**一，利用的数据**

本场赛题共有六张表的数据：a.   连接WIFI AP （Access Point）的人数表 airport\_gz\_wifi\_ap，b.   安检旅客过关人数表 airport\_gz\_security\_check，c.   旅客进入－离开机场的行程表 airport\_gz\_departure，d.    航班排班表airport\_gz\_flights。e.  机场登机口区域表airport\_gz\_gates,f.  机场WIFI接入点坐标表 airport\_gz\_wifi\_coor考虑到要预测未来两天的WIFI AP 连接数量，连接WIFI AP （Access Point）的人数表无疑是最有利用价值的一张表，而安检旅客过关人数表和旅客进入－离开机场的行程表对于预测未来几个小时内的很有价值，但是对于预测未来两天这样长的时间段应该作用不大。 航班排班表与WIFI AP连接情况尤其是登机口附近的WIFI AP相关性很大，配合机场登机口区域表和机场WIFI接入点坐标表应该能较好地从时间和空间上预测WIFI AP的连接情况。所以，我们选用a,d,e,f这四张表来进行建模。

**二，数据预处理**

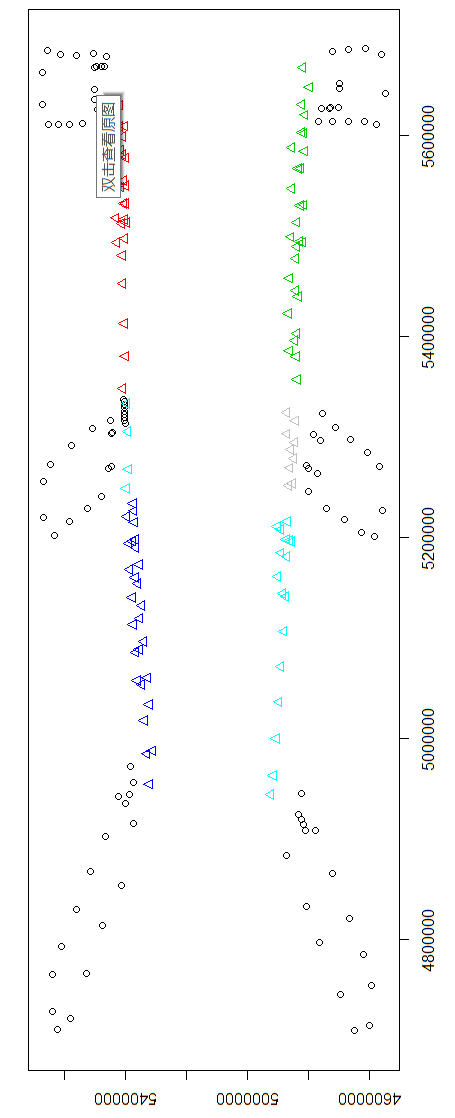
本赛题的数据预处理主要包含：1，连接WIFI AP （Access Point）的人数表存在缺失数据情况，对于个别时间段的数据缺失，可以利用填充0或者前后几天均值来进行填充，对于大段时间段的数据缺失，我们考虑以天为单位删除数据，不予利用。2，航班排班表中的计划起飞时间和实际起飞时间与其他表的时间有八小时误差，我们进行了处理。3，机场WIFI接入点坐标表中WIFI AP 坐标没有办法直接和登机口联系起来，我们先将这些点的坐标回归拟合成经纬度坐标，然后可以和登机口经纬度坐标联系起来，如图一。



图一：WIFI AP和登机口位置信息

**三，预测方案**

1. 整体预测，每个WIFI AP的在某一时间段的连接数具有很强的时间周期性，所以用历史数据来进行回归预测对于绝大多数AP来说就已经能很好地拟合，而且过拟合的情况不严重。
2. 通过第一步的整体预测，大部分的WIFI AP已经的到了一个相对较优的预测值，但是对于一些AP，尤其是在登机口附近的WIFI AP由于受到飞机航班起降的影响非常严重，预测地不是很好，需要进一步进行修正学习。我们将第一步的预测值和实际值作为一个残差，学习并预测这个残差，就可以进一步提高预测的精度，而每一个WIFI AP附近的登机口是不同的，而且由于其他空间因素的影响，不同WIFI AP之间的数据几乎无法互相利用，所以需要对于每个WIFI AP单独建模，回归出残差。对于E1，E2，E3，W1，W1，W3这六个区域的WIFI AP，是直接的候机休息区，会直接收到距离其很近的登机口的航班起落影响，于是我们根据WIFI AP和登机口的经纬度坐标，计算距离，求出每个WIFI AP可能会受到那几个登机口的影响，进一步提取特征，训练模型。对于EC和WC区域，是走向候机区的通道，会受到某一个区域所有航班的影响，我们将其细分成了六个区域（如图一），分别提取特征，每个WIFI AP单独训练,预测各自残差。



图二：EC和WC区域细分

1. 登机口附近AP的线性回归预测方案。在登机口附近的AP的wifi连接人数，可以近似认为和等待登机的人数成正比。当然某个AP可能受到不止一个登机口的影响。即

（1）

其中其中y(t)为t时刻某AP的wifi连接人数，zi(t)是登机口等待登机人数，ai是相关系数，i代表某AP附近的i个登机口。显然这是一个容易求解的线性回归问题。

那么是否可以通过航班的起飞时间和登机人数，大致估算出每一时刻，等待登机的人数呢？既求取函数f：

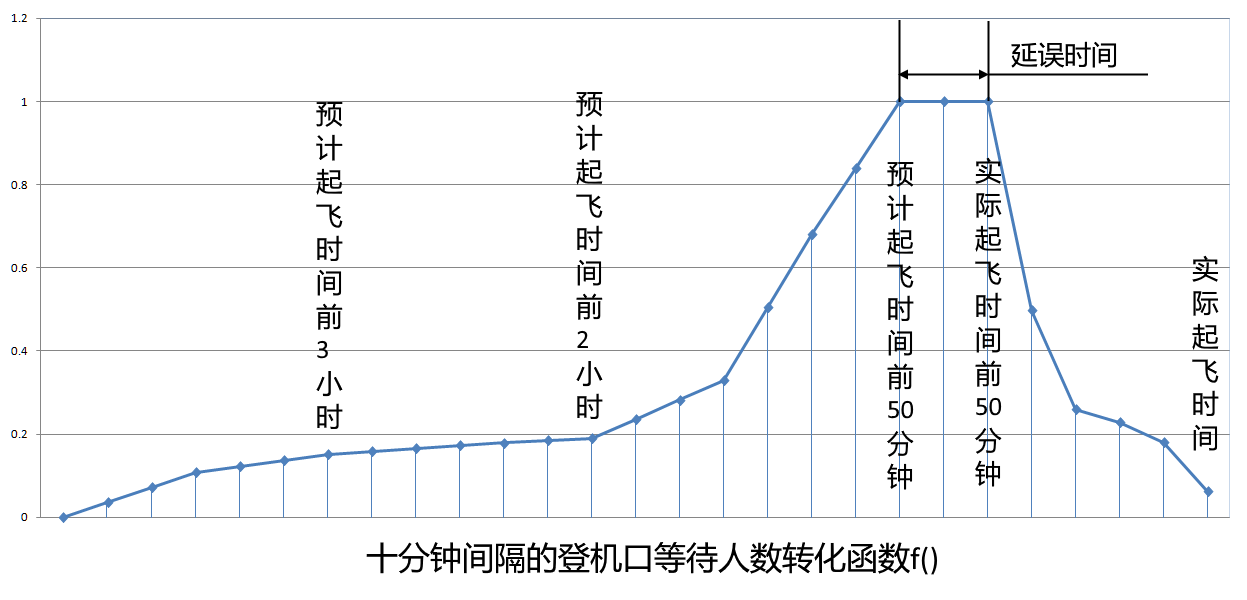
（2）

其中xi是登机口对应的航班信息，包括航班起飞时间和登机人数。我们注意到一层的登机口非常密集，而等待区是一个独立而封闭的区域。我们可以将一个区域看成是一个整体，那么有

（3）

其中j代表一个区域内的AP数量，k代表一个区域内登机口的数量。根据(3)，我们可以通过最优化问题求解f0：

其中cov是person相关系数。具体求解我们使用了粒子群优化算法(PSO)。求解结果



**四，特征提取**

考虑每个WIFI AP各自建模，特征也不尽相同，主要分为几个方面，该WIFI AP的历史连接情况，最大值，最小值，平均值，中位数等等；可能影响该WIFI AP的几个登机口的过去一段时间（过去三小时，过去一小时，过去半小时，过去十分钟……）,最近一段时间（最近三小时，最近一小时，最近半小时，最近十分钟……），未来一段时间（未来三小时，未来一小时，未来半小时，未来十分钟……），各有多少个航班起降；距离该WIFI AP的最近的登机口在过去，最近，未来的飞机起飞情况；距离该WIFI AP的第二近的登机口在过去，最近，未来的飞机起飞情况等等。

**五，模型选择**

由于每个WIFI AP需要单独建模，其实每个模型的数据实际上非常有限，数据的稀缺导致模型容易过拟合，所以，我们选择线性回归这样简单的模型来防止过拟合的发生。同时使用L2正则化项来进一步防止过拟合。