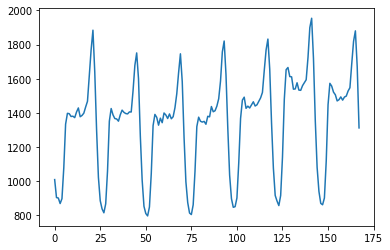
## 基本数据分析

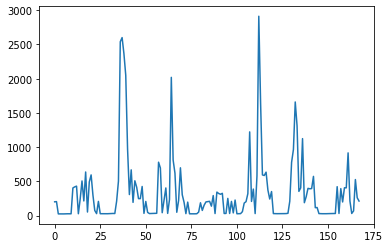
通过对数据进行初步分析，可以知道数据有120组，每组数据是一段长达168的序列。有少量机组数据存在缺失值，以“NA”表示。每组数据的范围各不相同，最大的从0到384410。

## 数据形态分析

随机选取数据进行绘图，可以发现有些数据呈现出明显的周期性。比如第0组数据，如下图所示：



而有些数据的周期性不是特别明显，具有强烈的震荡，如第3组数据：



## 模型训练

**数据插值**

由于少量数据存在缺失情况，采用线性插值填补缺失数据。具体插值方法如下：

1. 如数据从第0项开始缺失到第n项，则用第n+1项的值填补前面所有缺失值。
2. 如数据从第m项确实到第n项，则m到n之间的值用m-1项和n+1项的值进行线性插值。
3. 如数据从第n项缺失到数据末尾，则通过第n-1项的值填补之后的所有缺失数据。

**数据的归一化**

由于各组数据的范围各不相同，为了让每组数据的权重相接近，我先分别对每组数据进行min-max归一化，使其处于0-1的区间内。

**模型选择**

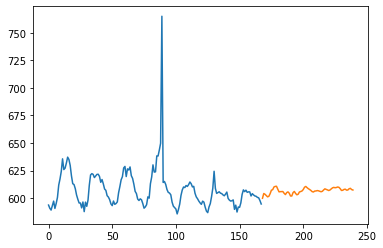
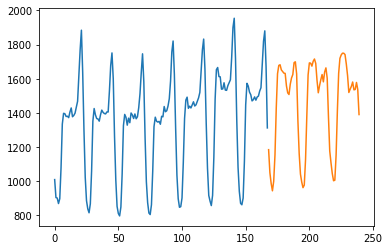
通过前面的分析可知，数据大多具有很强的周期性。即当前时刻的数据与周期T之前的数据有很强的相关性。因此可以采用基于卷积神经网络的序列预测模型。采用均方误差作为优化目标函数，

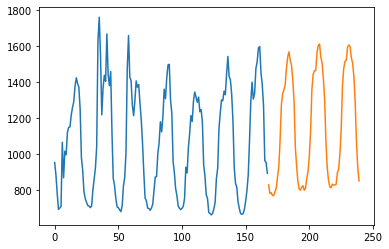
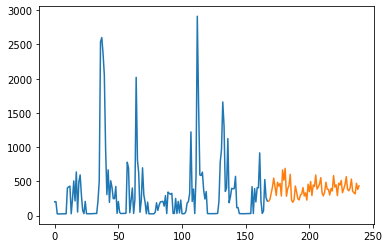
**模型训练**

为了防止模型出现过拟合的情况，把20组数据作为验证集。训练时采用100组数据进行训练。每次训练时用16组数据，从中随机截取长度为63的序列进行序列，预测下一个值。训练轮数为10000

## 预测结果

经过训练后，在测试集上的均方误差降到了0.03左右。因此停止训练，并对数据进行预测。预测的第1-4组数据如图（其中蓝色是原始数据，黄色表示预测的数据）：





可见对于周期性数据的预测较符合逻辑，而对于有些震荡较为明显的数据预测则会趋于平缓。整体预测效果尚可。