2020年河北省第三届研究生数学建模竞赛

题 目：交通检测器数据质量控制及预测

摘 要

本文针对交通检测器采集原始数据的数据质量问题，运用**XX法和XX法**建立了**XX**数学模型建立了储油罐的变位识别与罐容表标定的计算模型，分别对三种不同变位情况推导出的油位计所测油位高度与实际罐容量的数学模型，运用matlab软件编程得出合理的结论，最终对模型的结果做出了误差分析。

针对问题一要求根据给定数据（data1）建立识别问题数据的方法和流程，完成对问题数据的修正及缺失数据的填补。我们先根据题意设计了识别问题数据的方法，并转化为流程图展示出来，

针对问题二

针对问题三

本文主要应用**XX**软件对相关的模型进行编程求解，计算方便、快捷、准确，整篇文章图文并茂。文章最后根据所建立的模型用**XX数据**进行了误差分析，结果可靠，使得模型具有现实意义。

关键词

交通流参数；Python；多元回归。

1 问题重述

**1.1 背景知识**

1. 城市道路中大量应用交通检测器用于采集断面交通数据,采集得到的数据质量存在着诸如缺失、异常等诸多问题。如果直接使用包含质量问题的交通数据展开交通应用。将会给应用结果带来不稳定和不安全隐患。

2.交通流数据的三要素（流量、速度和时间占有率）对于交通的演变规律有着重要影响，通过研究交通量的变化规律以及三个参数之间的关系，对未来的交通的发展趋势和预测起了重要作用，同时也是改善现有交通道路设施的关键前提。

3.交通需求预测在交通规划领域中非常重要,交通需求预测的四阶段法更是城市地区规划的基石，更是交通规划领域符合未来发展的条件，如何利用不同场景所给出的交通数据来进行研究预测和建立模型，这对于实测交通流数据的模型检验结果，总结出实用的数据质量控制和预测经验，为交通管理部门的数据质量改善的实践工作提供指导提供很大的帮助。

**1.2 问题概括**

**问题一：**对于给定数据(data 1)，需要建立识别问题数据（包括异常数据、重复记录等）的方法和流程，并完成对问题数据的修正及缺失数据的填补。

**问题二：**首先需要在时间维度上对交通流参数（流量、速度和时间占有率）进行建模，用以刻画交通系统在时间上的演化过程。其次，需要对这三个参数之间的关系进行建模。对于给定的数据(data 2),将缺失的数据(NA)补全,即用数值代替 NA。

**问题三：**要求对给定的数据（data 3）,建立模型并对于给定的数据(data 3),将缺失的数据(NA)补全,即用数值代替 NA。

2 问题分析

本题是基于交通流参数（占有率、流量、速度）来建立交通检测器数据质量控制和预测的数学模型，本题提出了采集到的数据会有数据缺失的情况,首先依据数据缺失的处理方法（数据删除、数据插补）对数据进行第一步的处理，对异常数据进行数据清洗后再进行阈值检验，提高数据集的质量。其次，根据时间维度上建模的方法，得到它们在时间尺度的演化过程，再根据三者之间的关系建立三者关系公式的模型，补全数据。然后根据之前的数据清洗与时间建模，关系建模，从而建立预测的数学模型来实现对交通需求的预测。

对于问题一，要对所给出的数据利用python进行数据清洗，将给定的data1进行遍历，检查数据完整性，通过前后数据补全，缺的数据多的话，考虑平滑处理，去除重复记录，对三参数进行方差检验，设置阈值检验，确保数据合法性，以此得到质量提高的数据集。

对于问题二，交通系统已经在时间周期上表现出依赖性等显著特征，将data2数据清洗，重构数据集，对三个参数分别在时间尺度上利用拟合进行建模，得出模型1，推导三参数的关系公式，建立三维坐标从而建立模型2。

对于问题三，在解决问题一和二的前提下，重复前面问题一、二步骤，之后得到三者基于时间的关系，通过曲线以及邻近线性插值可以预测未来时间三要素的值。

3模型假设

1.所有数据均为原始数据。

2.阈值

3.

近外红光谱的肠溶片包衣厚度在当前条件下不可测量，只能确定何时包衣厚度合适。

3.样品中的特征峰均来自于片心和包衣材料，不来源于其他物质。

4.包衣材料和样品均有相同的特征吸收。

5.近红外光谱在测量吸收峰时，吸收峰没有其他耗损。

6.素片就是样品的片心，而样品=片心+包衣材料，样品不含其它不相关物质。

4名词解释和符号说明

**4.1名词解释**

样本点：某一个时刻的各个近红外线所有波长对应的吸收值。

样品点：一个样品对应的所有时刻的各个近红外线所有波长对应的吸收值。

训练集：提取经过波长降序处理的原始数据集的每时刻前面10个样本共100个样本。

测试集：提取经过波长降序处理的原始数据集的每时刻剩下的5个样本共50个样本。

平均训练集：训练集的每一时刻的所有样品平均值(10个样本点)

符号说明

表 4.1 这是表

|  |  |
| --- | --- |
| 符号 | 意义 |
|  | 对原始数据近红外线波长降序处理和按时刻、素片、最优分组的数据集 |
|  | 标准化处理的训练集 |
|  | 标准化处理的测试集 |
|  | 标准化处理的平均训练集 |

5 模型建立与求解

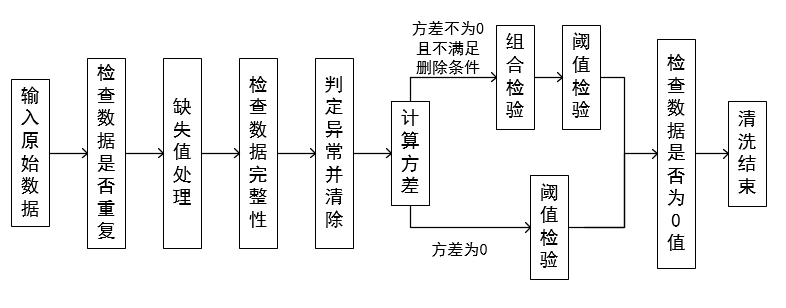
5.1 问题一的分析与求解

5.1.1 问题分析

题目要求的是对于给定数据(data 1)，需要建立识别问题数据（包括异常数据、重复记录等） 的方法和流程，并完成对问题数据的修正及缺失数据的填补，我们实际解决的是将数据遍历，检查数据的完整性，通过方差检验和阈值检验对异常数据进行清洗，方便后面的时间尺度上三要素模型的建立，同时也提高数据的质量，保证后续试验结果的正确性。

5.1.2 数据清洗流程

交通检测器识别问题数据的方法：

图1 交通检测器识别问题数据的流程

交通检测器数据清洗的步骤：

（1）将数据导入处理工具，原始数据遍历，查看是否重复，是就进行去重清洗。

（2）查看缺失值。确定缺失值范围，去除不需要字段，利用邻近线性插值填充缺失内容

（3）利用方差检验判断异常数据。分为为0和不为0两种情况处理，三参数方差为0的先组合检验，再阈值检验。另有一种则是直接阈值检验。

（4）将上述两种情况中不通过的数据直接删除。

（5）形成新的数据，数据清洗完成。

5.1.3 模型求解

正文

5.2 问题二的分析与求解

5.2.1 问题分析

题目要求的是先在时间维度上对交通流参数（流量、速度和时间占有率）进行建模，其次对这三个参数之间的关系进行建模，最后补全data2缺失的数据。而我们就按照这个顺序进行建模，从已知的数据出发，先进行数据清洗，再重构数据集，对三个参数分别在时间尺度上利用拟合进行建模，得出模型1，找出三参数的关系公式，建立三维坐标从而建立模型2。

5.2.2 模型1在时间维度上的交通流参数多元回归模型

1.模型分析

经过对问题二的分析，通过对各种模型的特点对比，这里使用多元回归模型进行曲线拟合。

2.模型建立

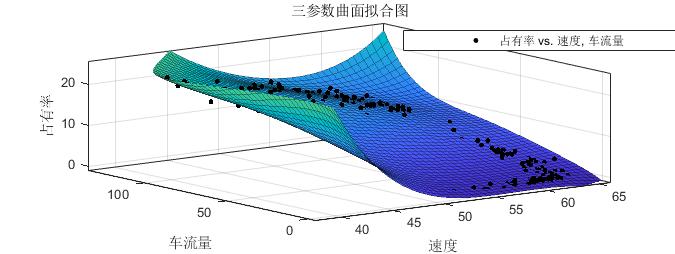
3.模型求解

正文

5.2.3 模型2

1.模型分析

由问题分析可知，我们的模型可以选用曲面拟合的形式进行，即采用三维坐标轴将三个参数量作为变量描绘出坐标系中的点的位置，通过曲面拟合的形式得出拟合曲面表达式，这样就可以得到三参数之间的关系表达，同时也可以通过三参数不同的取值得到预测。如下图所示：



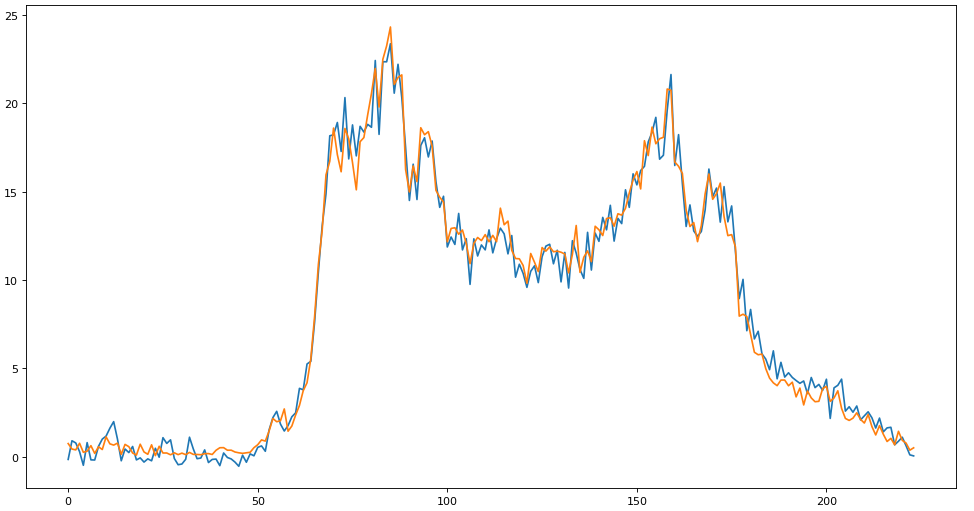
更好的，我们采用多元多项式进行拟合，同样的，由于需要寻求三参数之间的关系，我们讲占有率作为标签，将其余两个量速度和车流量作为变量，进行建模分析，采用Scikit-learn的liner-model和curve-fit进行变量的拟合，得到关系表达式，进而得到三者之间的关系。

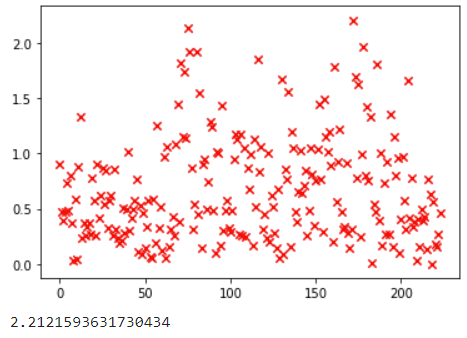
2.模型建立

1. 数据集重构
2. 初始化参数
3. 设置参数限制
4. 使用PolynomialFeatures进行特征构造
5. 特征转换与线性回归
6. 建立模型，开始训练
7. 预测值结果误差处理

3.模型求解

将预处理的数据带入上述模型中，提取需要用到的列作为变量，以dataframe格式存储，在运算时转为numpy的数组形式，通过python变成，得到对应阶数的特征构造多项式，接着我们对得到的特征进行转换，将转换后的特征数据作为输入值带入模型进行训练。将训练得到的结果再次进行特征转换得到我们期望的数据形式，然后与实际值进行误差对比。





5.3 问题三的分析与求解

正文

5.3.1 问题分析

正文

5.3.2 模型3

正文

5.3.3 模型求解

正文

6 误差分析

误差分析误差分析

7 模型评价与推广

7.1 模型优点

1.该模型深入研究了交通检测器数据的各项指标，充分利用数据，具有说服力。

2.该模型清晰明了地刻画了交通系统在时间上的演化过程，创新性地通过数据对交通流量进行了预测。

3.该模型利用了不同时间点检测出了交通流参数的数据的除了交通流参数在时间维度上的关系模型以及交通流参数之间的关系，十分巧妙。

7.2 模型缺点

交通流参数之间存在着联系，但不仅限于线性回归关系，从而会影响模型的正确性。

7.3 模型推广

本文所用的模型具有很强的通用性，具有多元限制条件的问题都可以类比使用，比如文中是交通检测数据参数之间的关系，在研究其他方面的数据参数之间的关系时也可套用本模型。不仅仅局限于交通领域，也可扩展到生活其他领域，适应我们的日常需求。

参考文献

引用别人的成果或其他公开的资料(包括网上甚至在“博客”上查到的资料) 必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。正文引用处用方括号标示参考文献的编号，如[1][3]等；引用书籍还必须指出页码。参考文献按正文中的引用次序列出， 其中书籍的表述方式为：

[编号] 作者，书名，出版地：出版社，起止页码，出版年。

参考文献中期刊杂志论文的表述方式为：

[编号] 作者，论文名，杂志名，卷期号：起止页码，出版年。

参考文献中网上资源的表述方式为：

[编号] 作者，资源标题，网址，访问时间（年月日）。

附录