时间序列数据预测

一、术语

时间序列是以规律的时间间隔采集的测量值的有序集合，例如，每日的股票价格或每周的销售数据。测量值可以是您感兴趣的任何内容。

时间间隔可以代表任何时间单位，但所有测量值的时间间隔必须相同。而且，没有测量值的任何时间间隔必须设置为缺失值。因此，有测量值的时间间隔数（包括测量值为缺失值的时间间隔）定义数据历史记录范围的时间长度。

二、背景

当今社会互联网发展越发成熟，其上运行着无法统计的各种系统，而这些系统每时每刻都在产生海量的数据。基于这些海量数据，我们可以做很多事情，比如预测未来数据走向。就像天气预报一样，准确的预测可以带来巨大的价值。

现在，对系统运行中产生的海量历史数据进行分析，进而准确预测未来一段时间内的数据走向，已经是很常见的大数据应用方式了。

其中关键的核心逻辑就是：基于一定的时间序列的历史数据，来预测未来一段时间的可能值（可以叫做预测值）；当预测值对应的时间点到来时，会采集到对应的测量值；然后通过计算测量值和预测值的偏差度，来衡量预测值对应的时间点是否发生异常。

三、 赛题

1.本赛题会给出一条时间序列数据，作为历史数据供参赛者分析。参赛者分析之后需要按时间顺序给出未来 10 个点的预测值。

2.赛题给出的时间序列数据为 CSV 文本格式，文件名为 data1.csv，示例部分会有样例数据供参考。

3.最终考核时给出的数据量可能会较大，所以需要参赛者以程序作答，输入为给出的 CSV 文本，输出为同样格式的 CSV 文本文件，输出文件名为 result1.csv。

4.给出的时间序列会存在噪声，噪声为一个或一些采集时间点相近的测量值，这些测量值与整段时间序列不具有相同的规律和特征，需要识别出并做降噪处理。

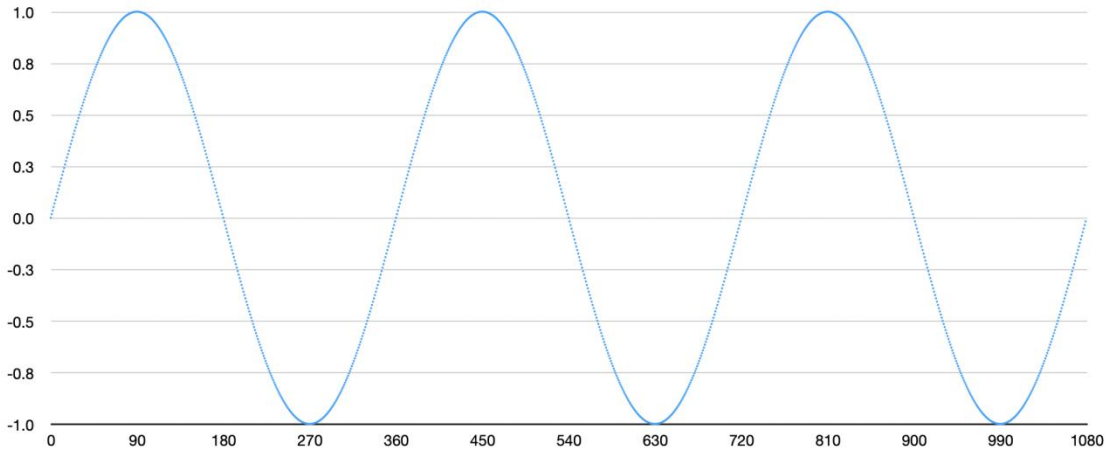
干扰点的判定、识别、降噪逻辑由参赛者自行确定方案并实现。

四、示例

示例一：

时间序列数据见附件 data1-1.csv

数据内容如图：



基于该时间序列历史数据，按顺序预测未来 10 个点的数据值。

示例答案：

输出的 CSV 文件内容：

0.0

0.01745240643728351

0.03489949670250097

0.05233595624294383

0.0697564737441253

0.08715574274765817

0.10452846326765346

0.12186934340514748

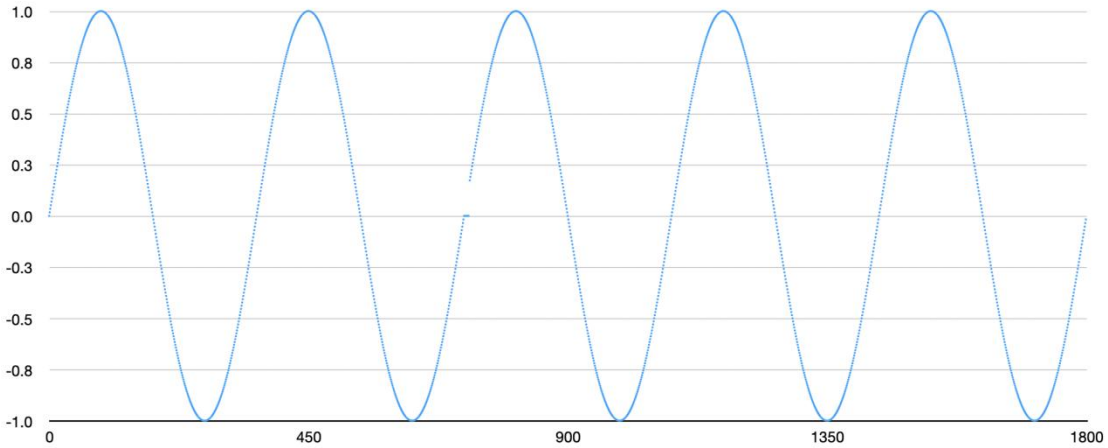
0.13917310096006544

0.15643446504023087

示例二：

时间序列数据见附件 data1-2.csv

数据内容如图：



基于该时间序列历史数据，按顺序预测未来 10 个点的数据值。

示例答案：

输出的 CSV 文件内容：

0.0

0.01745240643728351

0.03489949670250097

0.05233595624294383

0.0697564737441253

0.08715574274765817

0.10452846326765346

0.12186934340514748

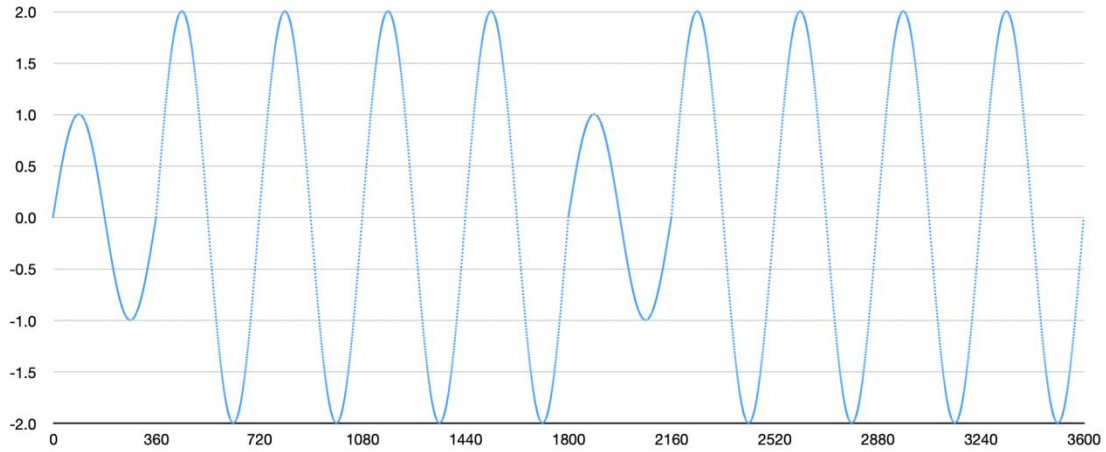
0.13917310096006544

0.15643446504023087

示例三：

时间序列数据见附件 data1-3.csv

数据内容如图：



基于该时间序列历史数据，按顺序预测未来 10 个点的数据值。

示例答案：

输出的 CSV 文件内容：

0.0

0.01745240643728351

0.03489949670250097

0.05233595624294383

0.0697564737441253

0.08715574274765817

0.10452846326765346

0.12186934340514748

0.13917310096006544

0.15643446504023087

五、提示

1. 考虑数据规律性，且不同的时间序列数据可能具有不同的规律。
2. 需要合理设计噪声判定及降噪逻辑，这会影响规律发现及预测逻辑。
3. 规律性可能不止一种，需要考虑多种规律同时存在的情况。

六、考核

1.参赛者需要在 5 月 26 日 23:59 前提交程序文件到指定大赛官网或发送至大赛邮箱 shxrt@fusionskye.com，程序需要包括运行说明、一键执行脚本。举办方会对提交的程序进行必要的审查。

2.考核时会给出一条全新的时间序列数据，包含不同规律、且含有不定数量噪声。大赛举办方会统一运行参赛者提交的所有合格程序，运行方式为执行程序中的一键执行脚本。之后会读取程序输出的 CSV 文件，并比较文件内容和参考答案的偏差，偏差越小则得分越高。

3.程序语言限定为 C、C++、Java、Python，且原则上不考核性能、资源消耗等因素，重点只关注预测的准确性。只有当多组答案相同时，会参考程序设计合理性、性能、资源消耗等因素。

4.程序运行环境为 CentOS7，具有 C（gcc 8.2）、C++（clang 11）、Java（Java SE 17）、Python（Python 3.9）运行环境，运行时不可联网。