《MATLAB 程序设计》大作业个人汇报

COVID-19 疫情数据的简要统计分析与预测*

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 整体分析思路与代码分工

众所周知，2019 年底开始的 COVID-19 疫情给人们的生产生活造成了极大的影响，如何利用疫情相关数据、如何通过数据挖掘、分析、预测，以指导人们作出更加有利于社会发展的决策，是当今较为热门的研究方向. 在此，我们将根据所给的有关数据，尽可能地贴合这个目的，做出一些简要的分析与预测.

1.1 数据观察与简要分析

首先，观察所给数据. 有两个数据集，一个是 `country.csv`（以下称「国家数据」），另一个是 `daily_info.csv`（以下称「每日数据」）. 每日数据中包括 21 个国家的有关数据，其中 Senegal 在国家数据中有数据缺失，遂只对余下 20 个国家进行分析，分别是 Australia, Canada, Chile, China, Egypt, Ethiopia, France, Germany, Hungary, Iceland, Japan, Malaysia, Saudi Arabia, South Africa, South Korea, Sweden, Thailand, United States, 以及 Zimbabwe.

这 20 个国家的国家数据中，每个国家都有 12 个特征，简要分类：

地理位置: 所属大洲 `continent`;

人口成分: 人口 `population`, 每平方千米人口密度 `population_density`, 中位年龄 `median_age`, 65 岁以上人口占比 `aged_65_older`, 70 岁以上人口占比 `aged_70_older`;

经济环境: 按购买力平价的人均国民生产总值 `gdp_per_capita`;

卫生水平: 男性吸烟率 `male_smokers`, 女性吸烟率 `female_smokers`, 每年每十万人心血管疾病死亡率 `cardiovasc_death_rate`, 每千人床位数 `hospital_beds_per_thousands`, 预期寿命 `life_expectancy`.

据此，我们可以对这些国家进行无监督聚类分析，在每一类中选取一个代表性的国家构建一分析预测模型（利用每日数据和/或国家特征），并对同一类中的其他国家验证该模型的有效性. 但是，需要指出，因为所给数据的有限性，所建立或选择的模型是有限的，这里的特征作为模型参数的个数也是有限的，因此这里的多数特征在本文中只能用于聚类.

对于每日数据，我们首先指出，在数据的导入与缺失值处理上，这个数据集遇到了一些困难，对于这些问题，我们将在节2与节3中分别尝试解决. 下面，对每日数据的各字段（除所属大洲与时间戳外），简要分类：

严重程度: 新增与累计病例：new_cases, total_cases; 住院与重症监护 ICU 人数：hosp_patients, icu_patients; 近七日平均检测阳性率：positive_rate;

检测能力: 新增与累计检测人次：new_tests, total_tests; 近七日平均检测阳性率：positive_rate;

疫苗接种: 新增与累计疫苗接种人次：new_vaccinations, total_vaccinations;

管控力度: 政府管控力度：stringency_index.

这些时间序列数据，(1) 可以用来构建一分析预测模型，(2) 也可以取某一时间段内的最大、平均等统计指标（时间序列特征），用来对不同国家进行监督或无监督聚类分析. 这里的聚类分析，得到的是不同国家的不同疫情发展模式，与前面的利用国家数据得到的关于国家特性的聚类不同.

另外，需要指出，因为所给数据的有限性，所能建立或选择的分析预测模型是极其有限的，比如，由于缺少治愈、死亡数据，不能采用传统的传染病模型 SIR 至 SEIR 等，又比如，由于缺少流调数据，无法统计计算基本再生数 R_0、潜伏期平均长度 $\bar{T_L}$ 等与传染病有关的统计指标. 此外，正因为所能建立或选择的模型是有限的，每日数据中的一些字段在本文中可能不会被用到.

1.2 整体分析思路

综合上面的简要分析，我们将本文的整体分析思路概括如下.

第一种路径 根据国家特征进行聚类，而后利用代表国家的每日数据构造一分析预测模型.

第一步:（聚类分析）对 20 个国家的特征数据，先进行主成分分析，选取前 5 个主特征后进行 5-mean 聚类分析.

第二步:（构建模型）基于 LSTM 神经网络模型，对代表国家一定时期内的每日新增病例数据进行学习.

第三步:（预测检验）利用上一步得到的模型，对时期外的数据进行预测检验，并对同一类中的其他国家的数据进行预测检验.

此路径能得到一种基于 LSTM 神经网络的分析预测模型.

第二种路径 根据每日数据和/或国家数据构造时间序列特征，而后进行谱系聚类.

第一步:（构造特征）

第二步:（聚类分析）

此路径能对不同国家的疫情发展模式进行分类，针对分类结果可以提出更有针对性的抑制疫情发展的建议. https://www.it610.com/article/1515470890224123904.htm.

第三种路径 拟合部分国家疫情发展初期缺失的数据.
https://www.it610.com/article/1515470890224123904.htm.

1.3 代码分工

因为疫情的关系，我们两位同学，不得不线上联系，我们也因此第一次尝试使用 github 平台. 仓库网址为https://github.com/maix00/StatisticsAnalysis. 有关代码分工的详细情况，在平台上可以清晰看到，这里就作简单介绍.

节2与节3数据导入、缺失值处理的有关代码主要是由王逸扬同学完成的，节4数据分析的有关代码主要是由杨耕智同学完成的.

2 数据的导入与检索

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) 以每日数据的 `new_vaccinations` 与 `total_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`，因此需要手动将导入参数修改为 `double`.

(2) 在导入每日数据时，通常最后需要检索出某个国家、某个时间段的数据，目的是形成时序数据. 但是，为此，MATLAB 可以通过索引检索，但是不同类型的数据检索的方式不同，不是很便利，没有专门的检索函数可以使用.

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与检索问题，我们利用了 MATLAB 的面向对象编程，编写了一个类 `StatisticsAnalysis`（之所以取这个名字，是因为之后还将赋予它更多的功能）. 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 年的新增与累计接种疫苗人次——

```
daily_SA = StatisticsAnalysis( ...
    'TablePath', './data/COVID19/daily_info.csv', ...
    'ImportOptions', { ...
        'VariableTypes', { ...
            'new_vaccinations', 'double',
            'total_vaccinations', 'double' ...
            }, ...
        'SelectedVariableNames', ...
            {'date', 'new_vaccinations', 'total_vaccinations'} ...
        }, ...
    'SelectTableOptions', { ...
        'location', 'France', ...
        'date', timerange("2020-01-01", "2020-12-31", 'closed') ...
        } ...
    );
daily = daily_SA.TimeTable;
```

其中，`"TablePath"`、`"ImportOptions"` 和`"SelectTableOptions"` 是函数参数名，`TimeTable` 是该类的一个属性. 如果我们还想导入法国在 2020 年的新增与累计病例数，我们只需——

```
daily2 = daily_SA.Update('ImportOptions', { ...
        'SelectedVariableNames', {'date', 'new_cases', 'total_cases'} ...
        }...
    ).TimeTable;
```

上面的方法不会重复导入、检索数据. 更多关于这个类的信息，请参见附录A.

3 数据的缺失值处理

以每日数据为例，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

本类存储在./functions/@StatisticsAnalysis. 本类目前有两个功能，一是数据的导入与检索，二是对表格进行初步的统计分析，即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. 其实现的主要过程见下.

```
TablePath作为参数被传入时——
Step 1: detectImportOptions
Step 2: 导入与检索表格
    Step (1): 用ImportOptions，修改detectImportOptions，并导入表格
    Step (2): 引用./function/selecttable.m函数检索表格
    Step (3): 去除多余的列
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
```

其中，各种 Options 参数，可以是 struct 类型，也可以是先选项名、后选项值的 cell 类型. 可以参见节2中的例子.

ImportOptions 所接受的选项名与选项值，请参考 detectImportOptions 中列明的属性名与属性值.

SelectTableOptions 所接受的选项名，须是表格的列名，所接受的选项值，可以是单一的值，可以是 timerange 类型，也可以是./functions/@arange.

前两种 Options 中，多个选项值可以 array 或 1x1 cell 的形式传入，不同的选项值将分别用于表格的导入、检索，最后尽可能地上下拼合表格.

TagsGenerateOptions 所接受的选项名，主要的即 CustomTagName 和 CustomTagFunction，前者给表格的列赋予标签、后者给不同的标签赋予统计指标（函数句柄）. 预设的变量标签、预设的统计指标，以及参数的格式，可以用 help 查看.

类方法中比较实用的有 TagsGenerate 与 Update,前者用作统计分析,后者用来更新各种 Options 参数.

B selecttable 函数

本函数存储在./functions.

C tableMissingValuesHelper 函数

本函数存储在./functions.

D 利用有关数据可以进行的可视化