Data 2050 Final Presentation

Zhirui Li

Introduction

- Focus on reproducible research in public health
- Replicate and extend a paper that involves public health using publicly available data
- Identify a topic and several research papers related to that topic
- Reproduce the model using publicly available datasets
- Go beyond the scope of the original paper, conducting more visualizations and analyses

Paper of Interest

- A forecast model for prevention of foodborne outbreaks of non-typhoidal salmonellosis
- Multivariate regression time series + Generalized auto-regressive and moving average models (GARIMA)
- Validated by analyzing the cases of Salmonella Enteritidis in Sydney Australia

GARIMA:

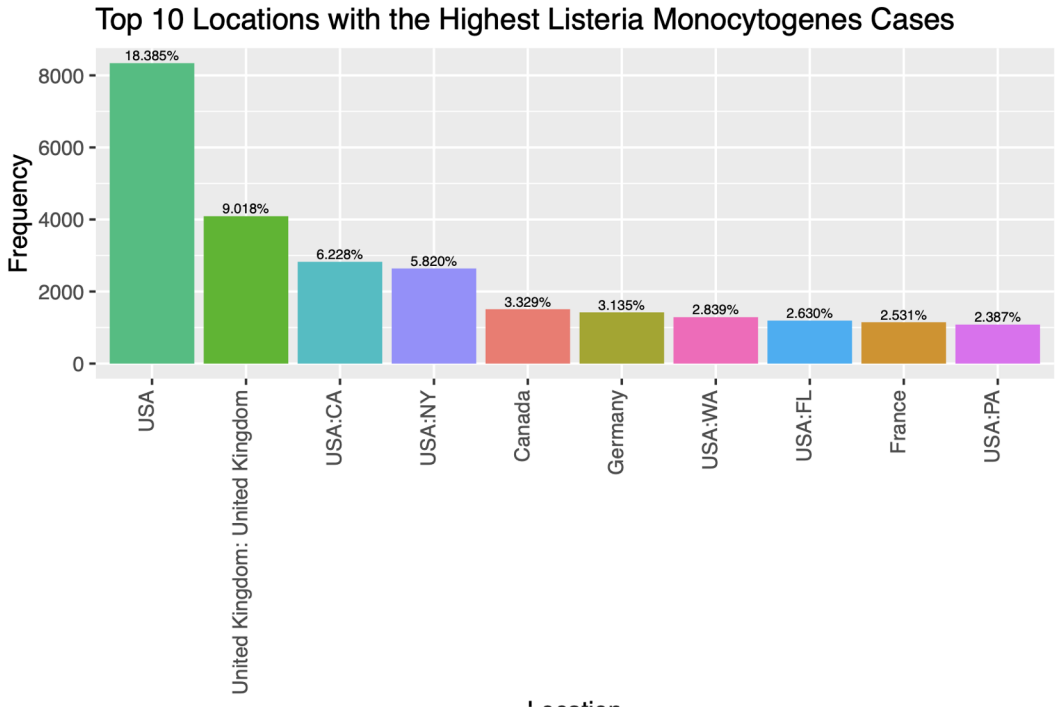
- G: generalization
- AR: autoregression
- I: integration
- MA: moving average

Listeria Monocytogenes Dataset

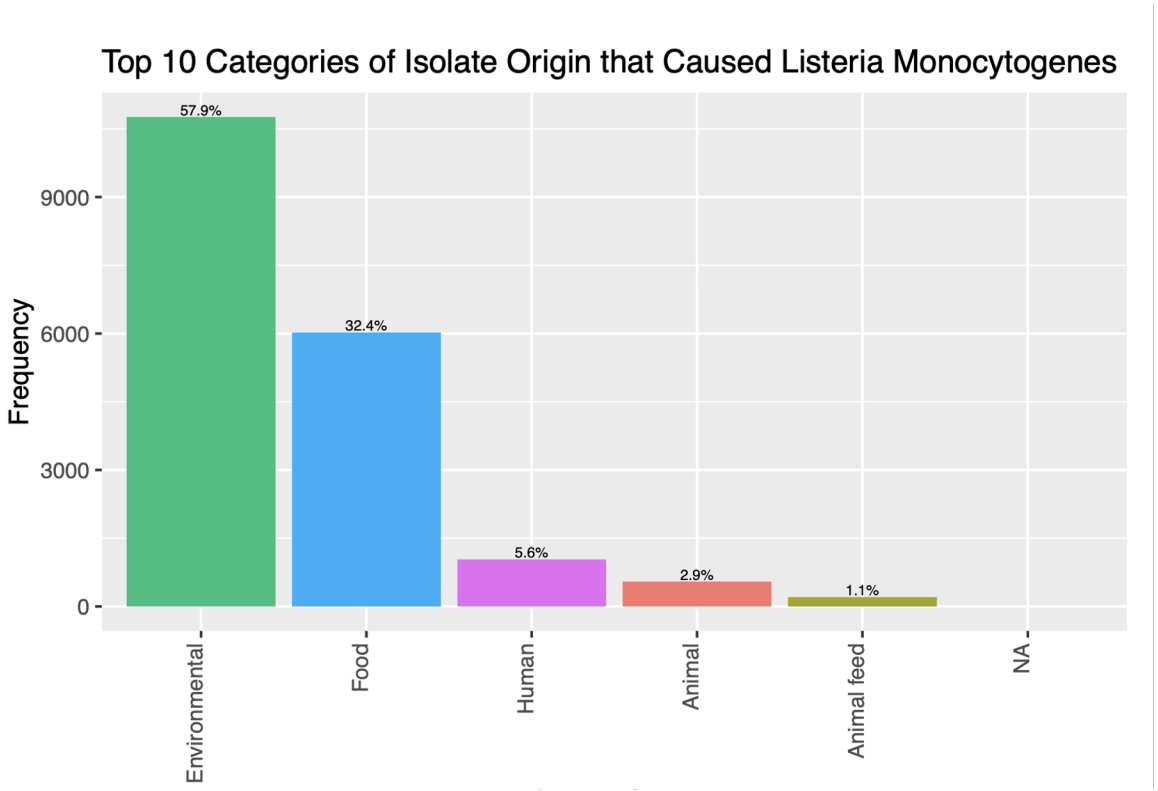
- National Center for Biotechnology Information
- Includes 51738 observations and 50 variables

```
 [1] "Library_layout"                "Method"                       "WGS_accession"
 [4] "WGS_prefix"                    "Host_disease"                 "TaxID"
 [7] "PFGE_secondary_enzyme_pattern" "PFGE_primary_enzyme_pattern"  "Level"
[10] "Species_TaxID"                 "Outbreak"                     "SRA_release_date"
[13] "SRA_Center"                    "Run"                          "Platform"
[16] "AMR_genotypes_core"            "Contigs"                      "N50"
[19] "Length"                        "BioProject"                   "Collection_date"
[22] "Stress_genotypes"              "Lat/Lon"                      "Collected_by"
[25] "PD_Ref_Gene_Catalog_version"   "AMRFinderPlus_analysis_type"  "AMRFinderPlus_version"
[28] "Scientific_name"               "Virulence_genotypes"          "AST_phenotypes"
[31] "K-mer_group"                   "Organism_group"               "Host"
[34] "Source_type"                   "IFSAC_category"               "Strain"
[37] "Isolate_identifiers"           "Serovar"                      "Isolate"
[40] "Create_date"                   "Location"                     "Isolation_source"
[43] "Isolation_type"                "SNP_cluster"                  "Min-same"
[46] "Min-diff"                      "BioSample"                    "Assembly"
[49] "AMR_genotypes"                 "Computed_types"
```

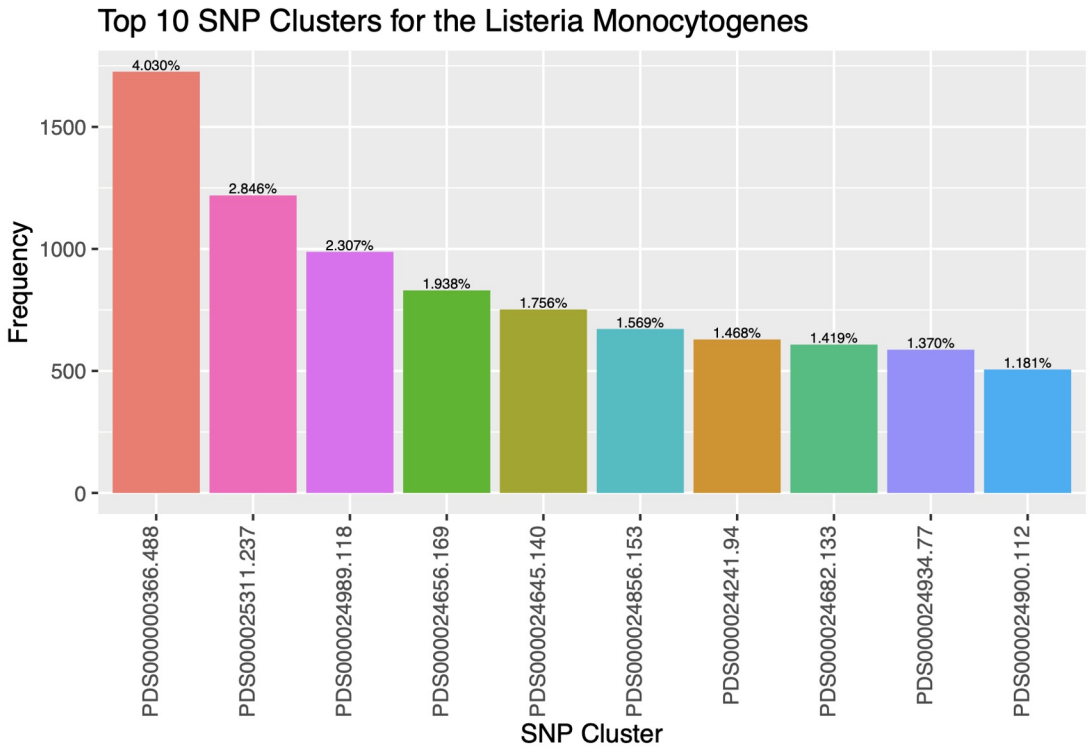
Variable “Location”



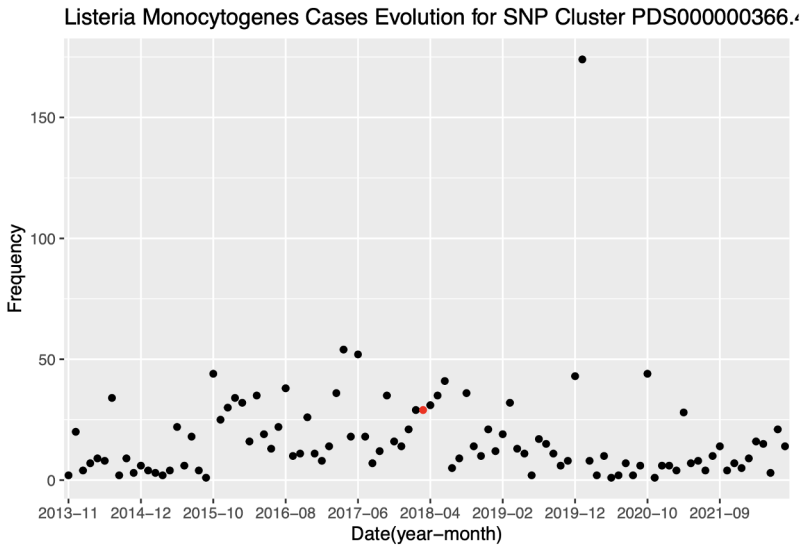
Variable “Source_type”

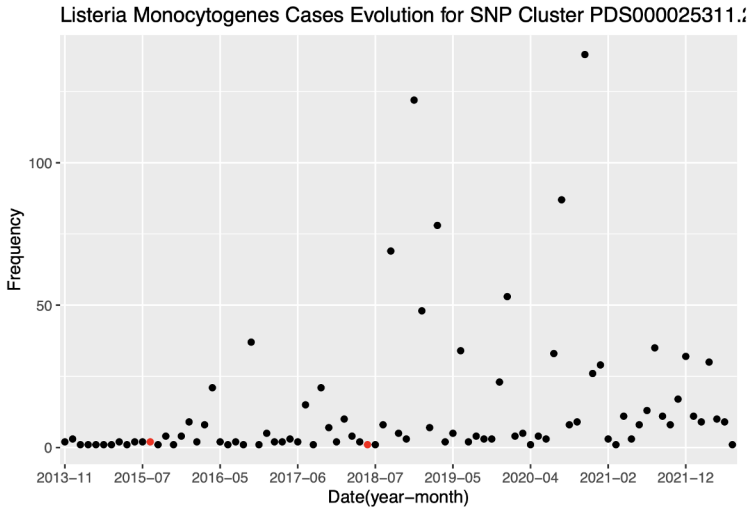


Variable “SNP_cluster”

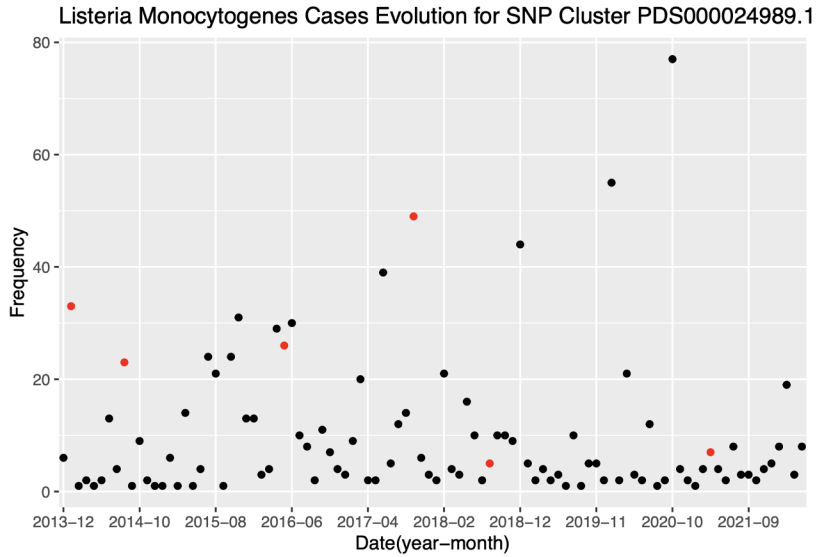


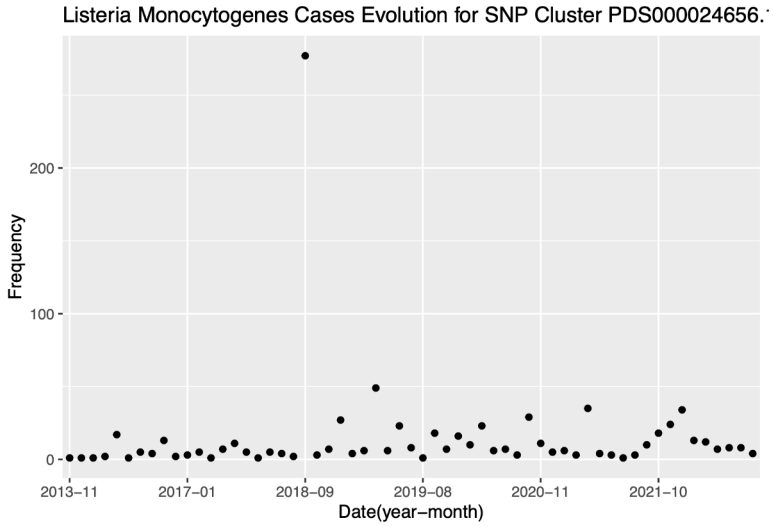
SNP Cluster Visualization by Month



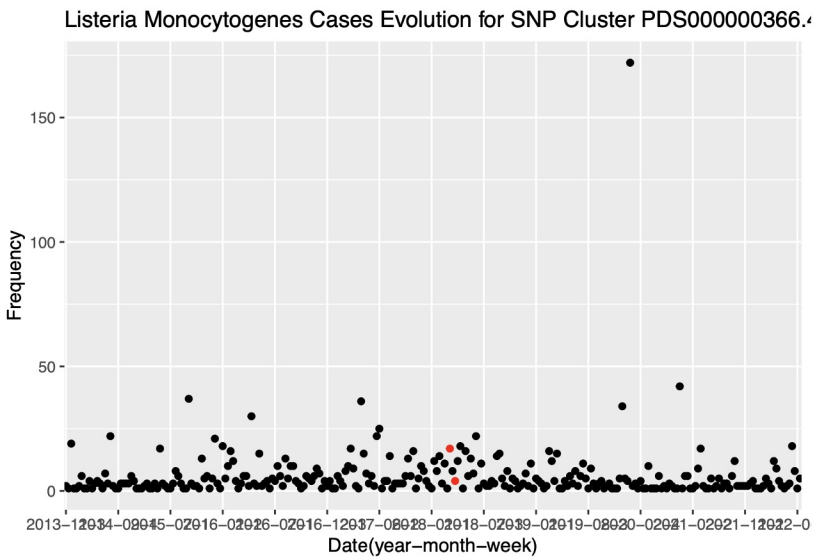


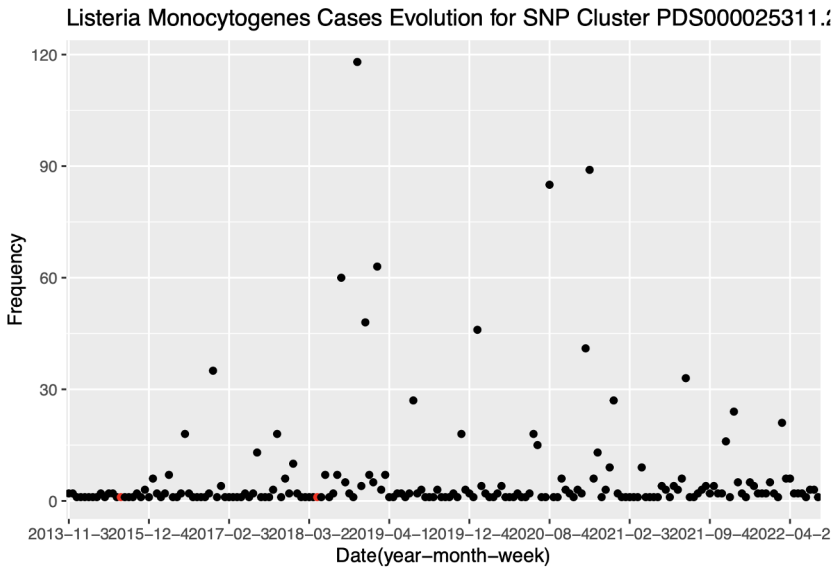
SNP Cluster Visualization by Month Continues



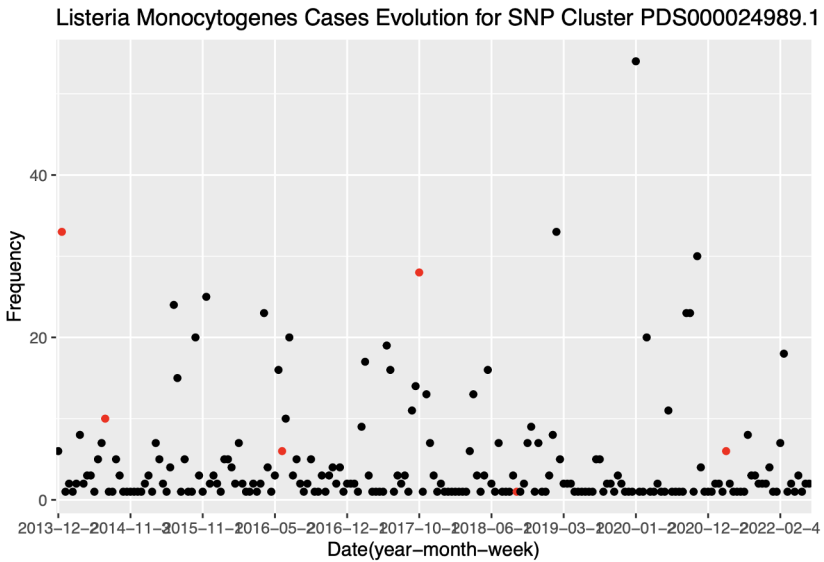


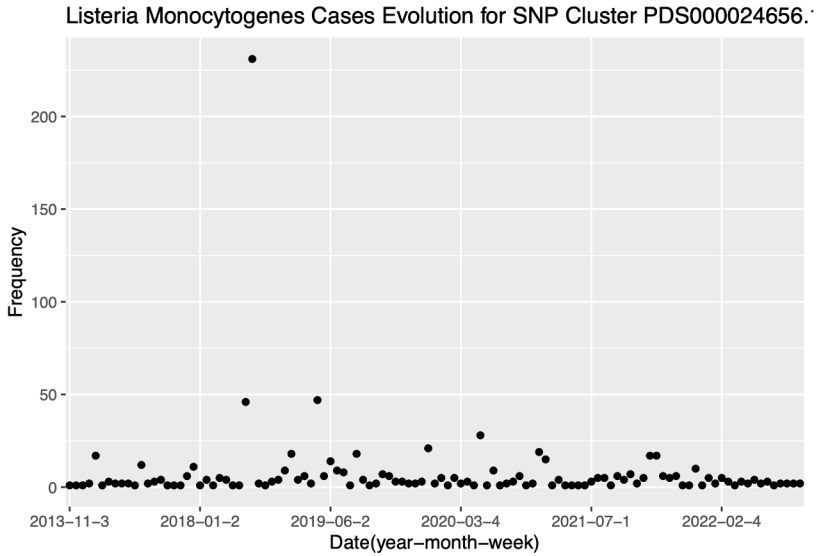
SNP Cluster Visualization by Week





SNP Cluster Visualization by Week Continues

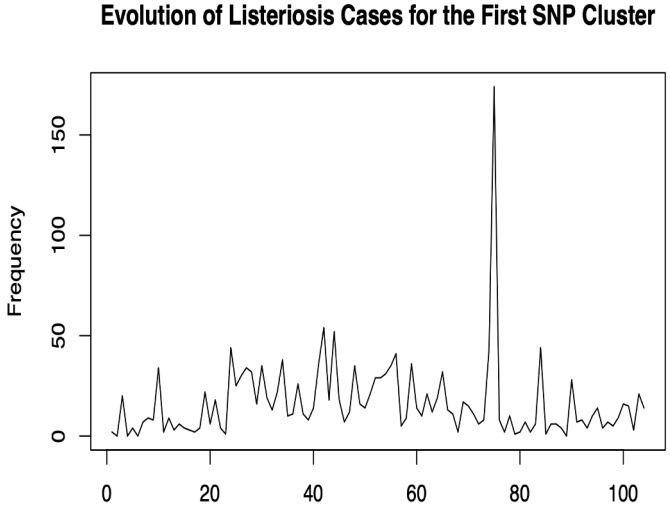


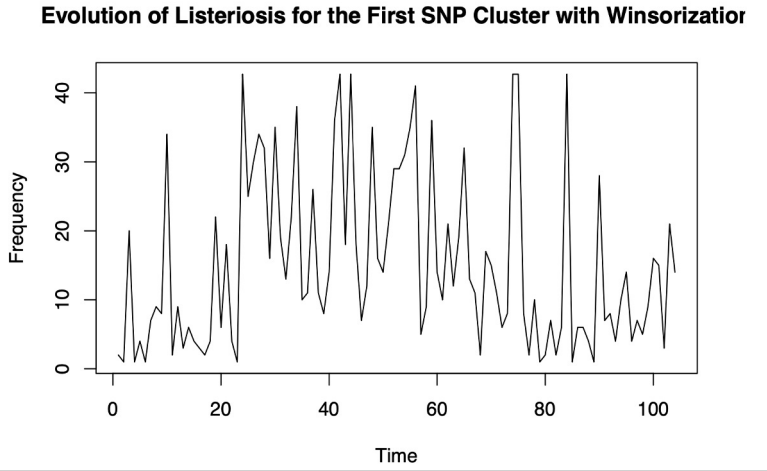


ARIMA and GARIMA

- I am going to build ARIMA and GARIMA models for both the SNP cluster contains the most Listeriosis cases as well as using the whole dataset
- In addition to using the largest SNP cluster and the whole dataset, I will use weekly count time series and monthly count time series to compare performance

Winsorization





XREG Matrix

```
##    February March April May June July August September October November
## 1         0     0     0   0    0    0      0         0       0        1
## 2         0     0     0   0    0    0      0         0       0        0
## 3         0     0     0   0    0    0      0         0       0        0
## 4         1     0     0   0    0    0      0         0       0        0
## 5         0     1     0   0    0    0      0         0       0        0
## 6         0     0     1   0    0    0      0         0       0        0
## 7         0     0     0   1    0    0      0         0       0        0
## 8         0     0     0   0    1    0      0         0       0        0
## 9         0     0     0   0    0    1      0         0       0        0
## 10        0     0     0   0    0    0      1         0       0        0
##    December
## 1         0
## 2         1
## 3         0
## 4         0
## 5         0
## 6         0
## 7         0
## 8         0
## 9         0
## 10        0
```

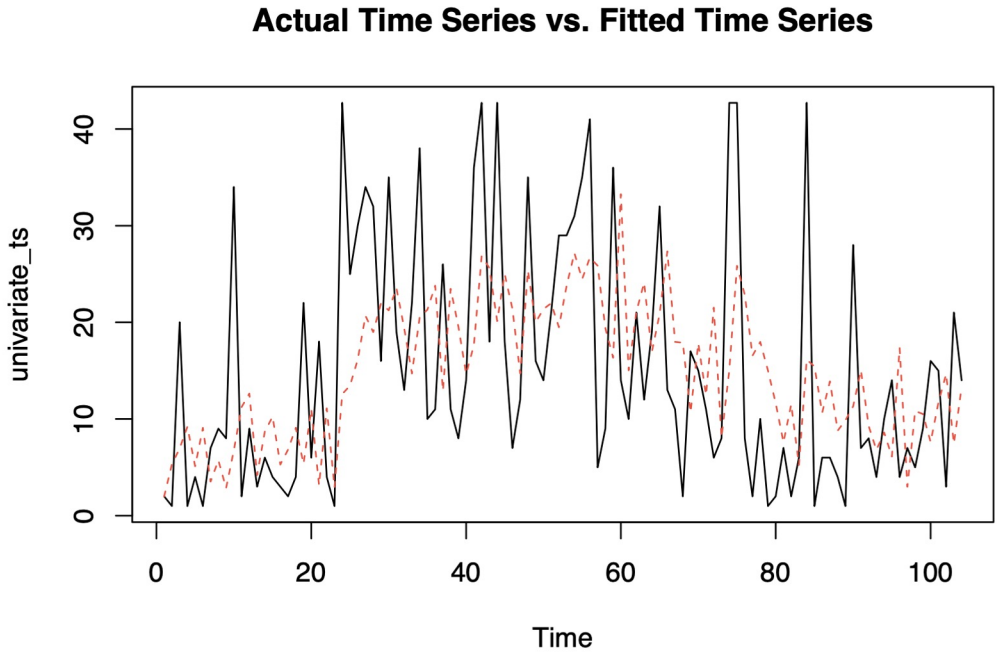
Time Series Analysis for the First SNP Cluster using Monthly Count Time Series (ARIMA)

- Nested cross-validation + Grid search to choose the best hyperparameters (p, d, and q)
- The best model with ... are deviation (RMSD) is the model with p = 0, d = 1, and q = 2

```
## [1] 22.21181
```

```
##        ma1        ma2   February      March      April        May       June
## -0.6889653 -0.1722320 -3.4419724 -2.2451469  1.1401353 -1.8485463 -2.2397674
##       July     August  September    October   November   December
## -5.7064539 -3.2508873 -6.6548814  2.3614807 -6.4797639 -2.2027550
```

Time Series Analysis for the First SNP Cluster using Monthly Count Time Series (ARIMA) Continues



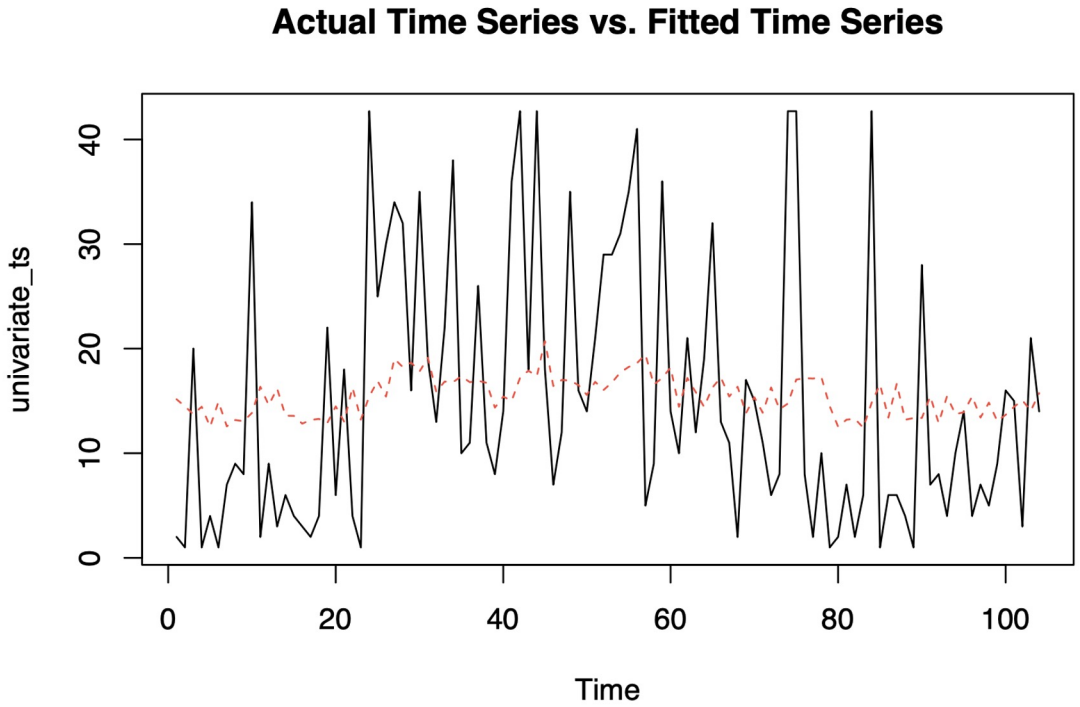
Time Series Analysis for the First SNP Cluster using Monthly Count Time Series (GARIMA)

- The best model with the lowest root-mean-square deviation (RMSD) is the model with p = 5 and q = 2

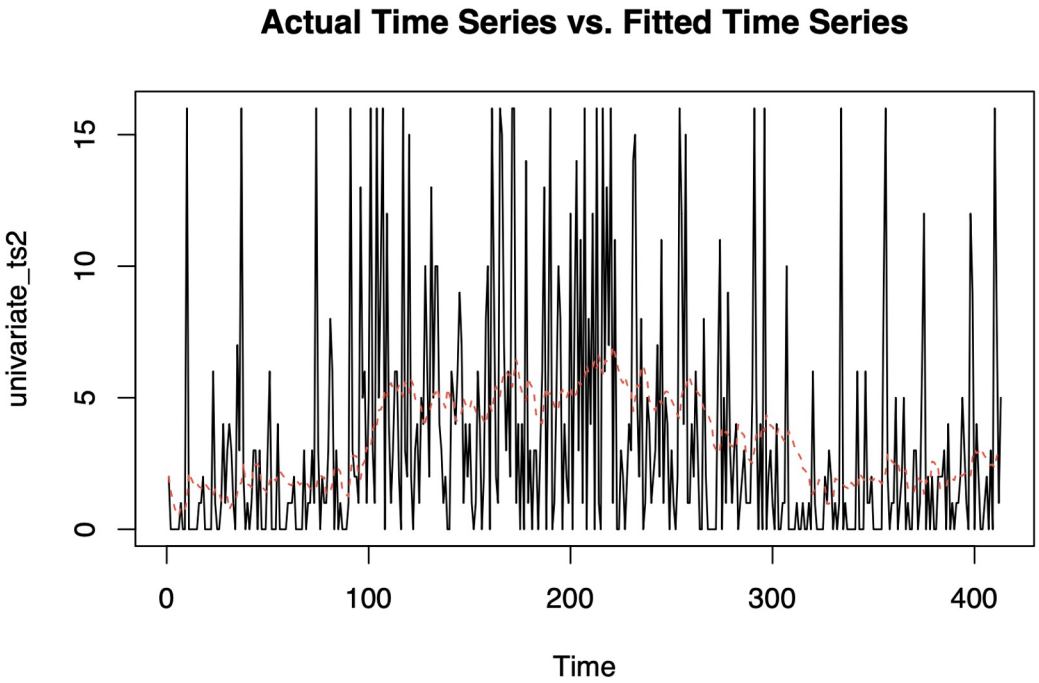
```
## [1] 20.50862
```

```
##  (Intercept)       beta_1       beta_2       beta_3       beta_4       beta_5
## 1.162981e+01 1.008412e-01 1.534862e-07 8.960506e-02 1.064954e-02 5.701031e-11
##      alpha_1      alpha_2     February        March        April          May
## 2.676415e-11 3.145270e-02 1.272345e-04 2.519806e-01 6.244795e-01 1.407905e-01
##         June         July       August    September      October     November
## 8.998533e-05 6.265522e-05 3.695883e-01 1.148869e-04 1.515120e+00 3.480620e-06
##     December
## 6.512570e-01
```

Time Series Analysis for the First SNP Cluster using Monthly Count Time Series (GARIMA)

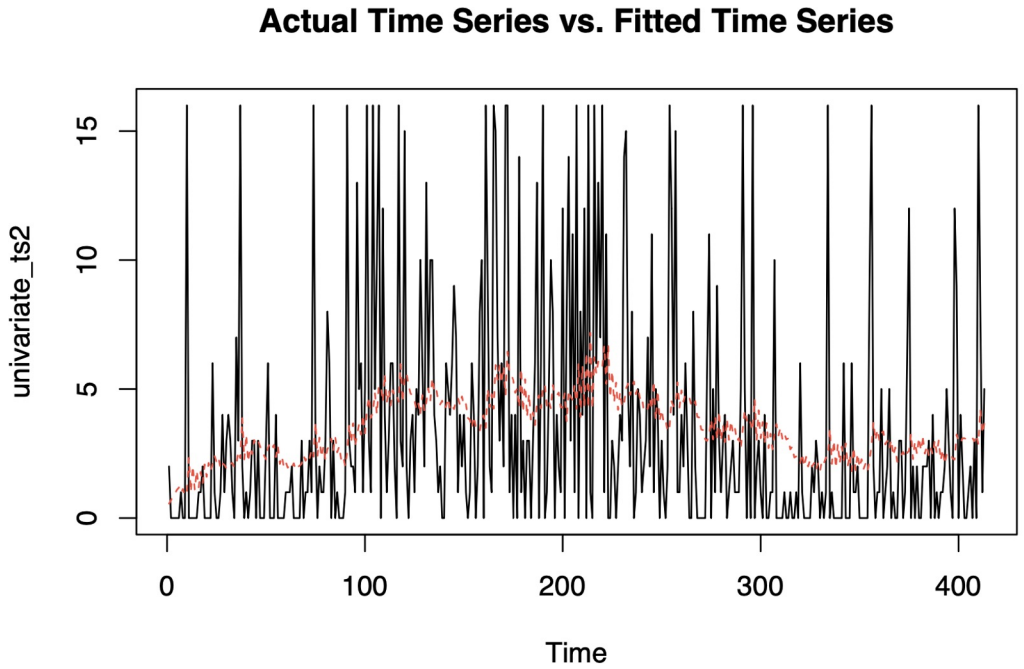


Time Series Analysis for the First SNP Cluster using Weekly Count Time Series (ARIMA)



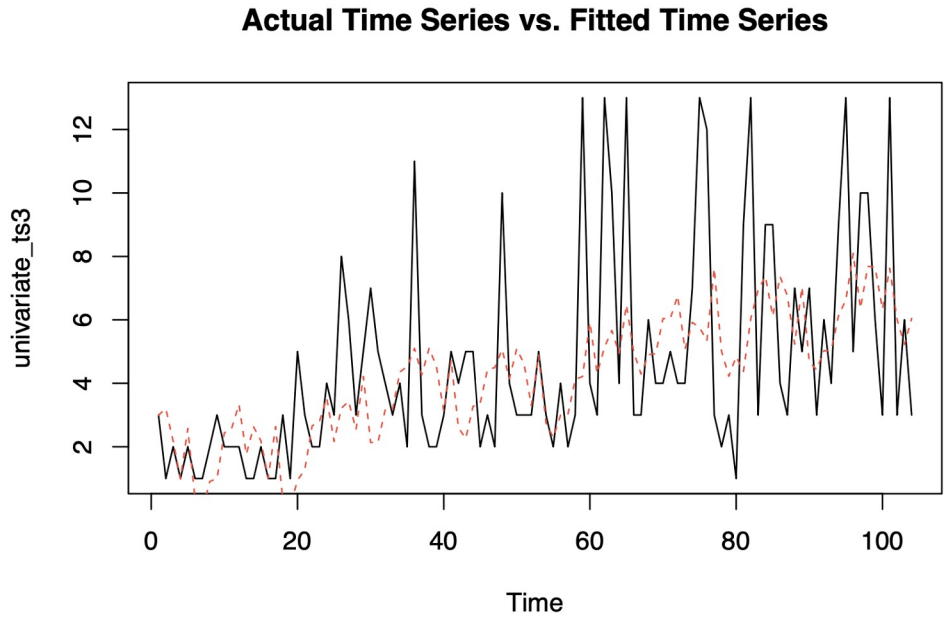
p = 0
d = 1
q = 1

Time Series Analysis for the First SNP Cluster using Weekly Count Time Series (GARIMA)



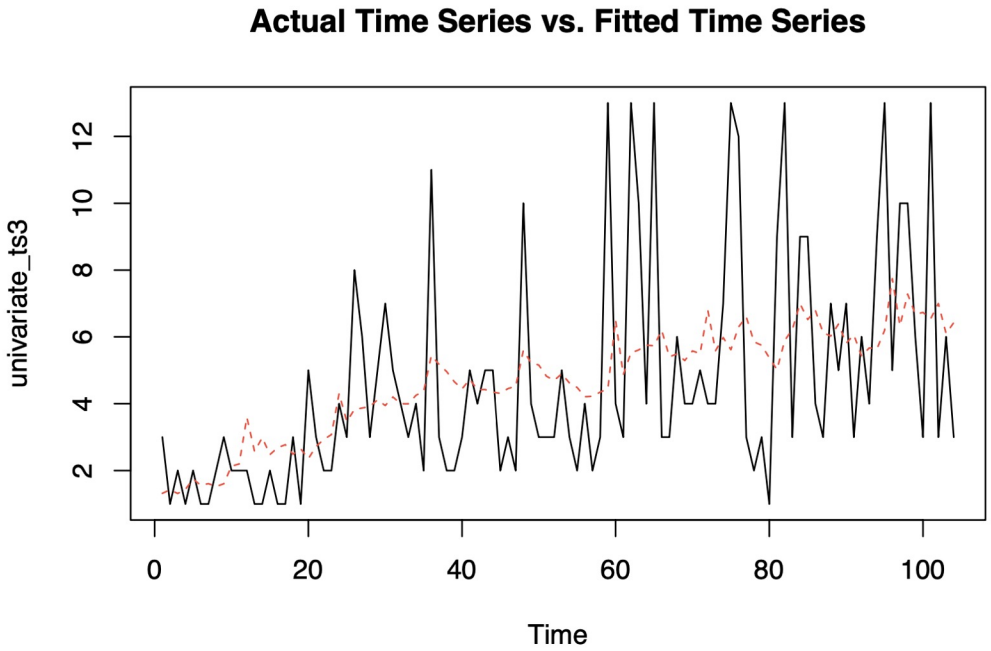
p = 1
q = 3

Time Series Analysis for the Whole Dataset using Monthly Count Time Series (ARIMA)



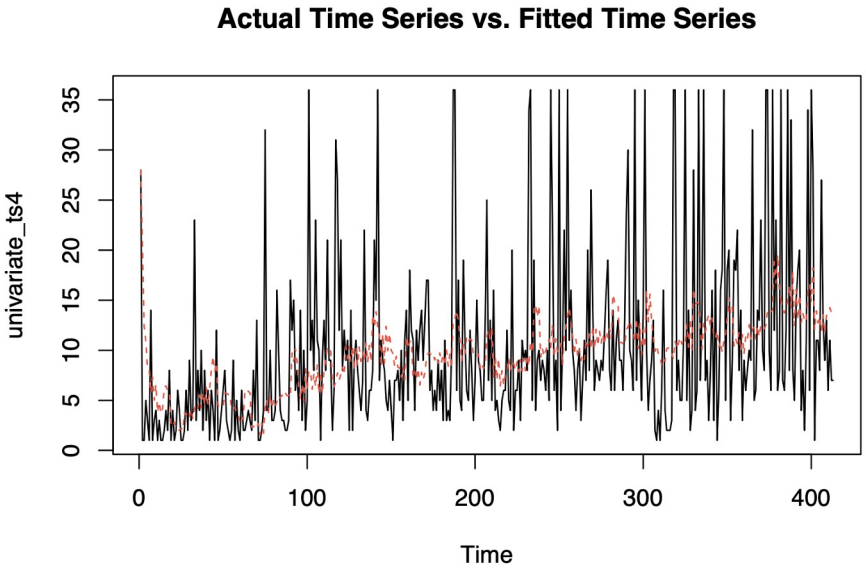
p = 0
d = 1
q = 1

Time Series Analysis for the Whole Dataset using Monthly Count Time Series (GARIMA)



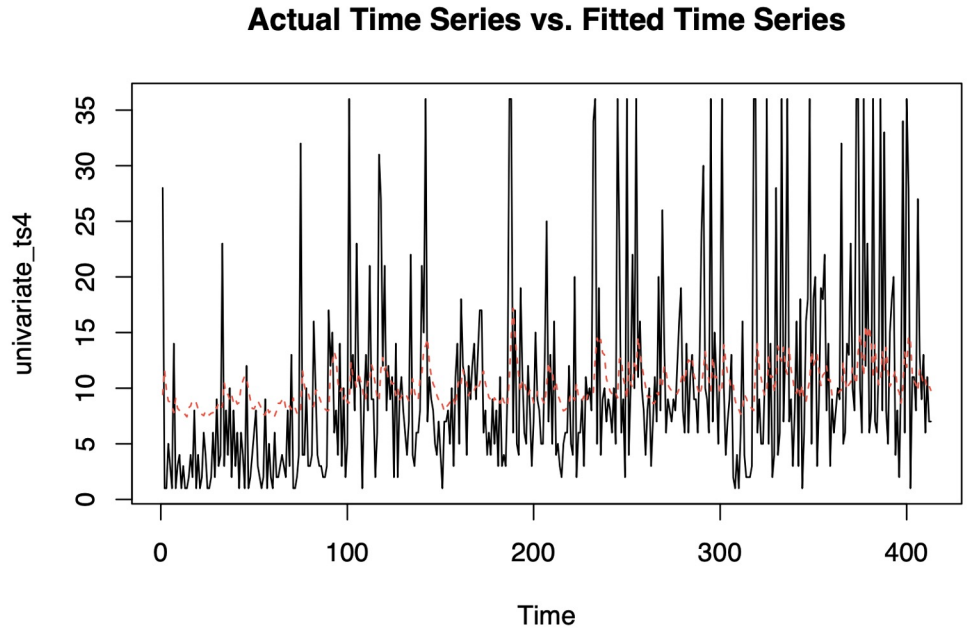
p = 1
q = 2

Time Series Analysis for the Whole Dataset using Weekly Count Time Series (ARIMA)



p = 1
d = 1
q = 3

Time Series Analysis for the Whole Dataset using Weekly Count Time Series (GARIMA)



p = 2
q = 2

Summary

```
##                     AIC       BIC      RMSD
## ARIMA(0,1,2)   822.7414  859.6276 11.338199
## GARIMA(5,2)         Inf       Inf 11.866818
## ARIMA(0,1,1) 2410.2531 2462.5264  4.351531
## GARIMA(1,3)  1931.7938 2000.1924  4.339950
## Arima(0,1,1)  540.7913  575.0427  2.907148
## GARIMA(1,2)   501.8040  544.1143  2.890416
## ARIMA(1,1,3) 2951.0196 3015.3560  8.323166
## GARIMA(2,2)  2733.2572 2801.6558  8.628902
```

Thank you for watching!