The School of Mathematics

THE UNIVERSITY
of EDINBURGH

Understanding the Epidemiology of Enteric Viruses in Dong Thap Province

by

Yifan Zeng, s1998233

July 2020

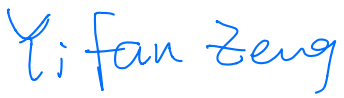
Supervised by
Dr. Serveh Sharifi, Dr. Simon Taylor and Steven Soutar

Acknowledgments

I am very grateful to the supervisors of this project: Dr. Serveh Sharifi, Dr. Simon Taylor and Steven Soutar. I also wanna say thanks Dr. Lu Lu, she provides the data and the background information of this subject.

Own Work Declaration

I confirm the work is done on my own and only supported by declared resources.
Edinburgh, 30th June 2020



Contents

Executive summary

Diarrhoea diseases as the second most cause of mortality in children younger than 5 years in 2014 are most caused by enteric viruses and are of high attention by researchers. The Vietnam Initiative on Zoonotic Infections (VIZIONS) is a project collaborate with Vietnamese organizations and provides resource for research on zoonotic diseases.

This report study the epidemiology of enteric viruses in Dong Thap province, a province in the Mekong Delta and Dong Thap Muoi region of southern Vietnam. QGIS is used to visualize the spatial distribution of viruses, the plot shows that most viruses are centering at Tp.Cao Lanh city and spreading towards the north of Dong Thap. Then risk factors corresponding to co-infections count are studied by using a hierarchy Bayesian Poisson generalized linear model. The risk factors are considered to be Age, Gender-Male, Killing_Animal and Water_Source: River/Tap/Well. Finally, the relationship between co-infections and disease severity is studied by using Kernelized Ridge Regression model, results show that both common viruses and uncommon viruses increase the severity and the former contributes much more.

1 Introduction

1.1 Background and Motivation

Diarrhoea diseases rank as the fifth leading cause of deaths worldwidePang et al. (2014). And it is also the second most common cause of mortality in children younger than 5 years, which account for 18% of 10.6 million of yearly deaths in 2000-03, studied by WHO in 2014Bryce et al. (2005). People in industrializing countries are disproportionately affected, and in such locations the symptoms are usually severe and requiring hospitalizationMy et al. (2011). Enteric viruses are the most important etiological agents of diarrhoea inducing high morbidity in developing countries. The most prevalent viruses causing gastroenteritis are Rotavirus, Norovirus, Mastadenovirus, Kobuvirus, Sapovirus and Mamastrovirus. And many of these viruses have zoonotic origins.

As the severe consequences of viruses, making better understanding of the epidemiology of enteric viruses is of vital important. Many effort about the virus genome or antibiotics has been taken. In this report, we pay more attention to the spatial distribution pattern and co-infection analysis of enteric viruses.

1.2 Data Source

The Vietnam Initiative on Zoonotic Infections (VIZIONS) is a countrywide project established by several international institutions collaborate with Vietnamese organizations. VIZIONS aims at address relevant one-health questions by combining clinical data, epidemiology, high-throughput sequencing, and social sciences. They collected data by Hospital-based surveillance. VIZIONS also provides platform and resource for further research on zoonotic disease agentsRabaa et al. (2015).

For this report, data is consist of 707 patients' demographic information, testing results and living behaviours. They are enrolled in VIZIONS during 2012-2016 and most (95%) of them are from Dong Thap, a province in the Mekong Delta and Dong Thap Muoi region of southern Vietnam as Figure.1 showing. Dong Thap is a rural province between latitude $10°58'01 \cdot 70$ and $10°0736 \cdot 56$ N and longitude $105°5647 \cdot 22$ and $105°1116 \cdot 30$. Covering an area of $3238km^2$, the province is home to about 1·6 million peopleCarrique-Mas et al. (2014).

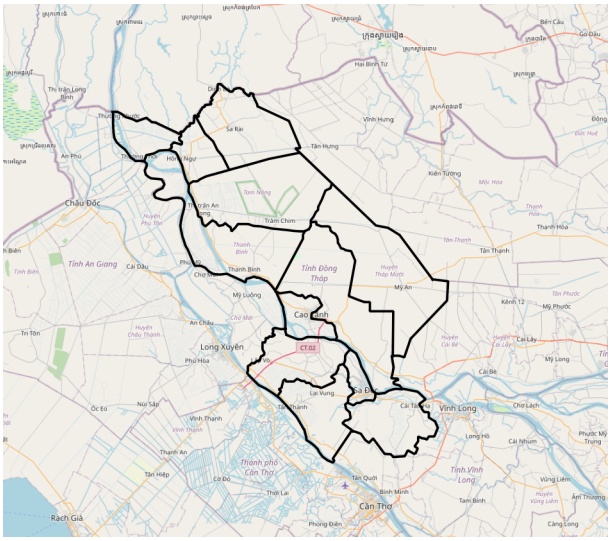


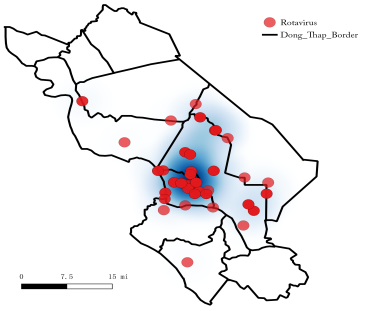
Figure 1: Location of Dong Thap province.

2 Spatial Distribution of Enteric Viruses

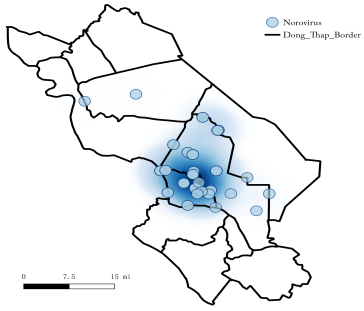
The data also record living locations of enrolled patients and these longitude and latitude were then visualized in QGIS v3.12.3 (http://www.qgis.org/en/site/) overlaid with province-specific geographic data Phan et al. (2016).

2.1 Distribution of Common Viruses

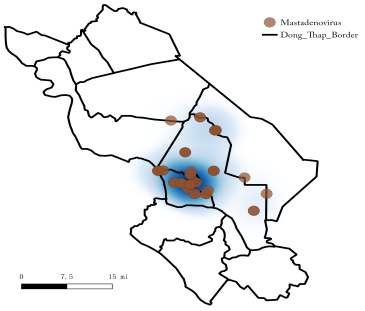
Figure.2 shows the spatial distribution of six types of common viruses inside Dong Thap city. For each separate plot the colorful points represent locations where there are patients infected with viruses tested by deep sequencing test, each point is of 30% transparency as many points are located closely with others. The blue heat maps below points are generated via the same test result data and are attached for observing the distribution easily.



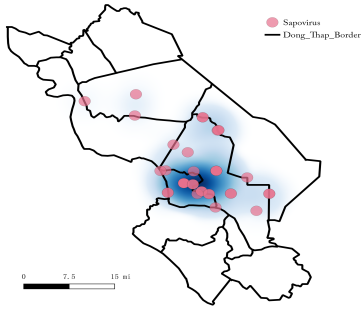
(a) Rotavirus



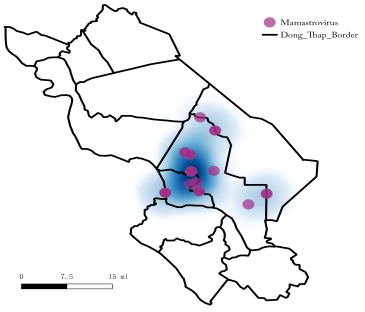
(b) Norovirus



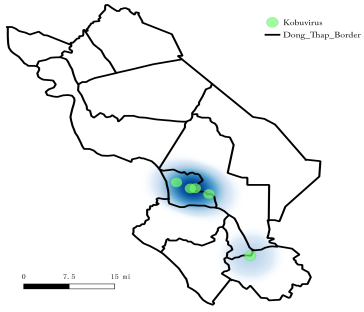
(c) Mastadenovirus



(d) Sapovirus



(e) Mamastrovirus



(f) Kobuvirus

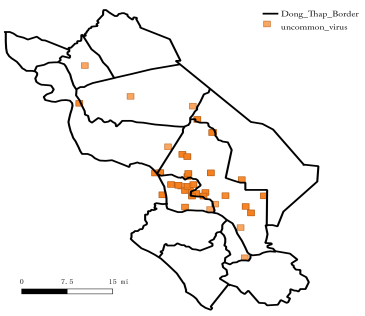
Figure 2: Spatial distribution of common viruses: Rotavirus, Norovirus, Mastadenovirus, Sapovirus, Kobuvirus and Mamastrovirus. Points are of 30% transparency.

Although these viruses are all of common virus. The frequencies of each one are varied which can be easily seen from the distribution figures. The patient counts of getting infected of each virus are respectively 151(Rotavirus), 89(Norovirus), 51(Mastadenovirus), 57(Sapovirus), 22(Mamastrovirus), and 5(Kobuvirus).

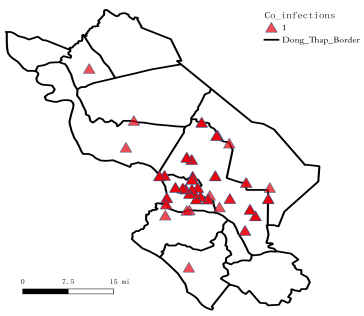
It's easy to find that all common viruses have similar distribution pattern which is centered at the mid of Dong Thap province, where is Tp.Cao Lanh city. Tp.Cao Lanh is the capital city of Dong Thap and has the second highest population density(1398 per km^2, https://en.wikipedia.org). Southern part of Dong Thap is least serious for common viruses infection, and the northern part is relatively less serious.

2.2 Distribution of Uncommon Viruses and Co-infections

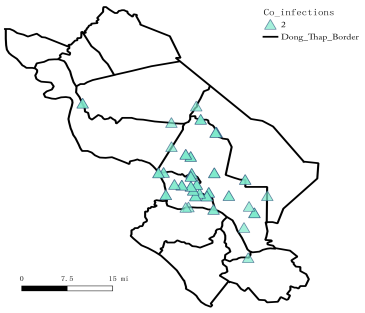
There are 25 types of uncommon viruses tested by deep sequencing technique. In Figure.3 (a) each point means patient at this location has been tested positive for at least one type of uncommon viruses.



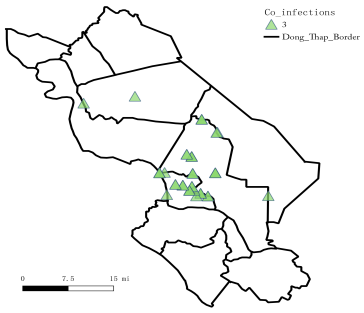
(a) Uncommon viruses



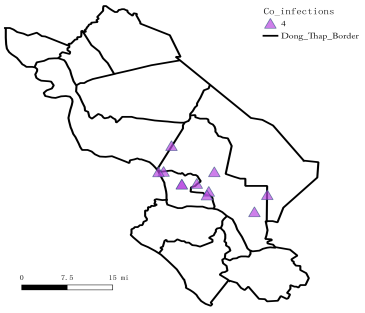
(b) Co-infections 1



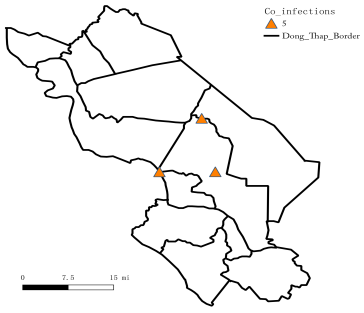
(c) Co-infections 2



(d) Co-infections 3



(e) Co-infections 4



(f) Co-infections 5

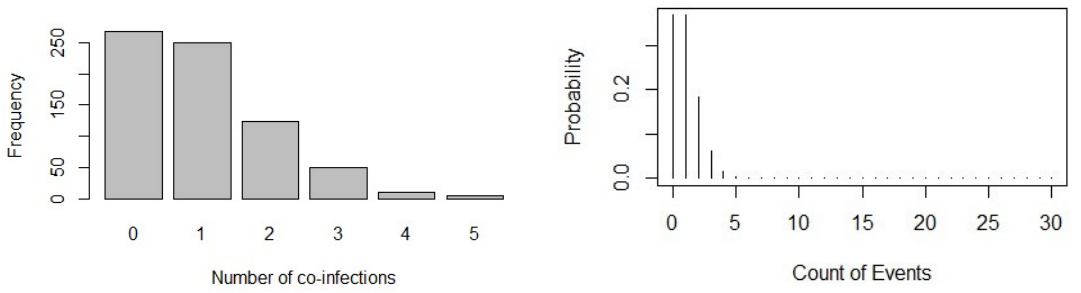
Figure 3: Spatial distribution of uncommon virus and co-infections. Points are of 30% transparency.

Compared with distribution pattern of common viruses, it's similar for that of uncommon virus, also centered in Tp.Cao Lanh district and spreading towards the north. Figure.3(b)-(f) show the distribution of different count of co-infections and points are also of 30% transparency to avoid overlapping. It's easy to see that co-infections counts are almost 1 or 2, only a few patients get large number of co-infections. As co-infections count distribution is another expression of viruses distribution, their distribution patterns are certainly similar.

As a result, the severity of center part of Dong Thap, or in other words the Tp.Cao Lanh city, is the most serious. The severity of northern part is relatively mild. The south of Dong Thap has the lowest infection rate. Studying the distribution of viruses among whole province gives the idea of including site information into model while looking deep inside the epidemiology of enteric viruses in Dong Thap province.

3 Bayesian GLM Analysis for Risk Factors Corresponding to Virus Co-infections

VIZIONS uses deep sequencing technique to test 6 common viruses, 25 uncommon viruses and unclassified viruses. A variable called co-infections is set as the sum value of common viruses count and uncommon viruses count. Deep sequencing test results tell that there are 38% patients have no viruses, 35% have single virus and 27% have co-infections. As co-infections may increase the difficulty of diagnosis and medical treatment. In this part, the aim is to seek the risk factors corresponding to virus co-infections in a bayesian way.



(a) Frequency of virus co-infections

(b) Poisson distribution with lambda = 1

Figure 4

There are many variables associated with virus co-infections, and likely to include demographic, living behaviour and agricultural practices factors that induce high level co-infections. `Age`, `Gender`, `Site_Recruitment`, `Contact_Diar` , `Water_ Sources(Tap, Well, River, Rain, Pond, Bottled)`, `Keep_Animal`, `Killing_Animal`, and `Eat_Cook_Raw_Meat` are selected from data provided by VIZIONS as potential factors[1].

3.1 Data Imputation

For response variable `co_infections`, original `NA`s are transferred into 0 which indicating there is no virus infected. As a result, there are 6 co-infections situations as Figure.4 showing.

[1]
`Contact_Diar` :whether or not patient has contacted other diarrhoea patients
`Keep_Animal` :whether or not patient keep animal
`Killing_Animal` :whether or not patient has killed animal
`Eat_Cook_Raw_Meat` :Whether has eaten/handled raw meat from any animal within 2 weeks of illness onset

For explanatory variables `Contact_Diar`, 18 out of 707 patients have contacted with diarrhoea and 675 out of 707 have not. For the other 14 unknown patients, their records are imputed with the most common one which is 'have not contacted'. And one kind of `Water_Sources` binary variable is removed from this model, which is `Other_WS`, because all patients show `FALSE` for this variable which provides no information and induces mistake while fitting models.

3.2 Exploratory Analysis

Before feeding the data into models, data visualization is carried out for exploratory analysis. Figure.5 shows violin plots of co-infections counts vs six different water sources. As there are just 3 and 13 patients who have taken `Pond` and `Rain` water respectively, which means the data source is quite unbalanced, the expectation may be biased when just looking at the plot. For `Well`, median value of co-infections are slightly larger for patients who have taken it. And the density of patients who have used `Tap` is also slighter higher. However, patients who have taken `Bottled` water source are showing a lower co-infections count level. Besides, pairplots for other

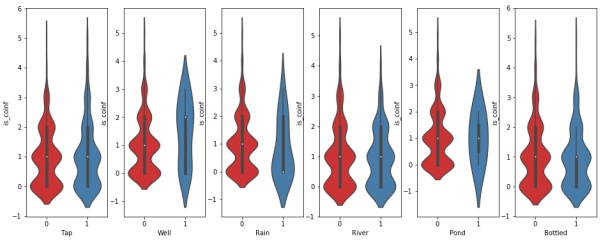


Figure 5: Violin plot of Co-infections vs water source.

variables are shown in Figure.6 . It's easy to see that there is a decreasing trend for co-infections while age increasing which is shown in subplot (1,2). It is reasonable as compared with adults children are more likely to be infected with viruses. In subplot(1,3), the number for male(0) is slightly higher than that of female(1), this phenomena may be induced by different jobs people are liable to take when gender varied. There's no significant trend for variables `Keep_Animal` and `Eat_Cook_Raw_Meat`. As subplot(4,4) and subplot(6,6) showing, which is the density plot of variables `Contact_Diar` and `Killing_Animal` respectively, the data set is quite unbalance for these two variables.

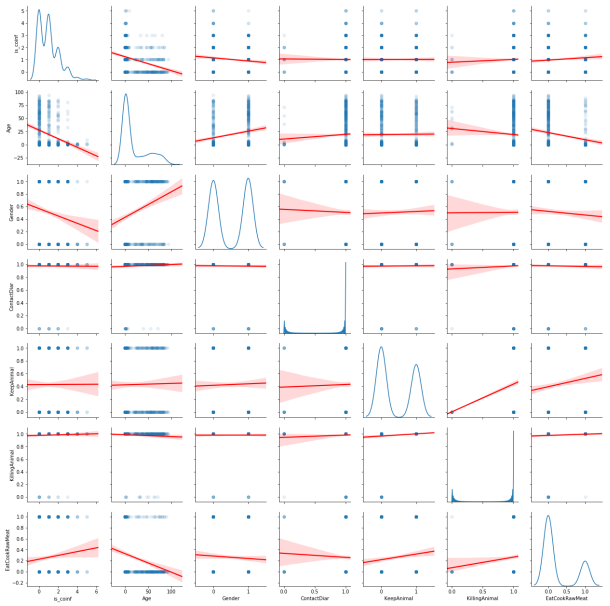


Figure 6: Pairplots of Co-infections vs potential risk factor variables.

3.3 Model

Count data variable takes discrete values which reflects the number of occurrences, it takes only on positive integer values or zero. The regression coefficients for that kind of response variable can be very unstable and will have large variance while using standard ordinary least square estimation. At the same time, Poisson distribution differs from Normal distribution as it takes on a probability value only for non negative integers.Coxe et al. (2009) As Figure.4 showing, the response variable `co_infections` can be taken as count data and the shape of sample data is similar to that of Poisson distribution with $\lambda = 1$. Besides, the link function for Poisson regression is natural log.

3.3.1 Poisson Generalized Linear Model

Generalized Linear Model(GLM) refers to the models where individuals y_i are independently distributed with exponential family pdf,

$$f(y_i|\theta_i) = exp[a^{-1}(\phi_i)\{y_i\theta_i - \psi(\theta_i)\} + c(y_i;\phi_i)]$$

where θ_i is unknown but $a(\phi_i)(> 0)$ are known. The assumption for θ_i is

$$\theta_i = h(x_i^T b)$$

where $h(\cdot)$ is a a strictly increasing sufficient smooth function, b is the vector of unknown regression coefficients and $x_i(p \times 1)$ are known design vectors of dimension p.

Maximum likelihood estimation is the classical way to calculate the coefficients. Let $X^T = (x_1, ..., x_n)$ has rank p, the likelihood function is given by

$$L(b) \propto exp[\sum_{i=1}^{n} a^{-1}(\phi_i) y_i h(x_i^T b) - \psi(h(x_i^T b))]$$

then the score vector is

$$\frac{dlog(L(b))}{db} = \sum_{i=1}^{n} a_i^{-1}(\phi_i)\{y_i - \psi'(h(x_i^T b))\}h'(x_i^T b)x_i$$

The maximum likelihood estimators are obtained by $\frac{dlog(L(b))}{db} = 0$Dey et al. (2000).

When it comes to count data, Poisson distribution is always considered to be the exponential family distribution and mean $\lambda = \theta = exp(x_i^T b)$, $a(\phi_i) = 1$. It has to be mentioned that as the link function between distribution mean and linear combination is $log(\cdot)$, the mean is multiplied by factor $exp(b_i)$ when increasing one unit of x_i .

3.3.2 GLM From Bayesian Perspective

From Bayesian perspective, regression coefficients b_i are assumed of independent prior distributions.When no information is available, prior distributions are of mean 0 and large value variance which make prior parameter scattered around zero and shows high uncertainty of priorsSpiegelhalter et al. (2004).

Then the posterior distribution of parameters by Bayes estimation is

$$p(\theta|Y) = \frac{p(\theta, Y)}{p(Y)} = \frac{p(Y|\theta)p(\theta)}{\int \cdots \int p(Y|\theta)p(\theta)d\theta}$$

But the above posterior is not analytically tractable. Also, finding posterior means, variances etc. by numeric integration is not easy even for moderate pDey et al. (2000). The most convenient approach seems to be Markov Chain Monte Carlo(MCMC), in other words, by generating samples from the posterior. To ascertain whether variable x_i has significant correlation with response variable or not, we just need to check if posterior b_i are still centered around zero or not.

For count data, which means using Poisson bayesian GLM, $\lambda = \theta = exp(x_i^T b)$, we also look at $exp(b_i)$ to learn how variable x_i influences response variable.

3.4 Model Setting

As patients are recruited from four sites: Dak Lak, Dont Thap, Hue and Khanh Hoa. Risk level of these four province are different. Hence `Site_Recruitment` is considered having random effect on `Co_infections`. Poisson hierarchy bayesian GLM is carried out to seek the risk factors corresponding to virus co-infections. The mathematical way to represent the model is:

- $\text{Co\_infections}_{ij} \sim Poisson(\mu_{ij})$
- $\log(\mu_{ij}) = \beta_0 + \beta_1(x_{i1} - \bar{x_1}) + ... + \beta_k(x_{ik} - \bar{x_k}) + \theta_j; \quad i = 1, ..., 707; j = 1, ..., 4$
- $\beta_k \sim N(0, 10); \quad k = 1, ..., 13$
- $\theta_j \sim N(0, (\sigma_j)^2); \quad \sigma_j \sim U(0, 10)$

As there's no enough information of the coefficients, the prior distribution of β_k is set to be $Normal(0, 10)$ which is flat and scattering around 0. Three chains starting from different initial nodes are used and the iterative times is 100000 with 6000 times burn-in. All continuous variables have been centered for better mixing. Model is carried out in `RStudio 1.2.1335`(package `rJAGs`).

3.5 Model Result

The model result is shown as below, where the `beta0` is the intercept value while calculating μ and `beta.ar[j]` is the random effect variable for different site recruitments.

Coefficient Name	Mean	SD	Naive SE	Time-series SE
beta0	-2.30729	4.727975	8.632e-03	1.514e+00
beta.Age	-0.01384	0.001923	3.511e-06	6.257e-06
beta.Gender	-0.08190	0.077440	1.414e-04	2.912e-04
beta.ContactDiar	0.08807	0.241245	4.405e-04	4.542e-03
beta.EatCookRawMeat	-0.03289	0.090441	1.651e-04	3.945e-04
beta.KeepAnimal	-0.00129	0.078381	1.431e-04	2.836e-04
beta.KillingAnimal	0.16222	0.326175	5.955e-04	8.202e-03
beta.Bottled	0.04097	0.242921	4.435e-04	4.423e-03
beta.Pond	-0.35510	0.668311	1.220e-03	4.478e-03
beta.Rain	-0.16508	0.380541	6.948e-04	4.154e-03
beta.River	0.06068	0.251546	4.593e-04	4.621e-03
beta.Tap	0.15068	0.243885	4.453e-04	4.563e-03
beta.Well	0.23057	0.316046	5.770e-04	4.536e-03
beta.ar[1]	2.07018	4.694686	8.571e-03	1.610e+00
beta.ar[2]	1.67758	4.682808	8.550e-03	1.438e+00
beta.ar[3]	2.12603	4.696619	8.575e-03	1.647e+00
beta.ar[4]	2.46355	4.731856	8.639e-03	1.655e+00

The estimated posterior mean of co-infections is 0.806, which is calculated by `R` command `E[exp(beta0)]`. The standard deviation of the random effect of the four recruitment sites captures a non negligible part of variability in the response compared to the magnitude of the other covariates.

For demographical variables, `Age` decreases the mean of co-infections slightly by 2.4%, it's reasonable cause children are more likely to get infected by research. `Female` also decreases the mean which is by 7.6%. That's because male have higher probability working on jobs contacting with animals, such as butchers or animal keepers.

For living behaviours, whether or not keeping animals has little(0.002%) influence on mean of co-infections, while killing animal induces much higher mean of co-infections(24.3%). However, the variable `EatCookRawMeat` decreases co-infections by 2.8%, this is not consistent with what before researches telling us, so this variable need reconsideration.

For water source variables, the results tell that `River`, `Tap` and`Well` increase co-infections by 9.6%, 19.7% and 32.3% respectively. `Pond` and `Rain` decrease by 14.3% and 9%, this result is not reliable because of the unbalance of data. The numbers of patients who have drunk pond water and rain water are 3 and 13 out of 707 respectively.

As a result, the risk factors for co-infections are `Age`, `Gender-Male`, `Killing_Animal` and `Water_Source: River/Tap/Well`.

3.6 Model Check

3.6.1 Gelman-Ruben-Brooks statistic

`GRB` statistic is a quantitative method to assess the mixing(or lack of). While running several chains of `MCMC`, the variability in the simulated values within each chain and that between chains are changing in different speed. Comparing these two variability is the idea of `GRB` statisticBrooks & Gelman (1998), and it is defined as:

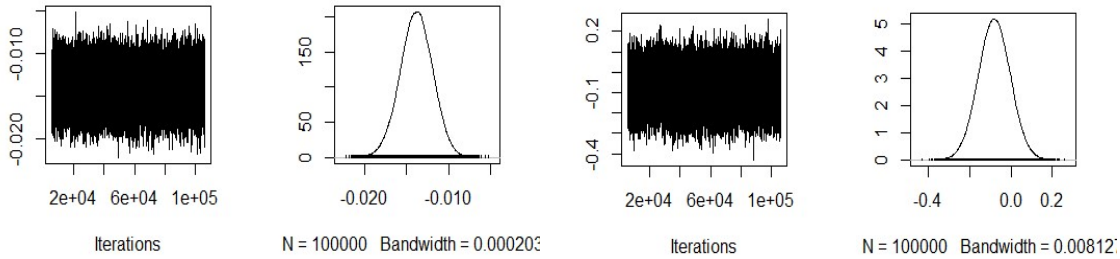
$$R = \frac{Width\_of\_80\%\_credible\_interval\_of\_all\_chains\_combined}{Average\_width\_of\_80\%\_credible\_intervals\_for\_the\_individual\_chains}$$

If the chain has converged, then R should be around 1. If the chains have not converged, then the width of the interval based on all chains combined will be greater than 1, as the between chain variation will be "large". The `GRB` statistics test results are shown as below, the values are all 1.00 which means well-converged.

Coefficient Name	Point est.	Upper C.I.
beta0	1.00	1.00
beta.Age	1.00	1.00
beta.Gender	1.00	1.00
beta.ContactDiar	1.00	1.00
beta.EatCookRawMeat	1.00	1.00
beta.KeepAnimal	1.00	1.00
beta.KillingAnimal	1.00	1.00
beta.Bottled	1.00	1.00
beta.Pond	1.00	1.00
beta.Rain	1.00	1.00
beta.River	1.00	1.00
beta.Tap	1.00	1.00
beta.Well	1.00	1.00

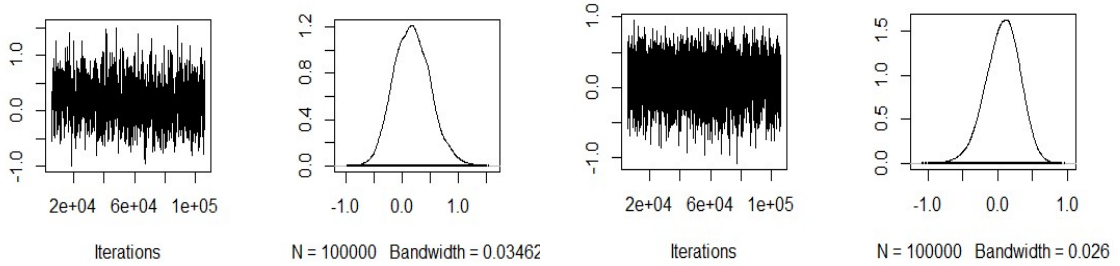
3.6.2 Trace Plot and Density

Another way to check the model is by looking at the trace plot and density plot of parameters. Below is some of the plots, there is no evidence showing any 'stuck' in these plots which means the model well-built.



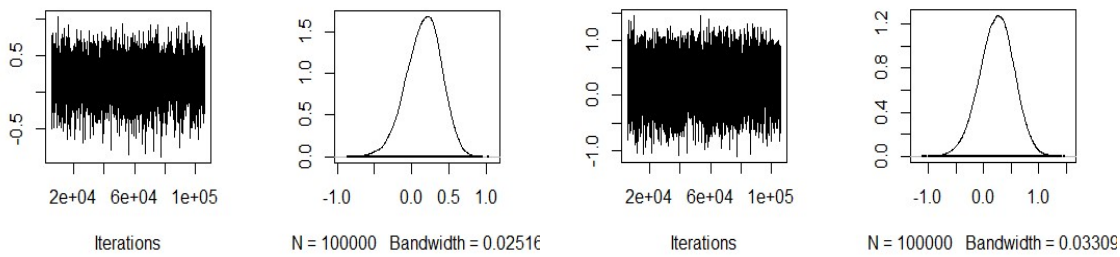
(a) Age

(b) Gender



(c) Killing Animal

(d) River



(e) Tap

(f) Well

Figure 7: Trace plots and density plots for some of parameters.

4 How Co-infections Impact Disease Severity

Symptoms induced by enteric viruses are also recorded by VIZIONS. As data showing, 25 out of 707 patients have blood in stool, 106 patients have mucoid in stool and 302 patients suffer abdominal pain. Besides, 131 patients get three days fever and all patients are showing varied diarrhoea episodes symptom. In a word, all patients are of different level of co-infections and showing varied disease severity. In this part, we consider studying the effect of co-infections to disease severity. As all patients infected with enteric viruses need to be hospitalized, variable `Length_ of_Stay`, which is the count of days that patient is in hospitalization, is taken to be the represent of severity which is also the response variable here.

Binary variables `Common_Single`, `Uncommon_Single`, `Unclass_virus` and `Between_Virus` are created and involved as explanatory variables. These four dummy variables indicate whether or not a patient is infected just by common viruses, uncommon viruses, unclassified virus or both common virus and uncommon virus respectively. Two count data variables `Common_count` and `Uncommon_count` are also created and involved, these two are the number of common viruses and uncommon viruses respectively patient get infected. In this part, we pay more attention to the relationship between disease severity and above six variables because the question is how co-infections affect severity.

Besides, variables that are considered related with severity, in other word the length of stay,

are also involved into model for better regression. These variables are `Age`, `"NumberDiarEpi"`, `BloodStool`, `MucoidStool`, `AbdominalPain` and `ThreeDaysFever`[2].

4.1 Data Imputation

In the origin data set, the first 21 records for `NumberDiarEpi` are missing. As the order of patients in data set is of no regularity, this missing pattern has been taken as MAR(Missing at Random) . This variable is of continuous value for which mean value imputation method is available considering the missing pattern. As a result, mean value 7.97 is used. For variables `BloodStool`, `MucoidStool`, `AbdominalPain` and `ThreeDaysFever`, numbers of missing are small and are respectively 3, 4, 11 and 4. As these four are all binary variables, which only take value 0 or 1 to indicate Yes or No, missing values are imputated with the most common inside each one.

4.2 Exploratory Analysis

For response variable `Length_of_Stay` and continuous explanatory variables `Age` and `NumberDiarEpi`, Figure.8 is a pairplot to show the relationship between each other from a qualitative way. As we can see from subplot(1,1) which is the density estimation plot of `Length_of_Stay`, the shape is similar to that of Normal distribution which is good for regression models. And from the subplot(2,2) and (3,3), most patients are of young age and small number of diarrhoea episode. But the relationship between `Length_of_Stay` and other two variables are hard to see by jus looking at this plot.

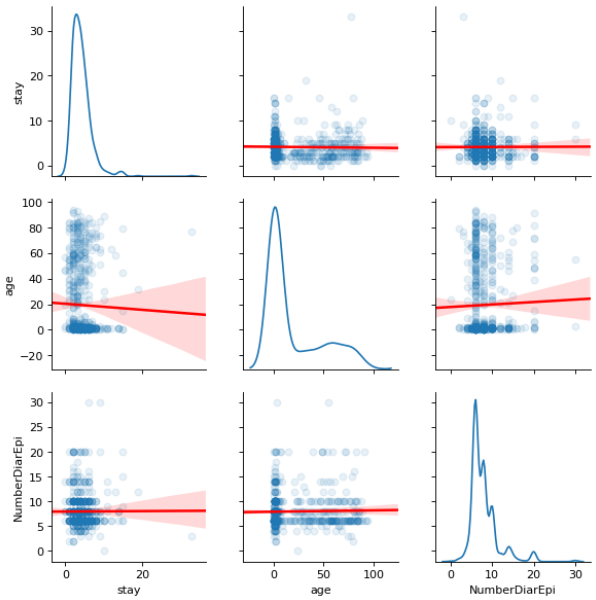


Figure 8: Pairplots of Length of stay vs continuous explanatory variables.

For the other binary variables or count data variables, violin plots of `Length_of_Stay` vs levels are more useful as Figure.9 showing. Among the symptoms in the first row, `ThreeDaysFever` seems to have a relatively significant influence, patients get fever are liable to stay longer in hospital. For other three symptom variables `BloodStool`, `MucoidStool` and `AbdominalPain`, the median values at varied levels are closely distributed. The long length of 'violin' at level

[2] `NumberDiarEpi` :number of diarrhoea episodes
`BloodStool` :whether there is or not blood in stool
`MucoidStool` :whether there is or not mucoid in stool
`AbdominalPain` :whether or not patient get abdominal pain
`ThreeDaysFever` :whether or not patient get three days fever

0(which means No) can tell us that there are few patients have stayed in hospital for a long period but without these symptoms which are taken serious by people.

At the second row of Figure.9, these variables are reflection of co-infections. The first plot at the second row, which is `Length_of_Stay` vs `Single_Common`, shows that most patients get infected only with common viruses stay in hospital for a longer period. And the forth plot at the second row, which is `Length_of_Stay` vs `Common_count`, also presents a increasing trend of stay period as common virus co-infections number boosting. As a result, getting infected with common viruses are expected to increasing the length of stay for patients. For the second and fifth subplots in the second row of Figure.9, although distribution of `Length_of_Stay` among varied levels of `uncommon_count` doesn't change obviously, more patients tend to have a longer stay at hospital when they get infected with just uncommon viruses.

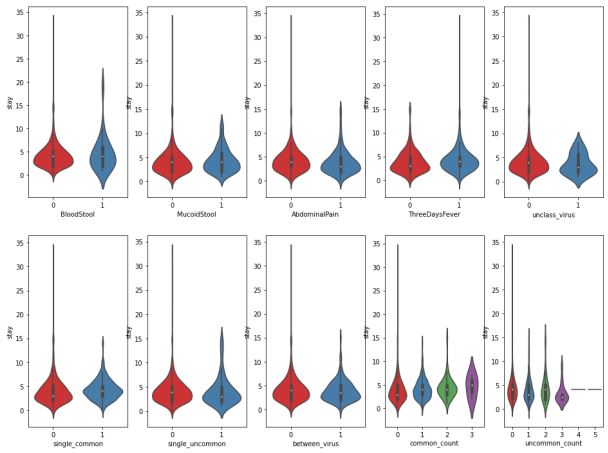


Figure 9: Violin plots of Length of stay vs levels of binary explanatory variables.

4.3 Model

4.3.1 Ridge Regression

After comparing Linear Regression model, Ridge Regression model, Lasso Regression model and Kernelized Ridge Regression model in terms of RMSE(Root Mean Square Error) value, Kernelized Ridge Regression model is used to find the relationship between co-infections and disease severity.

As we know, the Linear Regression coefficients are estimated by minimizing cost function, which is the residual sum of squares:

$$RSS = \sum_{n=1}^{N} (y_n - w^T x_n)^2$$

where N is the data size, x_n and y_n is the real value of explanatory variables and response variable, w^T is the coefficient vector to be estimated. And the estimators are:

$$\hat{w} = (X^T X)^{-1} X^T y$$

As all features that we consider might have influence on response variable are included into regression model, regularization is applied to overcome overfitting. Regularization is achieved by including penalization term in cost function, which is:

$$Cost\_Function = \sum_{n=1}^{N} (y_n - w^T x_n)^2 + \sum_{d=1}^{D} p_\lambda(w_d)$$

where $p_\lambda(w_d)$ is a penalization function for large weight values and λ is the tuning parameters. Different penalization function leads to different regularization method, one of them is Ridge Regression:

$$Ridge\_Cost\_Function = \sum_{n=1}^{N}(y_n - w^T x_n)^2 + \lambda \sum_{d=1}^{D} w_d^2$$

Horel (1962) we can see, the penalization strength depends on the tuning parameter λ in Ridge Regression. Larger value induce smaller weight in magnitude but less fitting accuracy. Then the Ridge Regression estimators are:

$$\hat{w}_{ridge} = (\lambda I_D + X^T X)^{-1} X^T y$$

where I_D is a $D \times D$ identity matrix.

4.3.2 Kernelized Ridge Regression

As the relationship between response variable and explanatory variables might be nonlinear, Kernel Trick technique is applied to Ridge Regression. Kernel Trick is an approach to explaining nonlinear mapping by replace all inner products $x^T x'$ with kernel function $k(x, x')$. For kernel trick to work, a positive definite kernel which is called *Gram Matrix* is need, it must be positive definite for all sets of inputs $\{x_n\}_{n=1}^N$, and the matrix is defined by:

$$K = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_N) \\ \cdots & & \cdots \\ k(x_N, x_1) & \cdots & k(x_N, x_N) \end{bmatrix}$$

The estimators of Ridge Regression model can be rewritten as:

$$\hat{w} = X^T(\lambda I_N + XX^T)^{-1} y$$

Then $\hat{w}$ can be kernelized by replacing XX^T with K.
Data fed into the model has been scaled to avoid the influence of magnitude and units. And the model is carried out in `Anaconda3` and the kernel function used by package`scikit-learn` is as follows:

$$K(x, y) = \exp(-\gamma \|x - y\|^2).$$

which is provided by the `KernelRidge` function in the `sklearn.kernel_ridge` submodule. All parameters needed are attained by cross validation method.

4.4 Model Result

Dual_coefficients are used to check the contribution of each variables. It is the coefficient in the Kernel space and can be used to check variable contribution, defined by:

$$Dual = (\lambda I_N + XX^T)^{-1} y$$

And if the kernel is linear, the dual coefficients from Kernelized Ridge Regression are same with the estimators from Ridge Regression model. The sign means positive or negative relationship between contributors and mean response variable, the value means how much the contribution is[3].

The result table is shown below. As we can see from the table, symptoms getting fever, having mucoid in stool, having blood in stool, later diarrhoea episode all increase the length of stay, among which getting fever increases most. It is consistent with what we got from the exploratory

[3]https://stackoverflow.com/questions/37045251/get-parameters-of-fitted-model-of-kernelridge-class-scikit-learn-library

analysing plots. Then variable `Age` has most negative contribution to mean length of stay, which means younger patients are in hospitalization for a longer period. It's also reasonable cause enteric viruses are more severe for children. For co-infections variables, the common virus count and single_common variables both do a positive contribution to disease severity while the latter has more serious effect. At the same time, getting infected only by uncommon viruses or common viruses also increase the disease severity.

The abdominal pain gets slightly negative contribution to mean length of stay, and also do getting infected with unclassified virus or between viruses indicators. That may because these symptoms have little effect on disease severity, shorter hospitalization periods are induced by other variables.

We can see that contribution of variables related with common viruses are much more than that of uncommon viruses. That may because as the length of stay doesn't only depend on disease severity, it also depend on how effective the medical treatment is and the speed patients be diagnosed accurately. The characteristic of common viruses are inducing more severe disease but are relatively easily tested by doctors, as a consequence faster and more effective medical treatment. The characteristic for uncommon viruses and unclassified viruses are inducing less serious symptoms but are more difficult to be tested.

As a result, both common virus co-infections and uncommon virus co-infections increase the disease severity but the former contributes much more. And when patients get infected with common viruses, disease severity increases because of virus itself. When patients get just uncommon viruses, disease severity also increases but because of late diagnose or incorrect medical treatment.

Variable	Dual Coefficient
ThreeDaysFever	80.2296
Number Diar Epi	55.4383
Common_count	50.2866
Single_common	46.3653
MucoidStool	26.0849
Single_uncommon	11.4629
BloodStool	10.4609
Unclass_virus	-11.3419
Between_virus	-14.3970
Uncommon_count	-14.8281
AbdominalPain	-108.9770
Age	-1030.9845

5 Conclusions

This paper present a report of the epidemiology of enteric viruses in Dong Thap province, a province in the Mekong Delta and Dong Thap Muoi region of southern Vietnam. Research data of 707 patients' demographic information, clinical data and living behaviors information are collected by VIZIONS (The Vietnam Initiative on Zoonotic Infections). The enteric viruses cause high mortality and morbidity of diarrhoea diseases especially among children. VIZIONS using deep sequencing technique to test 6 common viruses (Rotavirus, Norovirus, Mastadenovirus, Kobuvirus, Sapovirus and Mamastrovirus) and 25 uncommon viruses.

This report is consist of three parts. This first part is data visualization, spatial distribution of common viruses, uncommon viruses and virus co-infections are shown by scattering point above Dong Thap map. The plot shows that Rotavirus, Norovirus and Sapovirus have relatively higher frequency among common viruses. And most patients are infected with 1-3 co-infections, only 3 patient got five co-infections. The distribution pattern of all kind of viruses are similar, centering at Tp.Cao Lanh city, which is the capital city of Dong Thap province, and spreading towards the North of Dong Thap province.

The second part of this report study the risk factors corresponding to virus co-infections by using a hierarchy Bayesian Poisson generalized linear regression model. Demographical information and living behaviors variables are included into as potential risk factors of co-infection count. The prior distribution of coefficient is set to be $Normal(0, 10)$ and three chains staring from different initial nodes are used. The iterative times is 100000 with 6000 times burn-in. By looking at the estimated posterior mean of coefficients, the risk factor are Age, Gender-Male, Killing_Animal and Water_Source: River/Tap/Well.

The third part of this report is about how co-infections impact disease severity. As almost all patients infected with enteric virus need to stay at hospital, variable Length_of_Stay records the length of period of hospitalization for each patient and is considered as disease severity in model. Kernelized Ridge Regression model is used here. The count numbers of common viruses and uncommon viruses each patient get infected with are included as explanatory variables. Four binary variables indicating whether or not patient is infected with only common virus, only uncommon virus, both common virus and uncommon virus, and unclassified virus are also included as contributors. Besides, several symptom variables which may have influence on disease severity are included for better regression. The model result tells us both common virus co-infections and uncommon virus co-infections increase disease severity and the former contributes much more.

I hope this result can help prevent getting infected with enteric viruses and reduce disease severity for patients.

References

Brooks, S. P. & Gelman, A. (1998), 'General methods for monitoring convergence of iterative simulations', *Journal of computational and graphical statistics* **7**(4), 434–455.

Bryce, J., Boschi-Pinto, C., Shibuya, K., Black, R. E., Group, W. C. H. E. R. et al. (2005), 'Who estimates of the causes of death in children', *The Lancet* **365**(9465), 1147–1152.

Carrique-Mas, J., Bryant, J., Cuong, N., Hoang, N., Campbell, J., Hoang, N., Dung, T., Duy, D., Hoa, N., Thompson, C. et al. (2014), 'An epidemiological investigation of campylobacter in pig and poultry farms in the mekong delta of vietnam', *Epidemiology & Infection* **142**(7), 1425–1436.

Coxe, S., West, S. G. & Aiken, L. S. (2009), 'The analysis of count data: A gentle introduction to poisson regression and its alternatives', *Journal of personality assessment* **91**(2), 121–136.

Dey, D. K., Ghosh, S. K. & Mallick, B. K. (2000), *Generalized linear models: A Bayesian perspective*, CRC Press.

Horel, A. (1962), 'Applications of ridge analysis toregression problems', *Chem. Eng. Progress.* **58**, 54–59.

My, P. V. T., Rabaa, M. A., Vinh, H., Holmes, E. C., Hoang, N. V. M., Vinh, N. T., Tham, N. T., Bay, P. V. B., Campbell, J. I., Farrar, J. et al. (2011), 'The emergence of rotavirus g12 and the prevalence of enteric viruses in hospitalized pediatric diarrheal patients in southern vietnam', *The American journal of tropical medicine and hygiene* **85**(4), 768–775.

Pang, X. L., Preiksaitis, J. K. & Lee, B. E. (2014), 'Enhanced enteric virus detection in sporadic gastroenteritis using a multi-target real-time pcr panel: a one-year study', *Journal of medical virology* **86**(9), 1594–1601.

Phan, M. V., Anh, P. H., Cuong, N. V., Munnink, B. B. O., van der Hoek, L., My, P. T., Tri, T. N., Bryant, J. E., Baker, S., Thwaites, G. et al. (2016), 'Unbiased whole-genome deep sequencing of human and porcine stool samples reveals circulation of multiple groups of rotaviruses and a putative zoonotic infection', *Virus evolution* **2**(2).

Rabaa, M. A., Tue, N. T., Phuc, T. M., Carrique-Mas, J., Saylors, K., Cotten, M., Bryant, J. E., Nghia, H. D. T., Van Cuong, N., Pham, H. A. et al. (2015), 'The vietnam initiative on zoonotic infections (vizions): a strategic approach to studying emerging zoonotic infectious diseases', *Ecohealth* **12**(4), 726–735.

Spiegelhalter, D. J., Abrams, K. R. & Myles, J. P. (2004), *Bayesian approaches to clinical trials and health-care evaluation*, Vol. 13, John Wiley & Sons.

Appendices

A An Appendix

Code available under: https://github.com/YifanZeng1874/Dissertaion$_1$.*git*