AMA4951 Capstone Project

Topic: Mechanistic Model for Infecticous Diseases

Title: Unexpected positive correlation between human development index and risk of infections and deaths of COVID-19 in Italy

Kai Liu1, Mu He2, Zian Zhuang3, Daihai He3 and Huaichen Li4,5

1 Department of Statistics and Mathematics, Shanghai Lixin University of Accounting and Finance.

2 Department of Mathematics and Statistics, McMaster University, Canada.

3 Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong SAR China

4 Department of Respiratory and Critical Care Medicine, Shandong Provincial Hospital Affiliated to Shandong University, Jinan, 250021, China. lihuaichen@163.com.

5 School of Medicine and Life Sciences, University of Jinan-Shandong Academy of Medical Sciences, Jinan, 250062, Shandong, China. lihuaichen@163.com.

Email of authors

Kai Liu [liuk0809@lixin.edu.cn](mailto:liuk0809@lixin.edu.cn)

Mu He [hem11@math.mcmaster.ca](mailto:hem11@math.mcmaster.ca)

Zian Zhuang [larry.zhuang@polyu.edu.hk](mailto:larry.zhuang@polyu.edu.hk)

Daihai He [daihai.he@polyu.edu.hk](mailto:daihai.he@polyu.edu.hk)

huaichen Li lihuaichen@163.com

Correspondence to

Daihai He, Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong SAR China, [daihai.he@polyu.edu.hk](mailto:daihai.he@polyu.edu.hk)

# ABSTRACT

The Covid-19 virus is a global pandemic virus that is caused by the virus called severe acute respiratory syndrome coronavirus 2 or (SARS-CoV-2). The virus was first reported in the Wuhan District of China and since then the various has spread to almost every corner of the globe due its unique mode of transmission. It is a known fact that flue like diseases spread like wildfire.

Over the centuries that mankind has walked the earth several pandemics have threatened his existence such as the Prehistoric Epidemic of Circa (300 B.C), the Spanish Influenza and even HIV and AIDs disease pandemic. But none have raged chaos such as the Covid-19 pandemic that has devastated and looted the world of its peace from and to all ends. Peoples livelihoods were disrupted, many were infected and died and even some people were deported back to their home countries.

Much of research was conducted into the causes and effects of the disease. An intrigue was taken into the dynamic causal of the disease as many were unsure as to whether the disease propagated itself or mostly the propagation was at relationship with other factors such as the human development index.

The aim of this research paper is to study the correlation between human development index and the risks of infections and death in Italy through a mechanistic model that is grounded on raw data that serves to prove the significance of the assumed hypothesis.

# 1.CHAPTER 1

## Introduction

In this analysis, I am going to observe that how human development index, including age, education and life expectation, correlates with number of cases and death rate among the population of COVID-19, especially in Hong Kong. After the analysis, it has shown that HDI is positively correlated with annual income and chronical ill, and is negatively correlated with smoking habit. I will explain why these factors can be observed between human development index and infection risks and mortality of COVID-19 in Hong Kong.

## Historical Background

The dynamic of infectious disease epidemics is divided into individuals with differing disease status. Starting from December 2019, there was a new virus called SARS-CoV-2. It first appeared in China and expanded globally quickly. Many patients infected with COVID-19 have suffered from high fever, a cough and being weird on sense of smell, which were classified as mild symptoms, while others might not have any symptoms. However, some would evolve into more severe and fatal symptoms affecting lungs, heart and brain.

COVID-19 used ACE2, which is a human enzyme, to submerge the cells of its host. ACE2 decreases blood pressure in human body and has opposite function with another enzyme called ACE. Thus, human body has to balance ACE and ACE2 level to ensure a normal standard of blood pressure. And it is not sure how COVID-19 affects ACE and ACE2 work.

When COVID-19 first discovered in China, their researchers collected cells and fluid from sufferers’ lungs, in order to identify the virus. The researchers analyzed the data collected former work to study if changes in body regulation of blood pressure would cause severe symptoms of COVID-19. They found that the ACE level would decrease and ACE2 level in lung cells would increase due to COVID-19. This will lead to rise in molecule level called bradykinin in cells, which will prompt pain and cause expansion of blood vessels while this will cause swelling and agnail of surrounding cell. Moreover, it was found that it would produce a substance called hyaluronic acid and the enzymes that could demote it greatly decreased. Hyaluronic acid can form a hydrogel from the absorption of more than 1,000 times its own weight in water. This will lead to fluid leakage into the lungs combined with the excess hyaluronic acid. This would result in the creation of Jelly-like substance to prevent COVID-19 patients suffering from severe symptoms from oxygen uptake and carbon dioxide release in the lungs. Therefore, the findings suggested that this will lead to more severe symptoms of COVID-19.

The COVID-19 has exploded quickly in Hong Kong since January 2020 while the pestilence in China was then controlled. Obvious differences in contagious diffusion and fatality rate due to COVID-19 across countries appear, however, factors and reasons about the differences are not clear, and is meaningful to study the relationship between the infectious disease and human development level and other factors.

It is assumed that people are independent in the same country or region, with same probability of infection and confirmation. A multiple logistic regression model was performed between human development index and confirmed cases. It is found that HDI is significantly large, p-value is less than 2 \* 10-16. If the HDI rose by 1, the number of a confirmed case would increase by exp(2.8648).

Death rate is defined as the scale of death among confirmed cases. Not all the patients of COVID-19 are counted into and diagnosed as confirmed cases. It is assumed the people in the same country or region have same probability to suffer from infection or death because of COVID-19 but the death risk is different in different regions. The regression model is to observe how the performance of relation between death and HDI is. And it is found that a rise in 0.1 HDI will increase 39 deaths.

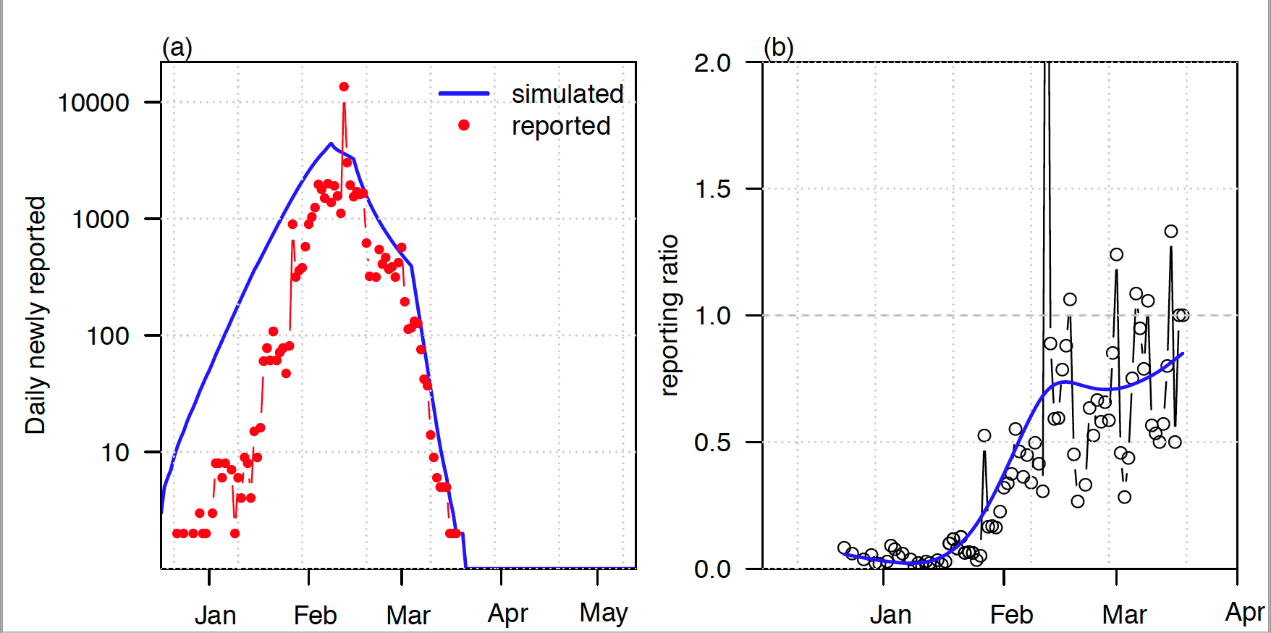
It is found that the higher the human development index in a region, the higher the COVID-19 case rate and death rate. HDI is a statistic index of life expectancy, education, living standard and gross income per capita.

Number of persons suffering from at least 1 chronic diseases per 50 persons and number of cigarettes per 50 persons with smokers aged at least 10 years are used to calculate life expectancy, since they may have health issues. And the median of life expectancy is 41.The gross income of a region is to measure a region’s living standard. And it median is USD27962. As a result, it shows that the higher the HDI in a region, the higher case rate and death rate of COVID-19.

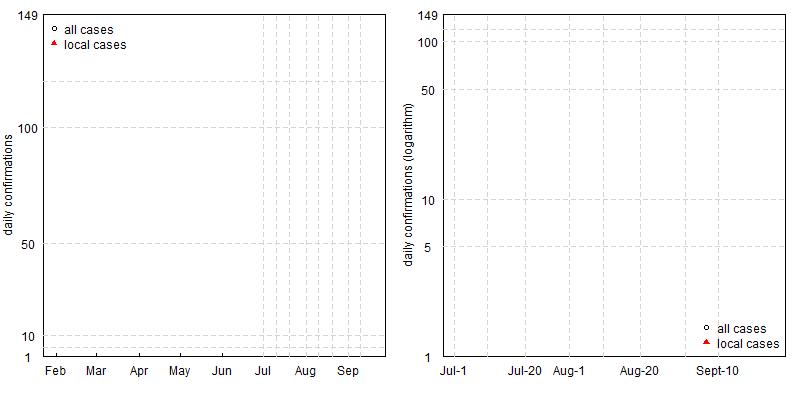
Simple regression model is to measure the effect HDI on case rate and fatal rate of COVID-19 directly. The results can be generated by standard errors.

Multiple regression model is to study the further effect of HDI on case rate and fatal rate after other factors’ adjustments. While more people consume cigarettes and suffer from chronic diseases, they will positively be correlated with case rate and fatal rate. And these factors are more significant than measuring gross income of a region. When HDI increases by 0.1, the number of confirmed cases will increase by 415.72 and the death rate will increase by more than 10000.

In conclusion, the higher the HDI means the case rate and death rate of COVID-19 are higher. Although higher HDI means improvement of living standards and better life expectancy with high gross income, it is found that there will be more persons suffering from more than 1 chronic diseases and smokers in a region.



Trend Graphs



# CHAPTER 2 THEORETICAL BACKGROUND

## 2. Related work

Mathematical modeling concedes rapid computation and estimation of pandemic outbreaks and plays a valuable role in decision making. Simulation techniques, on the other hand, are used when the data collection involves a large number of conditions to test, which leads to increased cost ([Siettos & Russo, 2013](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib52)). Several mathematical and statistical models have been derived recently such as the Multivariate linear regression ([Thomson et al., 2006](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib54)), time series models ([Kurbalija et al., 2014](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib31)), grey forecasting models ([Wang et al., 2018a](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib58); [Zhang et al., 2017](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib66)), back-propagation neural networks ([Liu et al., 2019](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib33); [Ren et al., 2013](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib45); [Zhang et al., 2013](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib65)), and simulation models ([Nsoesie et al., 2013](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib37); [Orbann et al., 2017](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib38)). The spread of an epidemic is unpredictable and random. Due to this reason, it becomes difficult to build mathematical models to analyze epidemic randomness. [Hethcote (1989](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib24), pp. 119–144) identifies three basic types of deterministic models – SIS endemic, SIR epidemic, SIR endemic for the mathematical modeling, and predicting the spread of infectious disease. Theorems that consist of “reproduction number R0, contact number σ and replacement number R″ are presented for mathematical models like SEIR and MSEIRS ([Hethcote, 1989](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib24), pp. 119–144). Compared with statistics methods, mathematical modeling based on dynamical equations receive relatively less attention ([Adam et al., 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib1)), though they can provide a more detailed mechanism for the epidemic dynamics.

The classical susceptible exposed infectious recovered model (SEIR) is one of the most widely adopted methods for characterizing the epidemic of COVID-19 outbreak in both China and other countries ([Hethcote, 1989](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib24), pp. 119–144). The SEIR model replicates the “time-history” of any epidemic or pandemic outbreak, and it presents the model of dynamic interaction between people with four different health conditions or phases of the pandemic, namely the susceptible (S), exposed (E), infective (I), and recovered (R). SEIRD model contains the 4 basic containers in the SEIR model: **S**usceptible, **E**xposed, **I**nfective, **R**ecovered, along with an added container - **D**ead. A “Formal Characterization and Model Comparison Validation” based on the SEIRD model which uses the data from Korea and Spain is proposed by Casas et al. ([Fonseca i Casas et al., 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib15)). The proposed model showed the predicted parameterization with empirical evidence and a decision support system (DSS) is implemented to study the nature of the pandemic in Catalonia ([Fonseca i Casas et al., 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032#bib15)).

A data-driven model to predict the spread of Covid-19 for an upcoming week using the SEIRD model is studied and tested for datasets obtained from Italy, India, and Russia ([Rapolu et al., 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib44)). The proposed model ([Rapolu et al., 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib44)) produces results in which the parameters are calculated from the data, to plan for the future requirement of PPEs for hospital staff and healthcare devices. Contrarily, the transmission dynamics of Covid-19 were evaluated based on a SEIRD compartmental modeling approach by Mukaddes et al. ([Mukaddes et al., 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib35)). This model was based on the “kinematic parameters” that describe the transmission, recovery, and death rate in Bangladesh. The study also highlights the dynamic factors and two parameters that refer to infection which derives the reproduction number R0. This study is groundbreaking for research work carried out in developing nations to reopen the businesses and to boost the economy back to the pre-covid period. However, external influences such as weather, herd immunity were not considered as a part of the study ([Mukaddes et al., 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib35)).

The COVID-19 pandemic is very dynamic and spreads rapidly, and hence there is a need to create robust modeling solutions to curb the outbreak. A forced SEIRD model with two different infection rate functions of the Covid-19 spread in Italy was investigated by Piccolomini et al. ([LoliPiccolomini & Zama, 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib34)), in which the “integration time was distributed into sub-intervals” to estimate the model parameters. The study was based on data collected from two regions in Italy, Lombardia and Emilia-Romagna. This model ([LoliPiccolomini & Zama, 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib34)) will be efficient to make predictions about different stages of the epidemic outbreak across various regions in Italy and Europe. Another popular and widely used statistical method for time-series forecasting is the **A**utomatic **R**egressive **I**ntegrated **M**oving **A**verage (ARIMA) model, which studies the series of temporal structures in time series data. Earlier study on disease management techniques with time series using ARIMA models is proposed by [Sato (2013)](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib47). The author [Sato (2013)](https://www.sciencedirect.com/science/article/pii/S2468042720301032#bib47) emphasizes the fact that the options to follow up and spot the difference in data patterns should be given importance in healthcare practices.

Forecasting a disease is essential for the healthcare department and policymakers to strengthen their vigilance and reallocate their resources. ARIMA time series model is a widely accepted method for the pandemic forecasting because of its simplicity and systematic structure ([Ceylan, 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib8); [Wang et al., 2018b](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib59)). The extent of the outbreak of the Covid-19 pandemic in Italy, Spain, and France was examined with the ARIMA model ([Ceylan, 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib8)). The proposed model consists of 4 steps in modeling which include “assessment, prediction of parameters, characteristic checking, and forecasting”. The outcome of the study can guide policymakers and healthcare authorities in the European nations to effectively allocate resources and plan for the future flare-up of the current situation ([Ceylan, 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib8)).

Alzahrani et al. ([Alzahrani et al., 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib3)) studied the spread of pandemic outbreak using the ARIMA prediction model in Saudi Arabia. The authors have used the “linear parametric model prediction approach”, in which the parameters of the ARIMA model were chosen based on the value of the “Akaike information criterion” ([Akaike, 1974](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib2); [Alzahrani et al., 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib3)). The dataset was divided into training and testing datasets and four statistical models were employed to predict the spread of Covid-19, by comparing the performance of each model with the evaluation metrics from which good fit is derived ([Alzahrani et al., 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib3)). Other related studies and research work on disease prediction using ARIMA and hybrid ARIMA modeling techniques found in the literature are presented in [Table 1](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "tbl1).

Table 1. Research work found in the literature for disease prediction using ARIMA/hybrid models.

| **Method (s)/Type of Modeling** | **Pandemic/Epidemic/Endemic** | **The research found in the literature** |
| --- | --- | --- |
| ARIMA | Malaria | Gaudart et al., ([Gaudart et al., 2009](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib16)) |
| ARIMA, Artificial Neural Networks (ANN) | HAV | Guan et al., ([Guan et al., 2004](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib20)) |
| ARIMA | SARS | Earnest et al., ([Earnest et al., 2005](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib13)) |
| ARIMA, Seasonal Autoregressive Integrated Moving Average (SARIMA) | Influenza | He et al., ([He & Tao, 2018](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib23)); Chen et al., ([Chen et al., 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib9)) |
| Multivariate Poisson Regression (MPR), ARIMA, and ANN | Dengue Fever | Polwiang ([Polwiang, 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib41)) |
| Random Forest (RF), ARIMA/X Models | Infectious Diarrhea | Fang et al., ([Fang et al., 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib14)) |
| Elman Recurrent Neural Networks (ERNN), ARIMA, and Jordan Neural Networks (JNN) | Brucellosis | Wu et al., ([Wu et al., 2019](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib60)) |
| ARIMA (or) SARIMA-NAR (Nonlinear Autoregressive Network) (or) hybrid model | Covid −19 | Ceylan ([Ceylan, 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib8)); Alzahrani et al. ([Alzahrani et al., 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib3)); Perone ([Perone, 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib40)); Kumar et al. ([Kumar et al., 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib30)); Sato ([Sato, 2013](https://www.sciencedirect.com/science/article/pii/S2468042720301032#bib47)); Wang et al. ([Wang et al., 2018b](https://www.sciencedirect.com/science/article/pii/S2468042720301032#bib59)); Benvenuto et al. ([Benvenuto et al., 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib6)); Hernandez-Matamoros et al. ([Hernandez-Matamoros et al., 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib22)); Kufel ([Kufel, 2020](https://www.sciencedirect.com/science/article/pii/S2468042720301032" \l "bib29)) |

This paper aims to build a hybrid model that will allow us to estimate the short and long-term dynamics of COVID disease and build confidence intervals of predictions by using a hybrid dynamic model based on SEIRD with ARIMA corrections. This model will help officials to be prepared for the waves of pandemic and reserve hospital beds in advance.

# CHAPTER 3 RESEARCH METHODLOGY

Code 1: Covid.R

x<-read.csv('covid\_hk.csv',as.is=T)

x[,2]<-as.Date(x[,2])

x$week<-round(as.numeric(x[,2]-x[1,2])/7)+1

i<-grep('Asymptomatic',x[,3])

y<-x[i,]

head(y)

i1<-table(y$week)

i2<-table(x$week)

months<-as.numeric(as.Date(paste(2020,2:12,1,sep="-"))-as.Date('2020-1-23'))/7\*1.24

i<-which(x$Case %in% c('I'))

z<-x[i,]

i3<-table(z$week)

wk<-sort(as.numeric(unique(c(names(i1),names(i2),names(i3)))))

wk1<-matrix(0,ncol=3,nrow=length(wk))

row.names(wk1)<-wk

wk1[match(names(i1),wk),1]<-i1

wk1[match(names(i2),wk),2]<-i2

wk1[match(names(i3),wk),3]<-i3

p<-wk1[,1]/wk1[,2]

se<-1.96\*sqrt(p\*(1-p))/sqrt(wk1[,2])

jpeg('hk\_asymp.jpeg',width=800,height=600)

par(las=1,mar=c(3,5,1,5))

barplot(wk1[,2],col='lightgrey',xaxt='n',ylim=c(0,max(wk1[,2])+20))

par(new=T)

barplot(wk1[,3],ylim=par('usr')[3:4],col='white',ann=F,axe=F,axisnames=F,xaxt='n')

legend("topleft",legend=c('Imported','Local'),pch=22:22,pt.bg=c('white','lightgrey'),bty='n')

axis(1,at=months,lab=month.abb[2:12],padj=-1)

par(new=T)

plot(as.numeric(names(p)),p,xlim=c(0.5,length(wk)+0.4),ann=F,axe=F,frame=T,ylim=c(0,1),col='red')

axis(4)

mtext(side=2,line=3,'Weekly confirmations',las=0)

mtext(side=4,line=3,'Asymptomatic ratio',las=0,col='red')

#mtext(side=1,line=2,'week since Jan 23')

for(i in 1:length(p))

lines(c(i,i),p[i]+c(-1,1)\*se[i],col='red')

smoothingSpline = smooth.spline(1:length(p),p, spar=0.8)

lines(smoothingSpline)

dev.off()

for(i in c(4,7)) x[,i]<-as.factor(x[,i])

x$asym<-0

x$asym[which(x$'Date of onset'=='Asymptomatic')]<-1

x$Date<-as.numeric(as.Date(x$'Report.date')-as.Date('2020-1-22'))

jpeg('trend.jpeg',width=800,height=400)

at<-as.numeric(as.Date(c('2020-6-20','2020-7-1','2020-7-10','2020-7-20','2020-8-1','2020-8-10','2020-8-20','2020-9-1','2020-9-10'))-as.Date('2020-1-22'))

at1<-as.numeric(as.Date(paste(2020,2:9,1,sep="-"))-as.Date('2020-1-22'))

x1<-x[x$Case=='L',]

y1<-table(x1$Date)

y<-table(x$Date)

par(las=1,mar=c(3,4,1,1),xaxs='i',yaxs='i')

plot2<-function(x){

plot(as.numeric(names(y)),y,log=x,type='b',xlim=c(ifelse(x=="",0,158),250),ann=F,axe=F,frame=T)

if(x=="y")

axis(1,tck=0.01,padj=-0.8,at,lab=c('Jun-20','Jul-1','Jul-10','Jul-20','Aug-1','Aug-10','Aug-20','Sept-1','Sept-10'))

if(x=="")

axis(1,tck=0.01,padj=-0.8,at1,lab=month.abb[2:9])

axis(2,tck=0.01,hadj=0.5,at=c(1,5,10,50,100,149))

abline(h=c(5,10,50,100,120),lty=2,col='lightgrey')

abline(v=at[-1],lty=2,col='lightgrey')

if(x=="")

mtext(side=2,line=2,las=0,'daily confirmations')

if(x=="y")

mtext(side=2,line=2,las=0,'daily confirmations (logarithm)')

points(as.numeric(names(y1)),y1,pch=17,type='b',col='red')

legend(ifelse(x=="","topleft","bottomright"),pch=c(21,17),col=c('black','red'),legend=c('all cases','local cases'),bty='n')

#mtext(side=1,line=-1,adj=0.8,Sys.Date())

}

par(mfrow=c(1,2),mar=c(3,3,1,1))

plot2("")

plot2("y")

dev.off()

Code 2 :Report.R

require('readxl')

load('flunet\_1995-2020.rdata')

month<-as.Date(paste(2020,2:9,1,sep="-"))

country<-unique(dat[,1])

x<-read.csv('WHO-COVID-19-global-data-2.csv')

country<-unique(x[,3])

names(x)[1] <-'time'

x$time<-as.Date(x$time)

cmp<-function(c){

y<-x[which(c==x[,3]),]

y$confirm<- diff(c(0,y$Cumulative\_cases))

y$dead<- diff(c(0,y$Cumulative\_deaths))

print(head(y))

if(c=='The United Kingdom')c<-'United Kingdom of Great Britain and Northern Ireland'

i<-which(c==dat[,1])

if(c=='United Kingdom of Great Britain and Northern Ireland')c<-'United Kingdom'

uk<-dat[i,]

print(head(uk))

print(unique(uk[,1]))

uk<-uk[sort(as.numeric(as.Date(uk$End.date)),index=T)$ix,]

sk<-y

par(las=1)

matplot(as.Date(sk$time),cbind(sk$confirm,sk$dead),type='b',log='y',pch=17:18,ylab='daily (weekly) confirmtions',axe=F)

for(year in 0:5){

uk1<-uk[grep(2015+year,uk$End.date),]

lines(as.Date(uk1$End.date)+365\*(5-year),uk1$Total.number.of.influenza.positive,col='blue',pch=19,lty=1,lwd=ifelse(year==5,3,1))

}

legend("topright",bty='n',legend=c('COVID-19 cases','COVID-19 deaths','Flu positive pre 5 years'),col=c(1:2,'blue'),pch=c(17:18,NA),lty=1)

axis(2)

axis(1,at=month,lab=month.abb[2:9])

mtext(side=3,c)

}

tiff('Figure1.tiff',,width=10,height=8,res=200,unit='in')

par(las=1,mar=c(3,4,1,1),mfrow=c(2,2))

cmp('Italy')

cmp('Spain')

cmp('Japan')

cmp('Republic of Korea')

dev.off()

tiff('FigureS1.tiff',width=10,height=8,res=200,unit='in')

par(las=1,mar=c(3,4,1,1),mfrow=c(2,2))

cmp('France')

cmp('The United Kingdom')

#cmp('South Africa')

cmp('United States of America')

cmp('Canada')

dev.off()

tiff('FigureS2.tiff',width=10,height=8,res=200,unit='in')

par(las=1,mar=c(3,4,1,1),mfrow=c(2,2))

cmp('Sweden')

cmp('Netherlands')

cmp('Brazil')

cmp('Iran (Islamic Republic of)')

dev.off()

# CHAPTER 4 EMPIRICAL DATA ANALYSIS

# CHAPTER 5 RESULTS

Highlights:

* Human development index is positively correlated infection and death rates of COVID-19;
* 0.1 increase in human development index results in 17.5448 and 39.6230 exponential increase in odds of infections and deaths, respectively
* Regions with smaller human development index consumes more cigarettes, but have lower average annual gross salaries and smaller chronic diseases population
* The positive correlation between human development index and infections and deaths rates is partially explained by the smoking habits, health status and income level.

Abstract (95 words)

In this analysis, we observed that human development index (an integrated index of life expectation, education and living standard) correlates with case rate (proportion of confirmed cases among the population) and the fatality rates of COVID-19 in Italy based on data as of May 15, 2020. Further analysis showed that HDI is negatively correlated with cigarette consumption, whereas it is positively correlated with chronic disease and average annual gross salary. These factors may partially explain why unexpected positive correlation is observed between human development index and risk of infections and deaths of COVID-19 in Italy.

Main text: 1077 words

To the editor:

The coronavirus disease 2019 (COVID-19) broke out quickly in Italy since March 2020 when the epidemic got controlled in China. Reasons of rapid breakout and overall case-fatality rate in Italy have been studied and reported in literature [1, 2, 3]. Obvious differences in epidemic spread and fatality rates among regions exist, but factors related these spatial differences are unclear. It is of interest to study this regional heterogeneity and the related factors.

Global data of COVID-19 have been integrated by researchers and available publicly from R package nCov2019 [4]. We downloaded and extracted the data of Italy by regions for our study. As of May 15, 2020, Lombardy ranks top 1 with 83820 cumulative confirmed cases among the 20 regions, while the number of cumulative confirmed cases in Basilicata is the smallest (389 cases). The number of death ranges from 22 to 15296, corresponding to regions of Molise and Lombardy, respectively. Demographical data including population, area, population density and human development index (HDI) by regions of Italy 2019 were downloaded from <https://en.wikipedia.org/wiki/Regions_of_Italy>. The case rates (the proportion of confirmed cases among regional population) range from 0.0006 to 0.009 with a median of 0.0025, while the death rate (proportion of deaths among regional population) ranges from 0.00005 to 0.00152 with a median of 0.00026. HDI [5] is an integrated index of healthy long life, education and living standard, measured by life expectancy, expected/mean years of schooling, Gross National Income per capita, respectively. The median HDI is 0.891 with a range from 0.845 to 0.919. Figure 1 presents the log odds of case rates and death rates as of May 15, 2020 against HDI by regions. It presents a linear pattern between the log odds and HDI. To quantify the association between cases rates and death rates with HDI, we perform univariate logistic regression.

It is reasonable to assume people in the same region are independent and identical with the same probability of being infected and confirmed. Under this assumption, we performed a univariate logistic regression between the cumulative confirmed cases and HDI. We found that HDI is statistically significant (log odds = 28.6476, p-value <2\*10-16). If HDI increases by 0.1, the odd of a confirmed case (that is, the probability that a person is a confirmed case against the probability that a person is not a confirmed infected) increase exponentially by exp(2.8648)=17.5448.

Many literatures have studied the case-fatality rate. Case-fatality rate is defined as the proportion of death among the confirmed cases. However, not all infected people are diagnosed and counted into the confirmed cases. We assume people in the same region has the same probability to get infected and die due to COVID-19 while the risk of death is different among different regions. A univariate logistic regression to study the relation between the cumulative death and HDI is also performed. HDI is again significant (log odds = 36.7946, p-value < 2\*10-16). An increase of 0.1 in HDI associates with an increase of 39.6230 in odds of death.

it is interesting to note that high HDI is associated with high case rate and high fatality rate. HDI is an integrated index of healthy life, education and living standard. To further explore how each of the components of HDI associates with case rates and fatality rates, we downloaded health data, including smoking data in 2019 and chronic disease data in 2018 from [http://dati.istat.it/?lang=en#](http://dati.istat.it/?lang=en). Average annual gross salaries by regions in 2019 is also downloaded from <https://www.statista.com/statistics/708972/average-annual-nominal-wages-of-employees-italy-by-region/>. Number of cigarettes per day per 100 persons with the same characteristics (cigarette smokers aged 14 years and over) and number of persons with at least one chronic disease per 100 people are used as surrogate index for healthy life for two reasons, one is they are public and the other reason is they are associated with life expectancy. The median of number of cigarettes per day per 100 persons is 11.10 (range from 9.2 to 12.5). The number of persons with at least oone chronic disease per 100 people ranges from 32.7 to 47.8 with a median of 41. Average annual gross salary by regions is used to present living standard. Its median is 27962 with a range from 24308 to 31446.

Initial graphical and quantitative analysis of relations between HDI and smoking data, chronic disease data and average annual salary are performed. It shows that HDI is negatively correlated with smoking data (correlation=-0.6428, p-value=0.0022), is positively correlated with chronic disease but not statistically significant (correlation=0.3275, p-value=0.1587), is positively correlated with average annual gross salary (correlation=0.6521, p-value=0.0018).

Simple logistic regressions are performed to study the direct effect of the three factors on case rate and death rate. The results are summarized in Table 1 in terms of log odds estimates and standard errors. It turns out all the factors are significantly associated with case rates and death rates.

Multiple logistic regression is further performed to investigate the effect of HDI after adjustment of other factors. The estimates of log odds are presented in Table 1. The effect of HDI decreased but remain positive. All the factors are statistically significant. The results are consistent with our knowledge that regions with more cigarettes consumption rates and more persons with chronic disease is positively associated with case rates and mortality rates. However, regions with higher average annual gross salary are also positively correlated with case rates and mortality rates, but with small magnitude. More specifically, assume other factors remain same, 0.1 increase in HDI results in 6.03 exponential increase in confirmation odds (p<0.001), and 9.78 exponential increase in death odds. Similarly, 1000 increase in average annual gross salary, 10 increase in number of cigarettes per day per 100 persons and 10% increase in the number of persons with at least one chronic disease results in 1.34, 1.35, 1.72 exponential increase in odds of confirmed case, respectively. The interpretation is similar for odds of confirmed case.

In summary, though high HDI means longer life expectation, better education and better living standard, it is surprising to note that it associates with higher case rates and mortality rate. By further study, we observe that regions with high HDI normally has higher number of persons with more than one chronic disease, less cigarette consumption and higher average annual gross salary. Multiple logistic regression analysis shows that these three factors take some effects of HDI on the case rates and death rates. This may partially explain why unexpected positive effect of HDI on case rate and mortality rate

# CHAPTER 6: CONCLUSIONS AND DISCUSSIONS

## 5. Conclusion

The proposed hybrid model consists of a dynamic SEIRD model with vital dynamics and decaying COVID mortality rate and three ARIMA models that cancel out dynamic model residuals and enhance prediction quality. Unlike pure dynamics models like SIR, SEIR, and SEIRD, this model allows us to make precise predictions for up to 2 months ahead.

The model was tested on US COVID statistic data. Obtained validation results allow us to conclude that the proposed hybrid model has good prediction ability and decent performance. Obtained long-term predictions reflect the general dynamic of the outbreak and are especially useful for the healthcare system workers and government officials. Obtained short-term predictions allows us not only to forecast the future number of infected, recovered, and deceased patients, but only estimate forecast error under adverse or optimistic circumstances. The proposed method can be used as an effective tool for prediction and analysis of the dynamics of the COVID-19 pandemic.

Here are some perspective ways of further development of the proposed method:

(a)

Parameter estimation with different algorithms and boundaries;

(b)

Testing the method of COVID statistics in other countries;

(c)

Develop alternative methods for residue prediction;

Enhancing of the proposed hybrid model depends on profound research results about COVID-19. That is why monitoring recent research in the field and quickly adjusting the model according to the new data is crucial.

## Declaration of competing interest

# REFERENCES

[1] Andrea Remuzzi and Giuseppe Remuzzi. COVID-19 and Italy: what next? *The Lancet.* 2020, 395:1225-28.

[2] Edward Livingston and Karen Bucher. Coronavirus Disease 2019 (COVID-19) in Italy. *JAMA.* 2020, 323(14):1335. https://jamanetwork.com/journals/jama/fullarticle/2763401

[3] Graziano Onder, Giovanni Rezza and Silvio Brusaferro. Case-fatality rate and characteristics of patients Dying in relation to COVID-19 in Italy. 2020, 323(18):1775-1776.https://jamanetwork.com/journals/jama/fullarticle/2763667

[4] Guangchuang YU, <https://github.com/GuangchuangYu/nCov2019>. Accessed on May 18, 2020.

[5] Average annual gross salary in Italy in 2019, by region <https://www.statista.com/statistics/708972/average-annual-nominal-wages-of-employees-italy-by-region/>