

# CHIC402/CHIC602 Coursework 2 - Variant Analysis

36112985

This project focuses on the analysis of the United Kingdom's genetically-typed SARS-CoV-2 infections data. The data set summarises the weekly infections fractions of 17 SARS-CoV-2 variants, starting *2021-01-31*, and ending *2021-10-17*; additionally, there's the field *Other*, each value of this field represents a pooled fraction w.r.t. (with respect to) an unknown number of unnamed variants.

Each week has 2 fractions records: one for cases associated with known travel links, the other without. The sum of each record's 18 infections fractions is **1**.

## 1 The Data

In this document the data frame variable *variants* stores the variants' data set. It consists of

```
'data.frame': 76 obs. of 20 variables:
 $ week      : Date, format: "2021-01-31" "2021-01-31" ...
 $ travel    : Factor w/ 2 levels "none","travel": 1 2 1 2 1 2 1 2 ...
 $ Alpha     : num  0.95 0.615 0.964 0.474 ...
 $ Beta      : num  0.00128 0.07692 0.00138 0.19298 ...
 $ Delta     : num  0.000107 0 0.000231 0 ...
 $ Eta       : num  0.0016 0.07692 0.00154 0.08772 ...
 $ Gamma     : num  0.00 0.00 7.69e-05 0.00 ...
 $ Kappa     : num  0 0 0 0 ...
 $ Theta     : num  0 0 0 0 0 0 0 0 ...
 $ Zeta      : num  0.00032 0 0 0 ...
 $ Lambda    : num  0 0 0 0 ...
 $ VUI.21FEB.01: num  0.000213 0 0.000231 0 ...
 $ VOC.21FEB.02: num  0.00032 0 0.000385 0 ...
 $ VUI.21FEB.04: num  0 0.038462 0.000231 0.035088 ...
 $ VUI.21APR.03: num  0 0 0 0 0 0 0 0 ...
 $ VUI.21MAY.01: num  0 0 0 0 0 0 0 0 ...
 $ VUI.21MAY.02: num  0 0 0 0 ...
 $ VUI.21OCT.01: num  0.000107 0 0 0 ...
 $ Mu        : num  0 0 0 0 0 0 0 0 ...
 $ Other     : num  0.0458 0.1923 0.0322 0.2105 ...
```

### 1.1 Long Format

Some parts of this document use the long form of *variants*, stored by the variable *melted*.

```
# The long form of data frame 'variants'
melted <- variants %>%
  tidyr::gather(key = 'variant', value = 'fraction', -c(week, travel))
```

The structure of *melted* is

```
'data.frame': 1368 obs. of 4 variables:
 $ week      : Date, format: "2021-01-31" "2021-01-31" ...
 $ travel    : Factor w/ 2 levels "none","travel": 1 2 1 2 1 2 1 2 ...
 $ variant   : chr  "Alpha" "Alpha" "Alpha" ...
 $ fraction  : num  0.95 0.615 0.964 0.474 ...
```

## 1.2 Infections Proportion Statistics by Variant

Parts of this document depend on infections proportion statistics by variant. The function *VariantsProportions()* calculates the statistics. In the code snippet below *proportions* stores the calculations.

```
# Infections proportion statistics
proportions <- VariantsProportions(melted = melted)
```

the variable *proportions* is a named list. The elements of *proportions* are discussed in detail at appropriate points in this document. In brief, *proportions* consists of four data frames:

```
# The contents of the named-list variable 'proportions'; each is a data frame.
names(proportions)
```

```
[1] "none"          "travel"        "leading"       "leadingseries"
```

The data frame *none* summarises the total contribution of each variant w.r.t. **(a)** infections without known travel links, and **(b)** the data's time span (*2021-01-31* → *2021-10-17*); the same logic applies to *travel*, for infections with travel links.

The data frame *leading* lists the 5 variants responsible for the most infections, w.r.t. both travel & non-travel linked infections, during the data's time span; *leadingseries* stores the infections' data of these five.

## 2 An Exploration of Non-travel Data

Each variant’s weekly data consists of two records. One represents the infections-fractions associated with known travel links (denoted *travel* in field *travel*), whilst the other represents the infections-fractions without known travel links (denoted *none* in field *travel*). This section focuses on non-travel data, and the variable *local* stores the applicable data.

```
# 3.0
# The non-travel data only
local <- variants %>%
  dplyr::filter(travel == 'none')
```

The truncated data description highlights the number of non-travel data records.

```
'data.frame': 38 obs. of 20 variables:
 $ week      : Date, format: "2021-01-31" "2021-02-07" ...
 $ travel     : Factor w/ 2 levels "none","travel": 1 1 1 1 1 1 1 1 ...
 $ Alpha      : num 0.95 0.964 0.974 0.982 ...
 $ Beta       : num 0.001278 0.001385 0.000663 0.001003 ...
 $ Delta      : num 1.07e-04 2.31e-04 2.21e-04 7.72e-05 ...
 [list output truncated]
```

### 2.1 The Progression of Delta (non-travel)

A variant of interest, amongst the variants that contribute to the non-travel infections quantities, is the **Delta** variant. During the data’s time span the largest proportion of non-travel infections was due to the Delta variant (*Table 1*).

Table 1: The variants responsible for the 5 largest non-travel infections proportion

| variant      | proportion |
|--------------|------------|
| Delta        | 22.333     |
| Alpha        | 14.553     |
| VUI.21OCT.01 | 0.707      |
| Beta         | 0.050      |
| Kappa        | 0.030      |
| VUI.21MAY.01 | 0.025      |

The graph of *Fig. 1* illustrates the progression of Delta variant infections over time. Each week’s Delta fraction is a fraction of the week’s total non-travel infections due to all recorded variants; named and otherwise. Delta’s fractions peak at 0.991 on 2021-07-04; hereafter, it seems that the Delta fraction values are continuously, but slowly, declining.

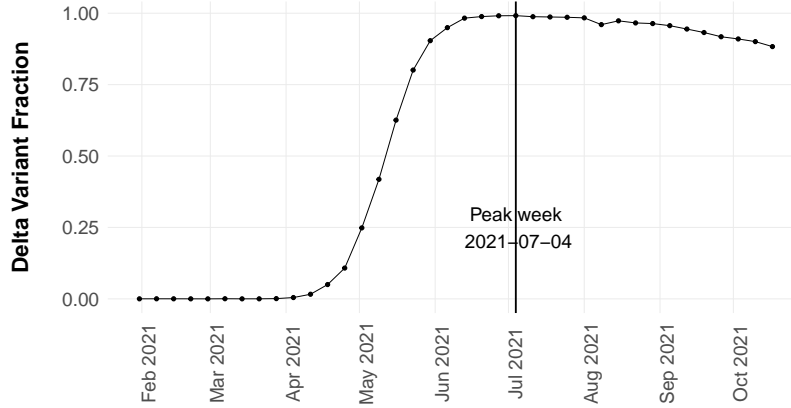


Figure 1: The Weekly Non-Travel Delta Variant Infections Fractions Over Time

## 2.2 Variant Peaks (non-travel)

In general, each variant has a fraction peak date w.r.t. the

- data's time span: *2021-01-31*  $\rightarrow$  *2021-10-17*
- non-travel data

The snippet below determines the peak date of each variant, and arranges the variants in descending peak-date fraction order.

```
# The peak date of each variant
peaks <- local %>%
  select(!(travel)) %>%
  tidyr::gather(key = 'variant', value = 'p', -week) %>%
  group_by(variant) %>%
  slice(which.max(p)) %>%
  arrange(desc(p))
```

Hence, *Table 2* lists the peak date & peak-date fraction per variant.<sup>1</sup>

Table 2: The peak date and peak-date fractions; descending fraction order

| Variant      | Peak Date  | Peak Date Fraction |
|--------------|------------|--------------------|
| Delta        | 2021-07-04 | 0.991              |
| Alpha        | 2021-03-28 | 0.989              |
| VUI.21OCT.01 | 2021-10-17 | 0.116              |
| Other        | 2021-01-31 | 0.046              |
| Kappa        | 2021-04-18 | 0.009              |

<sup>1</sup>Note: **Other** is not a variant, it represents a collection of variants, therefore excluded.

| Variant      | Peak Date  | Peak Date Fraction |
|--------------|------------|--------------------|
| Beta         | 2021-04-25 | 0.008              |
| VUI.21MAY.01 | 2021-05-09 | 0.007              |
| VUI.21FEB.04 | 2021-04-18 | 0.003              |
| Gamma        | 2021-04-25 | 0.003              |
| Eta          | 2021-05-02 | 0.003              |
| VUI.21MAY.02 | 2021-04-11 | 0.002              |
| Zeta         | 2021-02-14 | 0.001              |
| VUI.21APR.03 | 2021-04-18 | 0.001              |
| VOC.21FEB.02 | 2021-02-07 | 0.000              |
| Mu           | 2021-07-11 | 0.000              |
| VUI.21FEB.01 | 2021-02-07 | 0.000              |
| Lambda       | 2021-01-31 | 0.000              |
| Theta        | 2021-01-31 | 0.000              |

### 3 Variant Proportions by Known/Unknown Travel Links

The section *Infections Proportion Statistics by Variant* (Chapter 1: The Data) calculates a few variant proportions, and stores the results in named-list variable *proportions*.

```
# The contents of the named-list variable 'proportions'; each is a data frame.
names(proportions)
```

```
[1] "none"          "travel"        "leading"       "leadingseries"
```

The *proportions* element *leading* lists the five variants responsible for the five largest variant-infection-proportions, w.r.t. **(a)** known & unknown travel links, and **(b)** the data's time span; *Table 3*.

Table 3: The variants responsible for the five largest variant-infection-proportions w.r.t. entire data time span

| Variant      | Proportion |
|--------------|------------|
| Delta        | 44.1426727 |
| Alpha        | 25.5222505 |
| VUI.21OCT.01 | 1.1626944  |
| Beta         | 1.0029552  |
| Eta          | 0.6273858  |

#### 3.1 The Predominant Variants

The time series of each variant of *Table 3* are illustrated in *Fig. 2* & *Fig. 3*; *Fig. 3* has a rescaled *y-axis*, which makes it easier to observe the progression of curves that have small fraction values. A few observations:

- In relation to infections **with known travel links**, the Alpha variant dominated initially, but the Beta & Eta fractions are fairly substantive. During May 2021 the Delta variant overtook the Alpha variant - from May 2021, until the end of the data's time span, the Delta variant had the largest infections-fraction each week. By the end of the period (a) the Alpha variant barely contributes to each week's travel linked infections, and (b) the fractional contributions of VUI.21OCT.01 are increasing each week.
- In relation to infections **without known travel links**, again the Alpha variant dominated initially, and quite comprehensively; the Beta & Eta variants contribute small fractions. However, from May 2021 onward the Delta variant had the largest infections-fraction each week. In line with

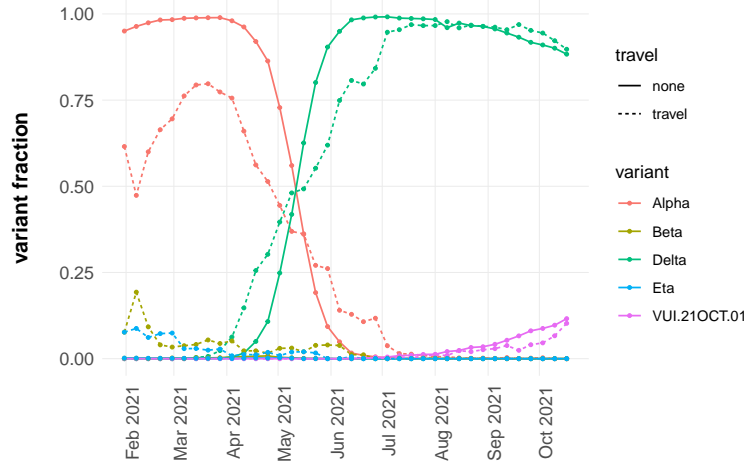


Figure 2: Weekly fraction proportions of the predominant variants listed in Table 3; w.r.t. (a) travel and non-travel infections, and (b) the data's time span. For a better view of the curves with generally small fractions, refer to Fig. 3; it is a version of this graph with a re-scaled y-axis.

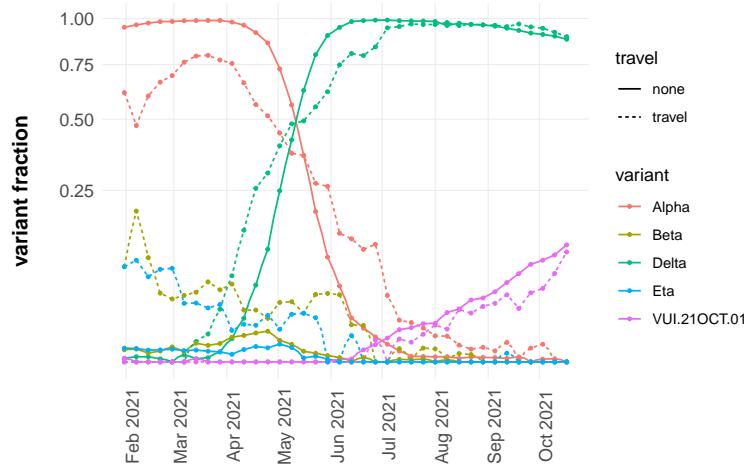


Figure 3: A version of Fig. 2 wherein the y-axis has been re-scaled, this makes it easier to observe the progression of curves that have generally small fraction values.

## 3.2 Progression Patterns

During the data set's time span Alpha initially dominated travel-linked & non-travel-linked infections. Its dominance is much more pronounced in the non-travel graph.

By May 2021 Delta starts contributing the largest weekly travel & non-travel infection fractions. The total travel & non-travel infection proportions due to Delta are quite similar. The fractions contributed by Delta continuously increase until the weekly fractions of VUI.21OCT.01 start becoming prominent. Thus far, the weekly travel-linked & non-travel-linked infection fractions of VUI.21OCT.01 have been relatively low.

The infections due to Beta, and those of Other (12 variants + Other), are predominantly travel linked.

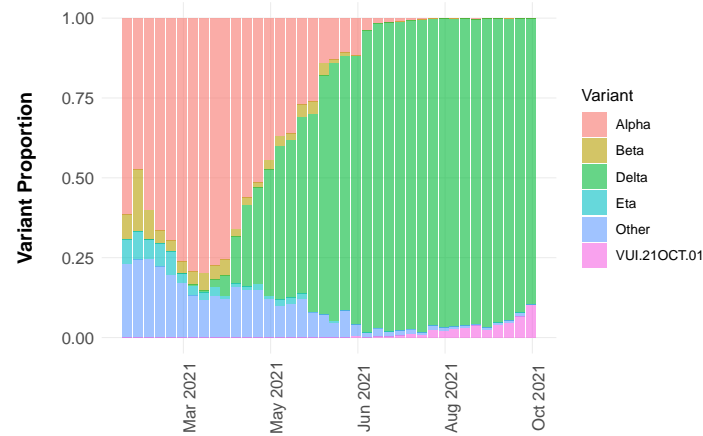


Figure 4: Bars of weekly travel-linked infection fractions over time; for each of the 5 predominant variants and Other (12 variants + Other).

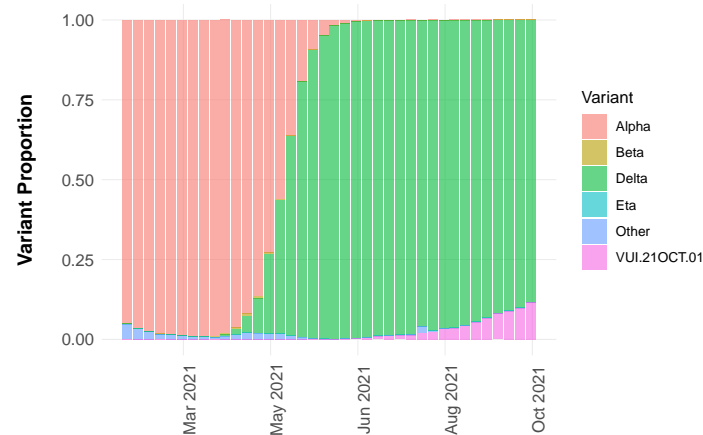


Figure 5: Bars of weekly non-travel-linked infection fractions over time; for each of the 5 predominant variants and Other (12 variants + Other).

## 4 The New Variant

This chapter focuses on non-travel VUI.21OCT.01 variant data; the variable `vui21oct01` stores this data.

```
# the non-travel VUI.21OCT.01 Data
vui21oct01 <- variants %>%
  filter(travel == 'none') %>%
  select(week, 'VUI.21OCT.01')
```

### 4.1 Non-travel VUI.21OCT.01 Cases Over Time

The graph of *Fig. 6* illustrates the non-travel VUI.21OCT.01 variant infection fractions over time.

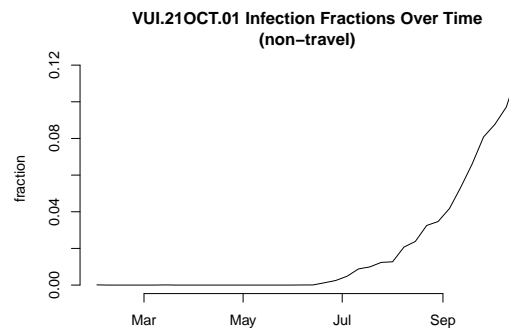


Figure 6: The non-travel VUI.21OCT.01 variant infection fractions over time.

### 4.2 A Prediction Model

An exponential growth model, for predicting daily infection fractions, has been developed. The model is

$$p = e^{(d\beta + \alpha)}$$

wherein

- $\alpha = -526.0385$
- $\beta = 0.0277$
- $d \rightarrow$  days since 1970-01-01

The function `VariantProgressionModel()` uses this model to predict daily infection fractions w.r.t. (a) a start date, and (b) a length of time. The variable `estimates` stores the predictions w.r.t. an entire year, starting from the starting date of the non-travel VUI.21OCT.01 data.

```
# predicting variant fractions
starting <- min(vui21oct01$week)
period <- lubridate::years(x = 1)
estimates <- VariantProgressionModel(starting = starting, period = period)
```



The graph of *Fig. 7* illustrates the model's predictions.

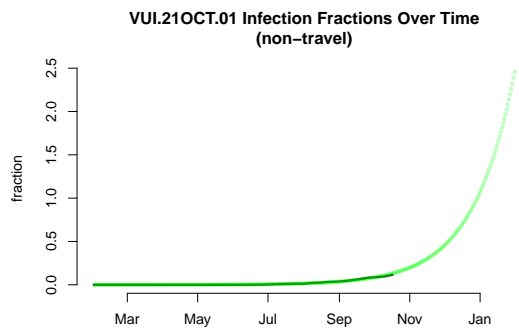


Figure 7: The original non-travel VUI.21OCT.01 variant infection fractions over time, and predictions. During January 2022 the predictions become implausible; a prediction should not exceed 1 because it is a fraction, and it can only equal 1 if, and only if, no other variant exists.

The predictions for *1 September 2021* & *1 January 2022* are extracted from *estimates*; *Table 4* summarises the results. The 1 January 2022 prediction is implausible because it is  $> 1$ ; herein, a prediction should not exceed 1 because it is a fraction, and it can only equal 1 if, and only if, no other variant exists.

Table 4: Predictions for 1 September 2021 & 1 January 2022

|     | date       | prediction |
|-----|------------|------------|
| 214 | 2021-09-01 | 0.036      |
| 336 | 2022-01-01 | 1.070      |