**确定数据的整定，样本筛选逻辑方法**

**1 数据采集**

原始数据采集来自于中石化高桥石化实时数据库（霍尼韦尔PHD）及LIMS实验数据库。其中操作变量数据来自于实时数据库，采集时间为2017年4月至2020年5月，采集操作位点数共354个。2017年4月至2019年9月，数据采集频次为3分钟/次；2019年10月至2020年5月，数据采集频次为6分钟/次。原料、产品和催化剂数据来自于LIMS实验数据库，数据时间范围为2017年4月至2020年5月。其*中原料及产品的辛烷值是重要的建模变量，该数据采集频次为每周2次*。

**2 数据整定**

原始数据中，大部分变量数据正常，但每套装置的数据均有部分位点存在问题：部分变量只含有部分时间段的数据，部分变量的数据全部为空值或部分数据为空值。因此对原始数据进行处理后才可以使用。数据处理方法如下：

（1）对于只含有部分时间点的位点，如果其残缺数据较多，无法补充，将此类位点删除；

（2）删除325个样本中数据全部为空值的位点；

（3）*对于部分数据为空值的位点，空值处用其前后两个小时数据的平均值代替*；

**（4）**根据工艺要求与操作经验，总结出原始数据变量的操作范围，然后采用最大最小的限幅方法剔除一部分不在此范围的样本；

**（5）**根据拉依达准则（3σ准则）去除异常值。

3σ准则：设对被测量变量进行等精度测量，得到x1，x2，……，xn，算出其算术平均值x及剩余误差vi=xi-x（i=1,2,...,n），并按贝塞尔公式算出标准误差σ，若某个测量值xb的剩余误差vb（1<=b<=n），满足|vb|=|xb-x|>3σ，则认为xb是含有粗大误差值的坏值，应予剔除。贝塞尔公式如下：



**3 样本确定**

本题目标为降低S Zorb装置产品辛烷值损失，故确定样本的主要依据为样品的辛烷值数据。由于辛烷值的测定数据相对于操作变量数据而言相对较少，而且辛烷值的测定往往滞后，因此确定某个样本的方法为：以辛烷值数据测定的时间点为基准时间，取其前2个小时的操作变量数据的平均值作为对应辛烷值的操作变量数据。