《MATLAB 程序设计》大作业个人汇报

COVID-19 疫情数据的统计分析*

王逸扬（19300180016）
杨耕智（19300180112）

2022 年 5 月 21 日

*相关代码及其版本演进托管在https://github.com/maix00/StatisticsAnalysis.

目录

1 整体分析思路

2 数据的导入

导入数据已有可以使用的 readtable 函数，且可以通过 detectImportOptions 函数，在导入数据前，预先探测并修改导入参数. 在此过程中，我们发现如下问题：

(1) 以 daily_info.csv 数据集的 new_vaccinations 变量为例. 由于疫情发生初期长时间没有此类数据，detectImportOptions 函数探测认为，此变量的变量类型是为 char，因此需要手动修改为 double.

(2) 在导入 daily_info.csv 数据集时，通常最后需要检索出某个国家、某个时间段的数据. 为此，MATLAB 没有专门的检索函数可以使用.

(3) 当数据集特别大，而我们只需要个别列、以及检索出的个别行的数据时，需要手动设置导入选项以节省存储空间.

为了解决上面发现的问题，并寄希望于能在单一函数中解决全部的导入问题，我们利用了 MATLAB 的面向对象编程. 下面用一个例子来阐述这个类的作用. 比如，我们想要导入法国在 2020 年的全体新增接种疫苗数与总接种疫苗数的数据——

```
path_daily = './data/COVID19/daily_info.csv';
data = StatisticsAnalysis( ...
    'TablePath', path_daily, ...
    'ImportOptions', { ...
        'VariableTypes', { ...
            'new_vaccinations', 'double';
            'total_vaccinations', 'double' ...
            }; ...
```

```
        'SelectedVariableNames', ...
            {'date', 'new_vaccinations', 'total_vaccinations'} ...
        }, ...
    'SelectTableOptions', { ...
        'location', 'France'; ... Country
        'date', timerange("2020-01-01", "2020-12-31", 'closed') ...
    } ... % Select Table Before Importing, Please Set 'SelectFirst' true
    ).Table;
```

其中，StatisticsAnalysis 是该类的生成函数，"TablePath"、"ImportOptions" 和"SelectTableOptions" 是函数参数名，.Table 是该类的一个属性. 之所以取 StatisticsAnalysis 这个名字，是因为后面将赋予它更多的功能. 更多关于这个类的信息，请参见附录A.

简单的数据可视化

3 数据的缺失值处理

以 daily_info.csv 数据集为例，我们发现如下问题:

(1) 如果要分析变量 new_vaccinations 或 total_vaccinations，那么在疫情初期，有很多缺失值，可能需要填充为 0.

(2) 变量 new_cases 可能会出现意外的缺失值，但能从 total_cases 变量中恢复.

(3) 一些国家的数据，可能由于更正数据的需要，部分日期的 total_cases 会比上一日的少，以致当日的 new_cases 成为缺失值.

(4) 一些不能恢复的变量，可能需要线性插值，或者其他的方法连接.

(5) 在统计上不可避免出现一些随机波动，可能需要光滑处理.

为了解决这些问题，我们编写了一个函数 tableMissingValuesHelper.m，可以用来分析缺失值的分布，在这之后再经过人工判断，可以选择不同的参数，以恢复数据. 更多关于这个函数的信息，请参见附录C.

4 数据分析过程

4.1 一般的统计指标

4.2 聚类分析

4.3 疫情数据的分析预测

A StatisticsAnalysis 类

本类的主体与相关方法存储在./functions/@StatisticsAnalysis. 本类目前只有两个功能，一是数据的导入与检索，二是对表格进行初步的统计分析，计算传入的或预设的统计指标.

StatisticsAnalysis 生成函数，主要的参数包括 ImportOptions，SelectTableOptions 和 TagsGenerateOptions. 其实现的伪代码见下.

```
TablePath作为参数被传入时——
Step 1: detectImportOptions
Step 2: 导入与检索表格
    Step (1): 将需要检索的变量，作为需要导入的列，更新到ImportOptions
    Step (2): 将ImportOptions作为参数，导入表格
    Step (3): 引用./function/selecttable.m函数检索表格
    Step (4): 去除多余的列（前面辅助检索的列）
Step 3: 添加变量标签Tag
        并将标签对应的统计指标添加到Table.Properties.CustomProperties
    Step (1): 引用类方法TagsGenerate以生成标签、计算统计指标
    Step (2): 引用addprop函数，将统计指标添加到..CustomProperties
```

```
Table作为参数被传入时——
Step 1: 检索表格
    if SelectTableOptions非空
        引用./function/selecttable.m函数检索表格
    end
Step 2: 添加变量标签Tag
        并将标签对应的统计指标添加到Table.Properties.CustomProperties
    Step (1): 引用类方法TagsGenerate以生成标签、计算统计指标
    Step (2): 引用addprop函数，将统计指标添加到..CustomProperties
```

各种 Options 参数，要求是左侧为选项名、右侧是选项值的 cell. 可以参见节2中的例子.

TagsGenerateOptions 所接受的选项名，或 TagsGenerate 方法所接受的传入参数名，主要包括 CustomTagName 和 CustomTagFunction等. 预设的变量标签 DefaultTagNames 包括：unique, invariant, logical, categorical, discrete, continuous. 可以参见4.1中的例子.

B selecttable 函数

本函数存储在./functions.

C tableMissingValuesHelper 函数

本函数存储在./functions.