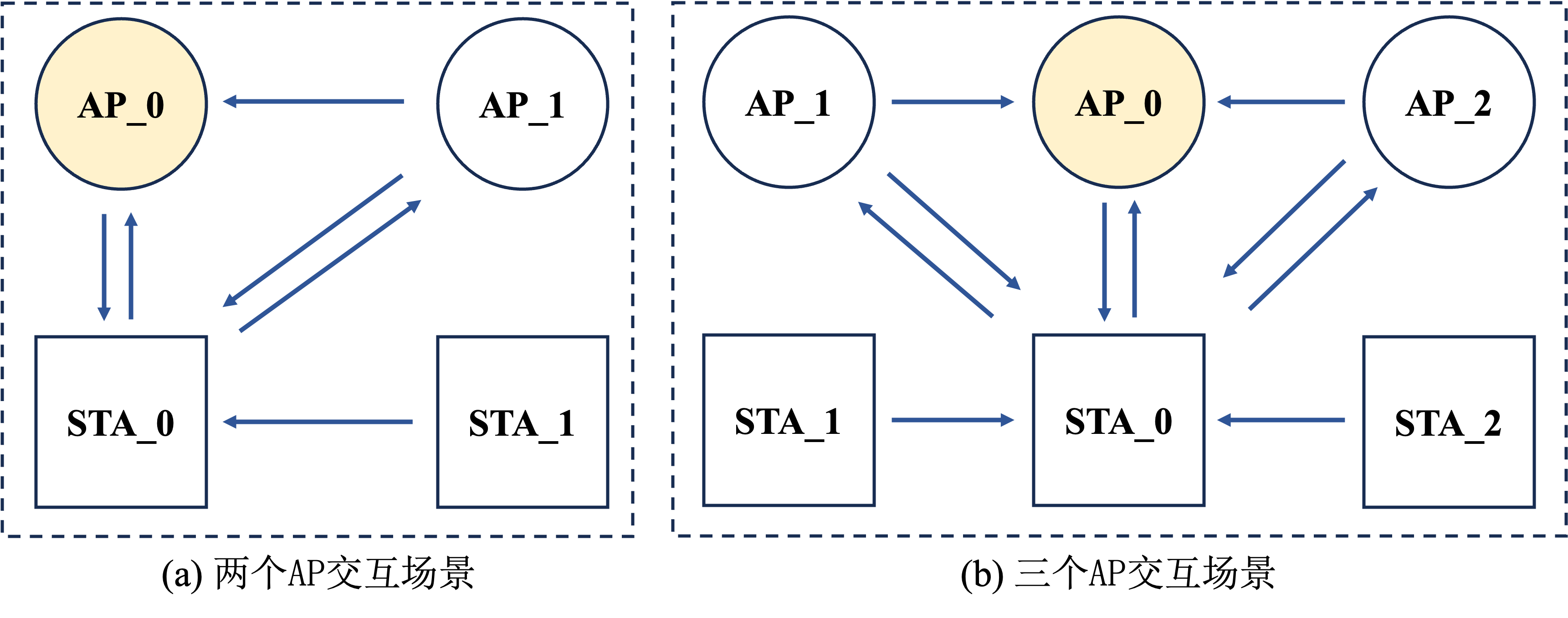
1. 不同的流量类型对AP发送机会的影响

**RSSI信号**：我们对数据集中的RSSI相关列进行分析，发现RSSI表达了两个节点（设备）间的信号强度，因此，我们将其视为设备间的距离因素。如下图X所示，基于数据集中的设备拓扑提取，主要分为两种交互场景：

（1）当有两个AP与两个STA设备时，如图X左侧，以AP\_0为主观测视角时的RSSI交互拓扑。结果表明，AP\_0的主要与STA\_0发生交互，同时会受到AP\_1的信号干扰，同时，考虑STA\_1的交互信息，可以将该场景视为AP\_0的空间位置信息，因此我们选用了图中的几条边的RSSI数据作为主要输入特征。

（2）当有三个AP和三个STA设备时，如图X右侧，以AP\_0为主观测视角时的RSSI交互拓扑。结果与两个AP的场景类似，我们选用了与之相邻的AP节点，以及STA\_0相邻的设备的RSSI数据作为主要输入特征。

因此RSSI信号与AP发送机会具有一定的相关性，作为我们主要关注的影响因子。



1. RSSI信号数据场景建模

## 影响性强弱排序模型

根据前面小节的分析，我们选取了与AP发送机会相关性最大的影响因子进行影响性强弱的排序，分别是部署位置（loc\_id），NAV门限（nav），发送功率（eirp），流量类型（protocol），RSSI。

**影响性强弱的定义**：根据对数据变化趋势的观察，我们发现AP的传输方式反映了AP间交互和干扰的情况。因此，我们将影响性强弱定义为影响因子使AP传输方式发生改变的临界点。

由于每个独立测试中，发送功率（eirp）为单调递增的变量，而发动功率的增加代表着AP间信号的增加，越高的功率将会使得AP间发生交互的概率增加。因此，我们通过寻找实测数据集中AP发送机会的明显转折点，统计AP发送机会变化的临界功率值，并记录部署位置（loc\_id）、NAV门限（nav）、发送功率（eirp）、流量类型（protocol）和RSSI，表X展示了部分的统计数据，其中RSSI选用ap\_from\_ap\_x\_mean的均值作为输入特征，其他特征则选用测试场景标注的信息。

1. 发送功率临界点统计信息

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **protocol** | **loc\_id** | **NAV** | **RSSI** | **erip** |
| tcp | loc0 | 82 | 85.02702703 | 21 |
| udp | loc0 | 82 | 85.02702703 | 21 |
| tcp | loc0 | 86 | 87.75 | 15 |
| udp | loc0 | 86 | 87.75 | 15 |
| tcp | loc1 | 82 | 83.54651163 | 18 |
| udp | loc1 | 82 | 80.59756098 | 21 |
| tcp | loc1 | 86 | 86.71428571 | 15 |
| udp | loc1 | 86 | 86.71428571 | 15 |
| tcp | loc2 | 82 | 84.30588235 | 21 |
| udp | loc2 | 82 | 84.45348837 | 20 |
| tcp | loc30 | 82 | 82.85 | 17 |
| udp | loc30 | 82 | 87.21568627 | 16 |
| tcp | loc30 | 86 | 81.27692308 | 18 |
| udp | loc30 | 86 | 88.85245902 | 15 |
| tcp | loc31 | 82 | 81.27692308 | 19 |
| udp | loc31 | 82 | 84.01639344 | 20 |
| tcp | loc31 | 86 | 78.95 | 18 |
| udp | loc31 | 86 | 86.09090909 | 16 |
| tcp | loc32 | 82 | 81.04038005 | 21 |
| udp | loc32 | 82 | 86.54545455 | 16 |
| tcp | loc32 | 86 | 84.19469027 | 17 |
| udp | loc32 | 86 | 85.34939759 | 16 |
| tcp | loc33 | 82 | 79.15776699 | 22 |
| udp | loc33 | 82 | 83.9516129 | 19 |
| tcp | loc33 | 88 | 85.07334963 | 15 |
| udp | loc33 | 88 | 88.45514223 | 15 |

**影响性强弱的排序**：根据上表的统计信息，我们使用主成分分析法（PCA）对这些影响因子进行了分析。表X展示了这些影响因子在每个主成分中的贡献情况，表X展示了每个主成分解释的方差比例。

1. 影响因子主成分分析的贡献度统计

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **影响因子** | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** |
| protocol | 0.2829 | 0.5022 | -0.6761 | -0.4407 | 0.1276 |
| loc\_id | 0.0345 | 0.6331 | 0.7027 | -0.3003 | 0.1177 |
| NAV | -0.4954 | 0.4538 | -0.1911 | 0.6035 | 0.3844 |
| RSSI | -0.5601 | -0.3195 | -0.0102 | -0.5564 | 0.5238 |
| eirp | 0.5995 | -0.1969 | 0.1112 | 0.2042 | 0.7401 |

1. 影响因子主成分解释的方差比例

|  |  |
| --- | --- |
| **主成分** | **解释方差比例** |
| PC1 | 0.4706 |
| PC2 | 0.2458 |
| PC3 | 0.1922 |
| PC4 | 0.0614 |
| PC5 | 0.0298 |

通过以上分析，结合我们在AP发送机会的预测模型测试结果（即RSSI与AP发送机会的预测具有更强的相关性这一结论，我们将在下一小节进行解释）。我们认为RSSI是影响最大的因素，其次，eirp在主成分分析中占比最大，因此排在第二。NAV在PC1与PC2中的加权贡献值大于loc\_id，排在第三。Loc\_id在第二个成分中排名第一，排在第四。Protocol则排在末尾。因此，我们认为最终的影响性强弱的关系为：

|  |
| --- |
| **RSSI** > **eirp** > **NAV** > **loc\_id** > **protocol** |

## AP发送机会预测模型

根据前两个小节的分析，我们选用了具有相关性的特征进行预处理，作为输入特征建立预测模型。同时，由于观测到测试数据集（test\_set\_1\_\*, test\_set\_2\_\*）中的部署位置与训练数据集中不一致，所以排除了部署位置这个输入特征。因此，我们最终选用了NAV门限（nav），发送功率（eirp），流量类型（protocol），RSSI作为模型的输入特征。AP发送的序列时间（seq\_time）作为模型预测的输出。

**特征预处理**：我们首先对选用的特征进行了预处理，其中NAV门限为每个实验的固定值，流量类型为TCP和UDP的二分类数据，进行0/1编码，发送功率则采用原始数值。针对RSSI，我们将其认为是节点间距离的表达，因此我们选用了AP与AP间，AP与STA间的RSSI均值作为输入特征。

**模型建立**：由于选择的输入特征为多维数据，是典型的数据预测问题。因此，我们选择了几个通用的基于机器学习和逻辑回归的预测模型进行对比。分别是线性回归、XGBoost、随机森林、支持向量回归、全连接神经网络。同时，由于观测到两个AP与三个AP交互场景下数据的差异较大，我们分别设计了2AP和3AP的预测模型。

**模型训练**：我们将训练集按照8:2的比例进行划分，其中训练集为80%，测试集为20%，并使用均方误差（MSE）评估模型在测试数据集上的效果。模型训练精度如下表X和表X所示。

1. 2AP发送机会预测模型均方误差对比

|  |  |
| --- | --- |
| 模型名称 | 均方误差（MSE） |
| 线性回归（LinearRegression） | 29.23 |
| XGBoost回归（XGBoostRegression） | 10.87 |
| 随机森林回归（RandomForestRegression） | 15.28 |
| 支持向量回归（SVR） | 33.12 |
| 全连接神经网络（MLP） | 13.18 |

1. 3AP发送机会预测模型均方误差对比

|  |  |
| --- | --- |
| 模型名称 | 均方误差（MSE） |
| 线性回归（LinearRegression） | 80.11 |
| XGBoost回归（XGBoostRegression） | 39.52 |
| 随机森林回归（RandomForestRegression） | 50.42 |
| 支持向量回归（SVR） | 73.81 |
| 全连接神经网络（MLP） | 39.69 |

根据上表中均方误差的对比，我们选用了XGBoost、随机森林、全连接神经网络作为主要优化的模型。经过模型调优我们发现，由于训练数据集较小，全连接神经网络容易出现过拟合，因此在该数据尺度下，我们选择参数调试后效果更优的XGBoost模型作为最终的预测模型。

## 问题一的预测结果

# 问题二的模型建立与求解

# 问题三的模型建立与求解

## 数据预处理

数据集分类按2ap、3ap以及training、testing分类，2ap training共5个文件。3ap traning共8个文件。2ap test共2个文件，3ap test共2个文件。其中test文件set\_1用于问题一和问题三，set\_2用于问题二。

首先对所有文件数据进行分析，情况如下：

训练集：

1. 公有数据有22种指标，包括：test\_dur,loc\_id,bss\_id,ap\_name,ap\_mac,ap\_id,sta\_mac,sta\_id,protocol,pkt\_len,pd,ed,nav,seq\_time,nss,mcs,per,num\_ampdu,ppdu\_dur,other\_air\_time,seq\_time,throughput。其中常量指标包括：test\_dur,pkt\_len,pd,ed,seq\_time，变量指标包括loc\_id,protocol,nav,rssi，测量统计指标包括：nss,mcs,per,num\_ampdu,ppdu\_dur,other\_air\_time,seq\_time,throughput。对于rssi，每种rssi均有sum、max、mean3类，且每类指标是时序数据，且数据长度不固定。

针对2ap，以ap\_0为例，含以下指标：ap\_from\_ap\_1\_sum\_ant\_rssi,ap\_from\_ap\_1\_max\_ant\_rssi,ap\_from\_ap\_1\_mean\_ant\_rssi,

sta\_to\_ap\_0\_sum\_ant\_rssi, sta\_to\_ap\_0\_max\_ant\_rssi, sta\_to\_ap\_0\_mean\_ant\_rssi

sta\_to\_ap\_1\_sum\_ant\_rssi, sta\_to\_ap\_1\_max\_ant\_rssi, sta\_to\_ap\_1\_mean\_ant\_rssi

sta\_from\_ap\_0\_sum\_ant\_rssi, sta\_from\_ap\_0\_max\_ant\_rssi, sta\_from\_ap\_0\_mean\_ant\_rssi

sta\_from\_ap\_1\_sum\_ant\_rssi, sta\_from\_ap\_1\_max\_ant\_rssi, sta\_from\_ap\_1\_mean\_ant\_rssi

sta\_from\_sta\_1\_rssi

针对3ap，以ap\_0为例，含以下特征：

ap\_from\_ap\_1\_sum\_ant\_rssi,ap\_from\_ap\_1\_max\_ant\_rssi,ap\_from\_ap\_1\_mean\_ant\_rssi,

ap\_from\_ap\_2\_sum\_ant\_rssi,ap\_from\_ap\_2\_max\_ant\_rssi,ap\_from\_ap\_2\_mean\_ant\_rssi,

sta\_to\_ap\_0\_sum\_ant\_rssi, sta\_to\_ap\_0\_max\_ant\_rssi, sta\_to\_ap\_0\_mean\_ant\_rssi

sta\_to\_ap\_1\_sum\_ant\_rssi, sta\_to\_ap\_1\_max\_ant\_rssi, sta\_to\_ap\_1\_mean\_ant\_rssi

sta\_to\_ap\_2\_sum\_ant\_rssi, sta\_to\_ap\_2\_max\_ant\_rssi, sta\_to\_ap\_2\_mean\_ant\_rssi

sta\_from\_ap\_0\_sum\_ant\_rssi, sta\_from\_ap\_0\_max\_ant\_rssi, sta\_from\_ap\_0\_mean\_ant\_rssi

sta\_from\_ap\_1\_sum\_ant\_rssi, sta\_from\_ap\_1\_max\_ant\_rssi, sta\_from\_ap\_1\_mean\_ant\_rssi

sta\_from\_sta\_1\_rssi

sta\_from\_ap\_2\_sum\_ant\_rssi, sta\_from\_ap\_2\_max\_ant\_rssi, sta\_from\_ap\_2\_mean\_ant\_rssi

sta\_from\_sta\_1\_rssi,sta\_from\_sta\_2\_rssi

测试集：

测试集中没有测量统计指标，但set\_1含有nss,mcs,per，用于问题三。

本文使用training文件，分为2ap、3ap两类情况分别建模解决各个问题。

问题一：protocol,nav,rssi作为模型输入集，seq\_time为标签集。

问题二：protocol,nav,rssi作为模型输入集，(nss,mcs)组合为标签集。

问题三：protocol,nav,rssi,phy作为模型输入集，throughput为标签集。此处phy是通过（MCS, NSS）与PHY Rate对应表得到。

数据分配，取训练集与测试集比例为9:1。

1. 异常值处理

本文使用Z-Score方法，以3为阈值，筛选异常值。



对于输入集：未发现异常值。

对于标签集：training\_set\_2ap\_loc2\_nav82.csv中test\_id=39的seq\_time为0.4，属于异常点，舍弃此数据。

1. 缺失值处理

使用pandas.info()查看各指标缺失值信息。训练集和测试集中均有缺失值，用均值函数（mean）处理。对于protocol缺失，因无法判定属于类别，舍弃此数据。对于标签集缺失的情况，无法用作训练，舍弃数据。

对于输入集：无protocol缺失情况，其余缺失已用均值函数处理。

对于标签集：training\_set\_2ap\_loc2\_nav82.csv中test\_id=40的点确实标签集，舍弃此数据。

1. 指标归一化

对于输入集：采用MinMax()归一化到[0,1]，消除特征量纲的影响。

对于标签集：无需归一化



1. rssi时间序列单值化

rssi类型为时间序列，长度不固定，且某些列为单个值。本文求均值将此指标单值化。

1. protocol与(nss,mcs)编码

对于protocol：只有两类udp,tcp，通过独热（One-Hot）编码转为[1,0],[0,1]的形式，加入到模型输入向量。

对于(nss,mcs)：两者组合能够枚举，将 nss 和 mcs 合并为一个标签nss\_mcs，使用LabelEncoder编码转为整数，方便模型训练预测，最后拆分为nss、mcs。

基于问题一特征分析，RSSI、发送功率（eirp）、NAV门限和流量类型（protocol）这些因子会影响发送机会，进而影响AP的调制编码方案（mcs）和空间流数（nss），所以对于问题二选用相同的特征。

对于问题三，为了预测系统吞吐量，在问题一、二的基础上，特征添加了实测中统计的数据帧真实空间流数和调制编码方案，最终选用特征RSSI、发送功率（eirp）、NAV门限和流量类型（protocol）、调制编码方案（mcs）和空间流数（nss）。

按照同频AP个数分类建模，将训练集进行8：2随机划分，前者用于算法训练，后者用于验证算法准确率。全面评估了线性回归、XGBoost回归、随机森林回归、全连接神经网络，预测发送数据帧序列的总时长（seq\_time）的表现，并通过4 种评估标准自适应地选择了最优的 XGBoost模型，多次实验表明，收敛后的模型在随机划分的测试上的均方根误差低于 0.07，同时 R2 Score 超过了 0.72。

# 模型评价与展望

## 创新点

## 优点

## 缺点

## 展望

# 参考文献

# 附录