5-2014

Early detection and control of potential pandemics.

Shengpeng Jin
University of Louisville

EARLY DETECTION AND CONTROL OF POTENTIAL PANDEMICS

By

Shengpeng Jin

A Dissertation

Submitted to the Faculty of the

J.B. Speed School of Engineering of the University of Louisville

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

Department of Industrial Engineering

University of Louisville

Louisville, Kentucky

May, 2014

EARLY DETECTION AND CONTROL OF POTENTIAL PANDEMICS

By

Shengpeng Jin

A Dissertation Approved on

April 11, 2014

By the following Dissertation Committee:

_____
Dr. Suraj M. Alexander

_____
Dr. Gerald Evans

_____
Dr. William Biles

_____
Dr. Weibin Zeng

!#+ . / 7 ,%$ '% - % . 43

I would like to thank my academic advisor, Professor Suraj Alexander, for his guidance and patience. He gave me this great opportunity to be involved in this interesting and meaningful research. His guidance and inspiration along the way were so helpful and his patience and encouragement always gave me confidence when I was lost. I also would like to thank the other committee members, Dr. Gerald Evans, Dr. William Biles and Dr. Weibin Zeng for their comments and assistance during this period of time.

I wish to thank Yepeng Sun, Yang Liu and other students in our department. Their insightful suggestions and comments helped to improve this research.

I would like to express my gratitude to all RTDSS research team members. Their intelligence and dynamic contributions has inspired my research and motivated my open mind to connect different theories and knowledge from different fields to achieve a goal.

I would also like to thank dear friends. Their friendship makes life much happier and more colorful. I wish to thank my dear great parents for their love and support.

ABSTRACT

EARLY DETECTION AND CONTROL OF POTENTIAL PANDEMICS

SHENGPENG JIN

April 11, 2014

Over the centuries, human beings have been inflicted with a variety of contagious diseases, resulting in tens of millions of respiratory illnesses and deaths worldwide. Early detection of disease spread facilitates timely responses that can greatly reduce its impact on a population. Therefore, this early information is a major public health objective and is crucial for policy makers and public health officials responsible for protecting the public from the spread of contagious diseases.

Current indicators of the spread of contagious outbreaks lag behind its actual spread, leaving no time for a planned response. The studies of Christakis et al. in 2010 have shown that social networks can provide more timely information for prediction. However, the reported social network methods used to monitor disease spread do not consider contact patterns of individuals over space and time, such as during their movement from place to place. In this dissertation we propose a more effective way to

chart the spread of contagious outbreaks, in a spatio-temporal sense, using “contact networks”. This enables more effective control of the spread of contagious outbreaks in their early stages so as to “nip a potential pandemic in the bud.”

In order to enhance the prediction model developed we introduce factors to consider the intensity of exposure to the disease, and the susceptibility of the individual. This would involve the consideration of both space and time factors, since diseases caused by either viruses or bacteria involve some type of contact, either direct (e.g. shaking hands) or through the atmosphere (e.g. coughing or sneezing) between the susceptible and infected individuals.

In this dissertation, we apply data mining methodologies and predictive modeling technologies, such as logistic regression, decision trees and neural networks to estimate the infection risk based on an individual’s demographic information and health status. The information used in the models can be obtained from a wide variety of data sources, including historical medical records from hospitals and clinics. Early information on the presence of a potential disease outbreak can be obtained from "sensors", such as, First Watch and EARS (Early Aberration Response Systems) and "central" individuals in “contact” networks.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

1.1 Background

Over the centuries, human beings have been inflicted with a variety of contagious diseases, including various forms of plague, typhoid fever, cholera, malaria, influenza and AIDS etc. At the beginning of the twentieth century, infectious diseases were the leading cause of death worldwide (Cohen, 2000). During the 20th century, three worldwide outbreaks of influenza occurred in 1918, 1957 and 1968. The pandemic in 1918 caused 40 to 50 million deaths worldwide and more than 500,000 deaths in the United States. The latter two were in an era of modern virology and were therefore most thoroughly characterized. All three outbreaks have been informally identified by their presumed sites of origin as Spanish, Asian, and Hong Kong influenza, respectively (Kilbourne, 2006). In the United States, three diseases – tuberculosis, pneumonia, and diarrheal disease – caused 30% of deaths (CDC 1994). Infectious diseases account for 29 out of the 96 major causes of human morbidity and mortality listed by the World Health Organization and the World Bank (Murray and Lopez, 1996) and 25% of global deaths (over 14 million deaths annually) (WHO 2000).

In 2009, the H1N1 influenza emerged out of Mexico and rapidly spread around the globe. Similarly in 2003 the respiratory illness SARS (Severe Acute Respiratory Syndrome) occurred in the Guangdong province of China and lead to the death of many people all over the world. Therefore, it is extremely important for public health

agencies to understand and to control the spread of infectious diseases and prevent potential pandemics (Dimitrov et al. INFORMS Tutorial 2010). Many infectious diseases are spread through populations via physical contacts among individuals. The patterns of these contacts tend to be highly heterogeneous. The traditional mathematical model used to understand the dynamics of epidemics is the compartmental SIR model, which assumes that the population groups are fully mixed and every individual has an equal chance of spreading the disease to another individual. However this is not the case in the real world (Meyers et al. 2005). Therefore, a more effective way is needed to chart the spread of infectious diseases, to allow for effective control. This dissertation illustrates a framework for accomplishing this objective.

1.2 Problem Statement

It is well established that random immunization requires immunizing a very large fraction of the population in order to arrest diseases that spread through contacts between infected and susceptible individuals (Cohen et al. 2003). Mathematical modeling, such as the compartmental SIR models, has long been utilized for predicting the spread of infectious diseases. These models are useful for defining the levels of resources needed to curtail the spread. However, as stated previously, the current models espoused in the literature do not properly address some important aspects of disease spread even though they have proven to be quite useful in understanding epidemic dynamics.

An appropriate strategy to slow down and control the spread rate of infectious diseases is needed so that there would be enough time for resource accumulation and allocation,

such as vaccination production and distribution. In recent years, the study of social networks and in particular the spread of disease through these networks has attracted considerable attention in the academic community (Newman, 2002). It is well known that individuals near the center of a social network are likely to get infected sooner during the course of an outbreak than those at the periphery on average (Christakis et al. 2010). Unfortunately, it is typically very difficult to map a whole network to identify central individuals who might be monitored for infection (Christley et al. 2005). Therefore, an alternative strategy which does not require ascertainment of the global social network structure has been proposed by Christakis et al, namely, i.e. to simply monitor the friends of randomly selected individuals. Also, the current social network methods used to monitor disease spread seldom takes both the space and time factors and the travel pattern of individuals into account.

1.3 Objective

Since the current social network methods used to monitor disease spread do not consider contact patterns of individuals over space and time, such as during their movement from place to place, in this dissertation we define "contact" networks, and suggest the use of contact network epidemiology to define effective control policies to arrest the spread of a contagious disease. We also propose a model that that utilizes spatio-temporal information for more effective control of disease spread. Additionally, we suggest a system that would provide proper directions to susceptible individuals so that they may reduce the likelihood of contracting the disease.

1.4 Dissertation Organization

Chapter 2 provides a literature review mainly associated with current approaches to disease prediction. Chapter 3 describes the basic methods of disease spread, including the research of Christakis et al on social network (Christakis et al. 2009), Hongbo Yu's research on spatio-temporal representation of travel patterns and interactions of individuals in GIS (Hongbo, 2008), data mining methodologies and proposed methods as well as some extension on the pandemic control. Chapter 4 presents our methodologies to predict the disease spread, such as risk analysis of infection, sensors to aid disease prediction and their relationships with the other parts of a decision support system. Chapter 5 evaluates the experimental results and conducts the analysis on the contact network and spatiotemporal GIS information. Chapter 6 presents the conclusions from this dissertation and Chapter 7 lists some points for the future research.

CHAPTER 2 LITERATURE REVIEW

2.1 Social Network

The careful collection of information from a sample of central individuals within human social networks could be used to mitigate the spread of contagious outbreaks before they happen in the population-at-large (Christakis et al. 2009). The social network itself will be an important conduit for the spread of an outbreak. However, mapping a whole network to identify particular individuals from whom to collect information is impractical, especially for large networks.

However, some other ways might be used to deal with this situation. Intuitively, it could enhance the population-level efficacy of a prophylactic intervention to vaccinate the central individuals in networks (Manhart and Holmes, 2005). Dr. Christakis et al monitored the spread of flu at Harvard College from September 1 to December 31, 2009 to evaluate the effectiveness of using nominated friends of randomly selected students as the central individuals in the social network.

They enrolled a total number of 744 undergraduate students from Harvard College as the randomly selected group and tracked whether they had the flu during that period of time. Another "nominated" group consisted of those who were named as a friend at least once by a member of the random group. And the demographic information, such as, whether they were infected and vaccination status were collected by a

completed brief questionnaire for each subject. The hypothesis is that the set of the nominated friends become infected earlier than the set of randomly chosen individuals (Christakis, 2010).

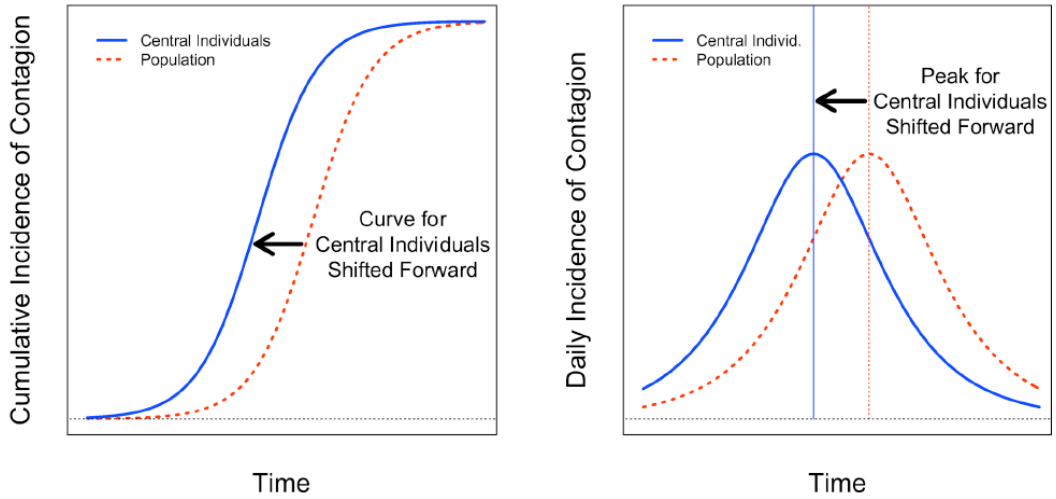


Figure 1. Theoretical differences in contagion between two groups

As hypothesized, the cumulative incidence curves for the friend group and the random group diverge and then converge and the friends curve for flu diagnosed by medical staff is shifted 13.9 days forward in time (95% C.I. 9.9–16.6), thus providing early detection. The friend group showed a significant lead time prior to the estimated peak which could be an effective technique for detecting outbreaks at early stages of an epidemic. See Figure 2 (Christakis, 2010).

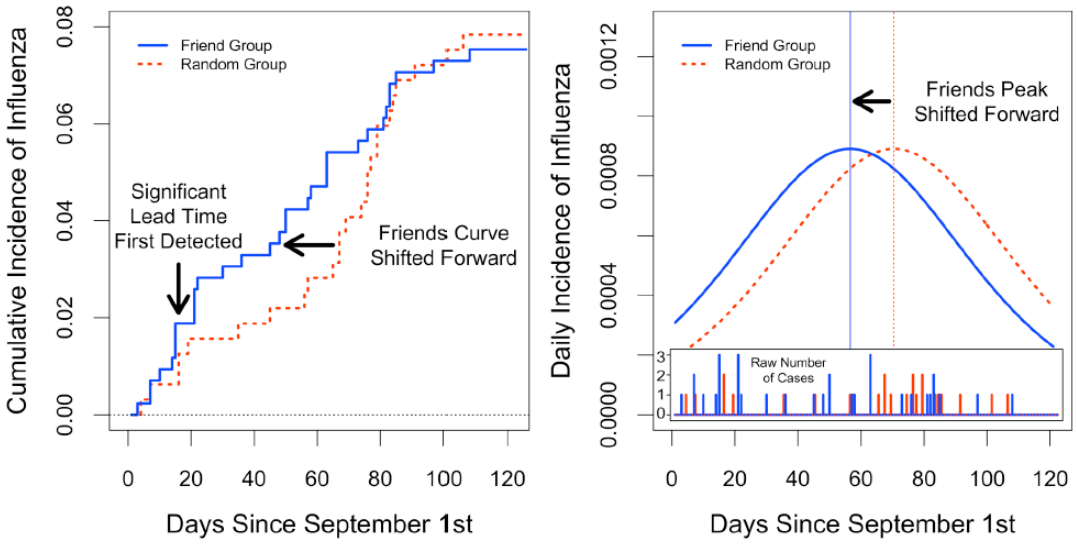


Figure 2. Empirical differences in contagion between two groups

For many contagious diseases, early knowledge of when – or whether – an epidemic is unfolding is crucial to policy makers and public health officials responsible for defined populations, whether small or large. In fact, with respect to flu, models assessing the impact of prophylactic vaccination in a metropolis such as New York City suggest that vaccinating even one-third of the population would save lives and shorten the course of the epidemic, but only if implemented a month earlier than usual (Khazen et al, 2009). Also in case of influenza it takes time to develop the vaccine. In addition, resource planning requires early knowledge of the pandemic spread, and when it is expected to peak, etc.

In fact, this method could be used to monitor targeted populations of any size, in real time. For example, a health service at a university (or other institution) could empanel a sample of subjects who are nominated as friends and who agree to be passively monitored for their health care use.

There are two main steps associated with using the social network epidemiology as an analytical framework to capture the disease transmission. The first step in this modeling approach is to build a realistic network model of contact pattern at an appropriate temporal and spatial scale. The second step is to predict the disease spread through the social network, based on the feature of both the disease and the network structure.

2.2 Spatio-temporal GIS Design

Human activities are performed within a spatial and temporal context (Golledge and Stimson, 1997). GIS has been used for representing human activity data, such as that obtained from travel and diary records for the exploration of their spatio-temporal characteristics (Shaw and Wang, 2000). The individual travel activities with their spatial, temporal and event attributes could be organized by using a path-based representation of trips in a relational GIS environment.

A person's daily activities include physical and virtual activities. Four types of communication modes have been suggested in the literature according to their spatial and temporal requirements (Janelle, 2004) in Table 1.

(1) Conventional face-to-face meetings require participants to be at the same location during the same time period. This communication mode requiring coincidence in both space and time is classified as Synchronous Presence (SP).

(2) Post-it notes or bulletin boards must have people visit the same location, but these visits can be at different times, to complete the information exchange. This type of

communications requires coincidence in space, but not in time, is called Asynchronous Presence (AP).

(3) With the use of information and communication technologies (ICT)s, people are no longer required to be present at the same physical location for communications. Synchronous Telepresence (ST) only requires coincidence in time (e.g., two friends at different locations doing instant messages over the Internet).

(4) Finally, Asynchronous Telepresence (AT) is free from coincidence requirements in either space or time. E-mail between people belongs to this type of communications.

This classification system can be used to describe different types of human interactions based on their spatial and temporal requirements. The SP and AP types of human interactions are carried out in physical space and they are also what we are interested in because only physical activities could lead to the disease spread in real world. Therefore, in this dissertation only SP is taken into consideration and AP is the extension of the dissertation.

Spatial / **Temporal**	**Physical presence**	**Telepresence**
Synchronous	**SP** Face to face (F2F)	**ST** Telephone Chat rooms Teleconferencing
Asynchronous	**AP** Post-it or notes Traditional hospital charts	**AT** Mail E-mail Web pages

Table 1. Communication modes based on spatial and temporal constraints

2.3 Space-time Path

Hagerstrand(1970) proposed a theoretical framework to study the constraints that affect an individual's presence in space and time and to portray individual activities in a space-time context, which is known as Time Geography. One fundamental concept is suggested under the time geographic framework to depict the capability of an individual to conduct activities in space and time which is space-time path.

A space-time path is the container of all activities performed by a person, since all activities take place at certain locations and time periods and each of them occupies a portion of the space-time path (Hagerstrand, 1970). It depicts the sequence of an individual's activities at various locations over a time period.

A space-time path offers a proper continuous representation of such a trajectory. Both physical activities and virtual activities performed by individuals leave traces in the physical space and time, which become contents of space-time paths. An individual's trajectory may pass through a location in the 2D space multiple times. When a space-time path is used to store the trajectory, every point on the space-time path possesses unique coordinates of (x, y, t) since a person only can be at a single physical location at any given time (Hongbo, 2008).

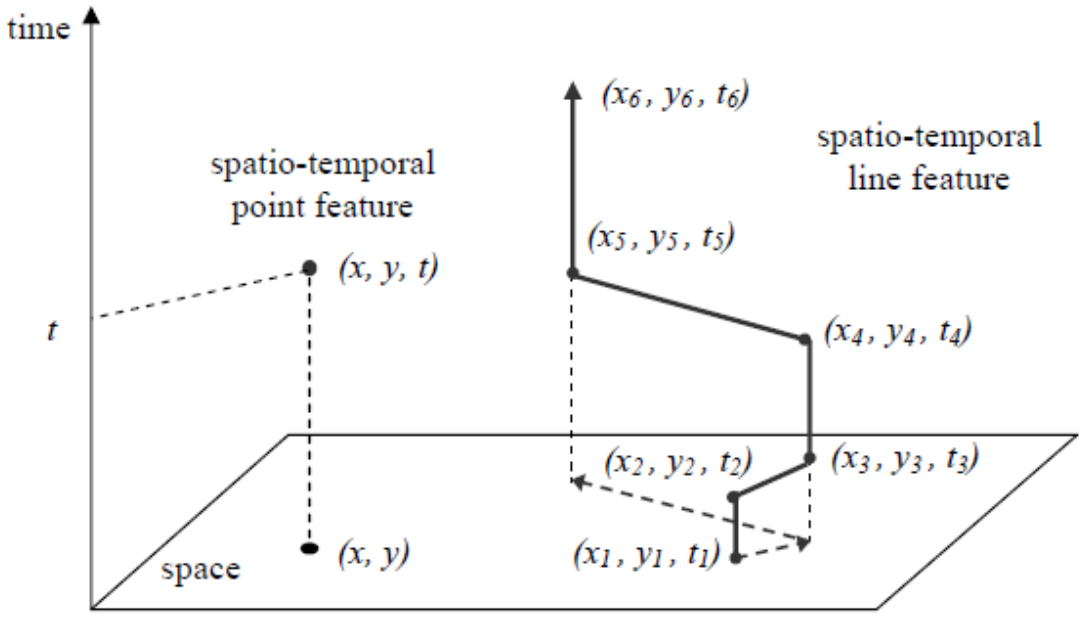


Figure 3. Spatio-temporal features in a 3D GIS framework

Figure 3 shows an example of typical activities of one individual.

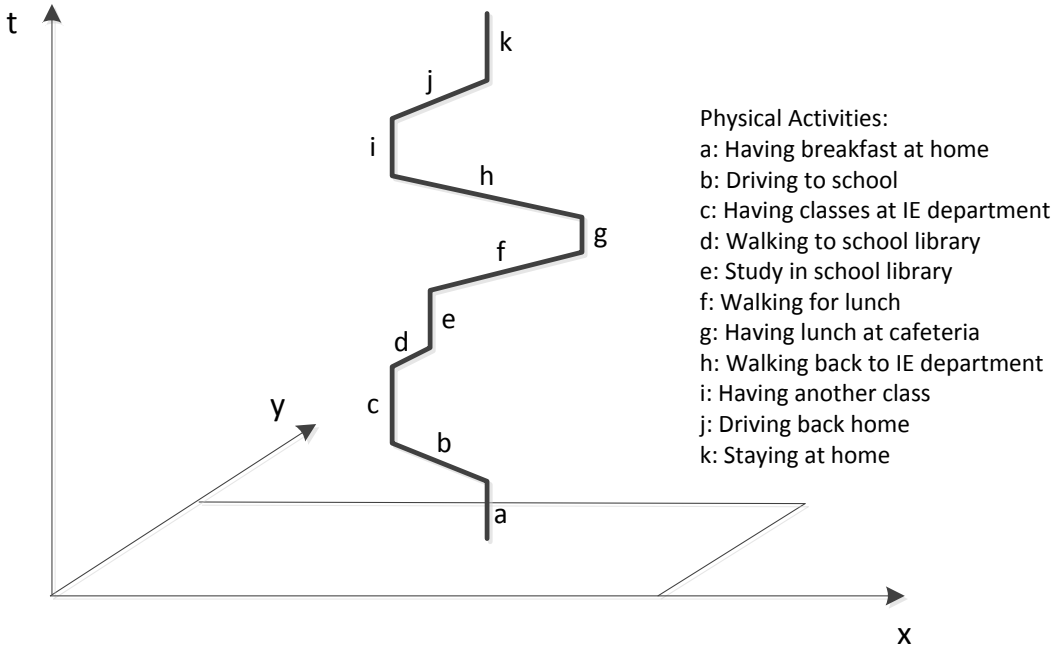


Figure 4. Locate individual activities on a space-time path

There exist three basic relationships of space-time paths between different individuals. (Hongbo, 2008) See Figure 4.

(1) Co-location in time represents activities in different space – time paths that interact with each other within a common time window.

(2) Co-location in space occurs when activities in different space – time paths occupy the same location in different time windows.

(3) Co-existence describes the cases when activities take place at the same location and within a common time window.

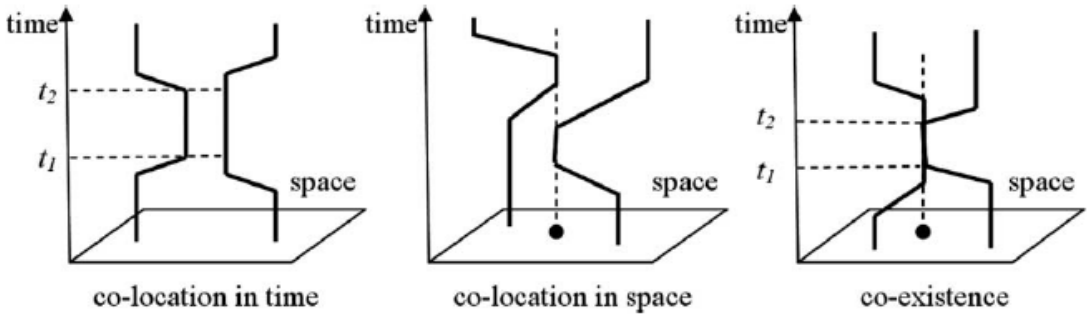


Figure 5. Space-time path relationships

Figure 5 illustrates the difference between these three basic relationships. The typical example for co-location in time is a phone call, instant messaging or a remote video meeting, while the use of a specific desk in an office, is an example of co-location in space. The third relationship, which is co-existence, is the one we will discuss in this paper, since it is the main factor that effects the spread of disease; a face to face talk might be a typical example. However, the audiences in a cinema, passengers on a plane, or the guests in a restaurant could also be classified as being in a co-existence relationship.

2.4 Transmission and Prevention of Infectious Disease

As is known to all, infectious disease could be transmitted easily from one individual to another. Pathogens can be spread by many methods other than direct contact, including through water, food, air, blood and so on. For instance, any time a person with an infection coughs or sneezes may be transmitting illness. And defining the means of transmission plays an important role in understanding the biology of an infectious agent and in addressing the disease it causes.

Transmission may occur through several different sources. Respiratory-borne diseases like influenza, tuberculosis, meningococcal meningitis and SARS, spread through the exchange of respiratory droplets between people in close physical proximity to each other. Sexually transmitted diseases like HIV, genital herpes, and syphilis spread through intimate sexual contact. Gastrointestinal diseases are often acquired by ingesting contaminated food and water. Some infectious agents may be spread as a result of contact with a contaminated, inanimate object (known as a fomite), such as a coin passed from one person to another, while other diseases penetrate the skin directly (Ryan and Ray, 2004). Explicit models of the patterns of contact among individuals in a community, contact network models, provide a powerful approach for predicting and controlling the spread of such infectious diseases (Longini, 1988; Sattenspiel and Simon, 1988; Morris, 1995; Kretzschmar et al., 1996; Ball et al., 1997; Morris and Kretzschmar, 1997; Ferguson and Garnett, 2000; Hethcote, 2000; Lloyd and May, 2001; Newman, 2002; Sander et al., 2002; Keeling et al., 2003; Meyers et al., 2003; Meyers et al., 2005).

Generally, there are several ways to prevent the spread of infectious disease. One of the ways to prevent or slow down the transmission of infectious diseases is to recognize the different characteristics of various diseases (Watts, 2003). Some critical disease characteristics that should be evaluated include virulence, season, age and gender of the susceptible, distance traveled by individuals, and levels of contagiousness. Another effective way to decrease the transmission rate of infectious diseases is to recognize the effects of small-world networks (Watts, 2003). And in epidemics, there are often extensive interactions within groups of infected individuals and other interactions within susceptible individuals. In this dissertation, we will only take the respiratory-borne diseases into consideration and do the risk analysis of getting infected for a susceptible individual by using both contact network model and spatio-temporal information.

2.5 Compartmental SIR Model

The compartmental SIR model, which is relatively simple and widely used, is a traditional approach to model infectious disease dynamics.

Consider a population of N individuals and the following simple discrete-time, discrete-state epidemic model. Each individual begins in one of the three possible states:

(1) susceptible, meaning that the individual has never had the disease and is susceptible to being infected;

(2) infected, meaning that the individual currently has the disease and can infect other people; and

(3) resistant, meaning that the individual does not have the disease, cannot infect others, and cannot be infected.

The model simulates the progression of the disease through the three states. Individuals are first susceptible, then infected, and then become resistant by acquiring immunity to the disease (Anderson et al, 1979; Bernoulli and Blower, 2004).

The model then evolves in discrete time steps, with all individuals simultaneously acting as follows in each time step:

(1) Each susceptible individual draws a uniformly random person from the population. If the person drawn is infected, then the susceptible individual changes his state to infected with probability β.

(2) Each infected individual changes his state to resistant with probability γ.

(3) Each resistant individual remains resistant.

The parameter β captures the ability of the disease to be transmitted from one person to another; the parameter γ is related to length of the period for which an individual can transmit the disease, called the infectious period. The population in this model is a homogeneously mixed population which interacts in such a uniformly random and independent way between time steps. In this model, there is a very important parameter R_0, called the basic reproduction number, which is the expected number of new infections created by an infected individual under the most favorable conditions for

transmission. For the compartmental SIR model, $R_0 = \beta/\gamma$ and generally the disease can become an epidemic only if $R_0 > 1$.

Here is a mass-action SIR compartmental model, where X(t), Y (t),and Z(t) denote the number of susceptible, infected, and resistant individuals in the population at time t. and X(t) + Y (t) + Z(t) = N.

$$\frac{dX(t)}{dt} = -\beta \cdot X(t) \cdot \frac{Y(t)}{N} \quad (1)$$

$$\frac{dY(t)}{dt} = \beta \cdot X(t) \cdot \frac{Y(t)}{N} - \gamma \cdot Y(t) \quad (2)$$

$$\frac{dZ(t)}{dt} = \gamma \cdot Y(t) \quad (3)$$

The model can also be extended to a more complex disease spread model with a more complex population structures. For example, a natural birth/death rate or a latent period of disease could be included in the model. For more information on the SIR model and its extension, see Dimitrov and Meyers, INFORMS Tutorial 2010.

Figure 6 provides an example of typical epidemic curves defined by the SIR model (Dimitrov and Meyers, INFORMS Tutorial 2010).

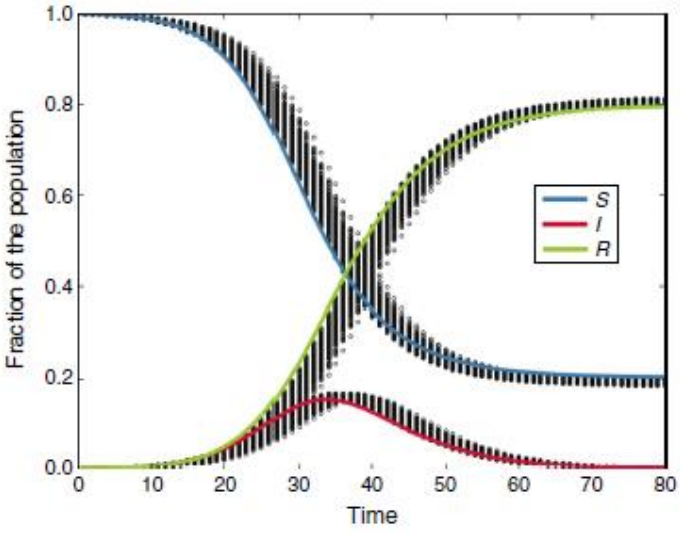


Figure 6. A typical curve of the compartmental SIR model

2.6 The Comparison of Three Disease Spread Models

Although the compartmental SIR models have proven to be quite useful in modeling epidemics, they do not properly model some important aspects of disease. Moreover, the compartmental SIR models assume a fully mixed, homogeneous population in which each individual has the same amount of contact as every other individual. Thus, simple SIR models do not accurately model the increased rate of contact in the hospitals and the decreased rate of contact of quarantined individuals. If the population at large had as many contacts as the population within a hospital, possibly the estimates of R_0 would have been more accurate, and SARS (Severe acute respiratory syndrome, see http://en.wikipedia.org/wiki/SARS) would have infected many more people.

Incorporating realistic contact patterns of the population is just one possible way to increase the fidelity of epidemic models. Diseases often spread at different rates based on age and the type of contact; they also have varying incubation periods in different age groups. For example, contacts at home tend to be more intimate than contacts at work. Hence, an infected person's family members are more susceptible to the disease. Also, disease spread is affected by both geographic location and seasonality.

Therefore, researchers have attempted to use high-fidelity agent-based simulation models, where each individual is tracked as they move from home to work and back. Such models involve complex parameterization and often require extensive computation that deems the models intractable and of limited usefulness. The social network modeling approach utilized in this research provides acceptable fidelity and

tractable formulations. This is illustrated in Figure 7 (Dimitrov and Meyers, INFORMS Tutorial 2010).

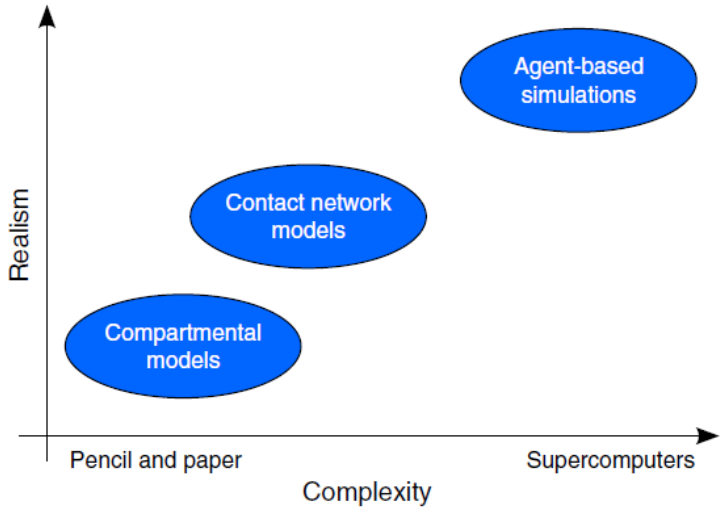


Figure 7. Complexity of epidemiological models

The Compartmental SIR models are easy to analyze but miss important, realistic details, such as heterogeneous patterns and types of contacts. Agent-based simulations are able to model reality with a great amount of detail, but are difficult to parameterize and analyze, and require large amounts of computation. Social network models capture disease transmission with a higher fidelity than compartmental models yet remain analytically tractable.

2.7 Travel Patterns

As we know, travelers are a rich source of information for infectious disease specialists. On returning from their journey, travelers can provide a representative sample of the diseases that abound in the places they have visited (Ross, 2006). There are a wide variety of travelers ranging from tourists and business people to immigrants, refugees and foreign-born citizens who have visited friends and relatives in their home countries. Moreover, a traveler who returns home with an unusual disease could be the first clue to a new outbreak (Ross, 2006).

Therefore, a disease tracking system could be built for sharing information among their networks of travel medicine clinics or hospitals such that doctors could record travel histories and symptoms of their patients and their diagnoses in standardized electronic forms and submit these to the system which in fact, is also a central database. Then the system regularly examines the data to detect the symptoms which might indicate a new outbreak and warrants a warning to the clinics or hospitals.

CHAPTER 3 BASIC METHODS OF DISEASE SPREAD

3.1 Basic Concepts of a Contact Network

The contact (or social) network is a hot concept across many disciplines, including sociology, epidemiology, biology, computer science and physics (Amaral and Ottino, 2004). A contact network model captures the patterns of interactions that can lead to the transmission of an infectious disease. And a social network, which focuses on the social relationship between nodes, is similar to a contact network in terms of analysis. A contact network is a contact structure made up of individuals (or organizations) represented as "nodes", that are tied (connected) by physical contacts (Horton, 2006). Contact network consists of nodes and ties (also called edges, links, arcs or connections). Nodes are the individuals within the networks, and ties are the physical contacts between the individuals. The resulting graph-based structures are often very complex. Contact network plays a critical role in determining the way problems are solved, organizations are run, and the degree to which individuals succeed in achieving their goals.

Moreover, contact network analysis has also been used in epidemiology to help understand how patterns of human contact aid or inhibit the spread of diseases such as HIV in a population (Parker, 2002). The evolution of contact networks can sometimes be modeled by the use of agent based models, providing insight into the interplay between communication rules, rumor spreading and social structure.

3.2 Network Centrality

3.2.1 Individual Centrality in Contact Network

The centrality in a contact network is a parameter used to measure how central an individual is in a contact network (Freeman, 1979). The concept of centrality was formally defined by Freeman. Specifically, Freeman identified three primary centrality measures: degree, closeness and betweenness.

(1) Degree centrality measures an individual's direct connectedness with other individuals;

(2) Closeness centrality provides a more global network prospective than degree centrality. Specifically, closeness centrality is a measure that indicates the degree to which an individual is near all the other individuals in the network not just those adjacent to them;

(3) Betweenness centrality is a measure of the strategic location of an individual along a potential communication path.

The study of the centrality in a contact network could determine the most central individuals who play a critical role in the disease spread. Obviously, the rate of spread of a potential pandemic could be mitigated to some extent by monitoring and immunizing those with higher centrality values.

3.2.2 The Traditional Method of the Centrality

Consider a contact network modeled as a direct graph, G(V, E). Let V = (v1, v2,…, v_n) denote the set of nodes in the network, and let E = (e1, e2, …, e_m) denote the set of edges between the nodes (Hamill et al. 2006). Specifically, consider the modified network from Hamill with n = 11 nodes with the social relationships (or paths of communication) depicted in Figure 8 (Schneider et al 2011).

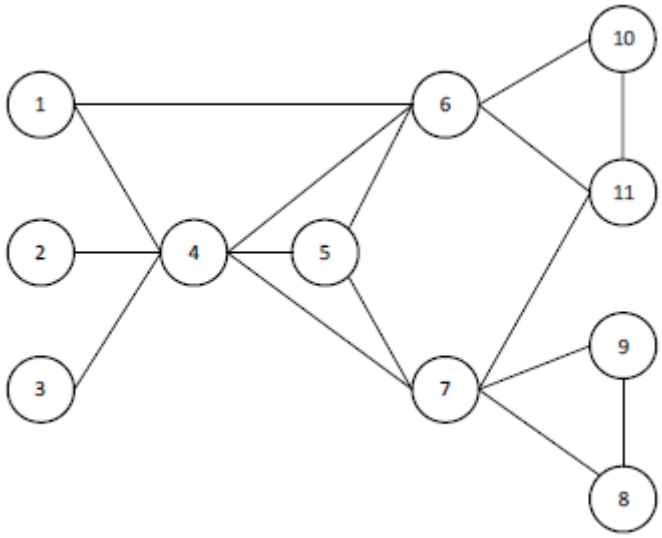


Figure 8. Sample Network

One of the most important metrics in the social network analysis (SNA) is the centrality of an individual (*each node in the network*). There are three main centrality measures: degree, closeness and betweenness (Freeman, 1979). The traditional way to calculate these three centrality measures is introduced below:

Degree centrality $C_D(v)$: measures an individual's direct connectedness with other individuals. The degree of a node (or vertex) is the number of edges connected to it. Let deg(v) denote the degree of an individual v in the network which have n individuals, and then the degree centrality is given by

$$C_D(v) = \frac{deg(v)}{n-1} \tag{4}$$

Closeness centrality $C_C(v)$: a measure that indicates the degree to which an individual is close to all the other individuals in the network not just those adjacent to them. It provides a more global network prospective than degree centrality. Let $d_G(v, c)$ denote the length of a shortest path connecting individual v with individual c, so that the closeness centrality of an individual v is given by

$$C_C(v) = \frac{1}{\sum_{c \in V} d_G(v,c)} \tag{5}$$

Betweenness centrality $C_B(v)$: a measure of the strategic location of an individual along a potential communication path. Let σ_{bc} denote the number of shortest paths from individual *b* to individual *c*, and let $\sigma_{bc}(v)$ denote the number of shortest paths from individual b to individual c that contain individual v. The betweenness centrality of an individual v is given by

$$C_B(v) = \sum_{b \neq v \neq c \in V} \frac{\sigma_{bc}(v)}{\sigma_{bc}} \tag{6}$$

The centrality measures for the network in Figure 8 are shown in Table 2 (Schneider et al 2011). From Table 2, we can see that individual 4 has the largest centrality values for each measure. Individuals 6 and 7 exhibit the second highest degree centrality and meanwhile individual 7 exhibits the second highest closeness and betweenness centrality as well. All these results could indicate which individuals are the most "central" people in a social network to some extent. However, these metrics treat each individual in the network identically and assume a perfect contact chain. In reality, certain individuals within the network may be more persuasive and the contact between individuals in the network may not be that perfect since their centrality values are so close such that we could not tell the differences among individuals based on their

centralities. In these instances, the centrality metrics may not adequately quantify the criticality of individuals within the network (Schneider et al 2011).

Individual	Degree	Closeness	Betweenness	Betweenness*
1	0.20	0.0500	0	0
2	0.10	0.0435	0	0
3	0.10	0.0435	0	0
4	**0.60**	**0.0714**	**43**	**1**
5	0.30	0.0588	2	0.0465
6	**0.50**	0.0588	17	0.3953
7	**0.50**	**0.0667**	**36**	**0.8372**
8	0.20	0.0435	0	0
9	0.20	0.0435	0	0
10	0.20	0.0455	0	0
11	0.30	0.0526	8	0.7273

Table 2. Traditional centrality measures of sample network

3.2.3 A Proposed Centrality Measure

Consider a contact network with n nodes with the connection probabilities between every two nodes given. The connection probability is the likelihood of contact between two individuals and it could be obtained by recent frequency of physical contact between the two individuals. The connection probability could reflect the possibility for one individual to get an infection from another individual to some extent, although the chance of getting infected also depends on other factors, such as the feature of the disease, the feature of the population and the contact network structure.

There are several steps to calculate the centrality value of a specific node i; these steps are illustrated using the simple network in Figure 9. R is the transmission rate between two specific nodes and it is a value between 0 and 1. Particularly, if R = 0, it indicates

that there is no direct connection between these two nodes; if R = 1, it indicates that two individuals will infect each other by all means.

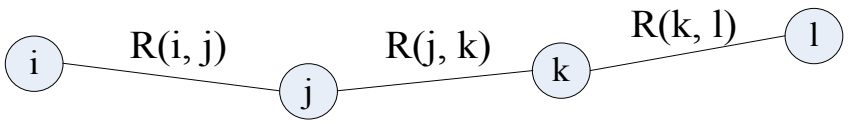


Figure 9. Sample arc in a network

1. For each node, find the first-nearest-layer (FNL) nodes which have direct connection with the original node i. In this case, node j is the only FNL nodes and the impact of FNL nodes (FNL values) could be obtained by the formula below.

$$FNL(i) = \sum_{j \neq i}^{n} R(i,j) \qquad i=1,2,\ldots,n \qquad (7)$$

2. For each FNL node, generate a second loop to their FNL nodes, which would be the second-nearest-layer (SNL) nodes of the initial node i. In this illustrative example, node k is the SNL node and their impact (SNL value) towards the initial node i could be obtained by the formula below:

$$SNL(i) = \sum_{j \neq i}^{n} \sum_{k \neq i, k \neq j}^{n} \{R(i,j) \times R(j,k)\} \qquad i=1,2,\ldots,n \qquad (8)$$

3. Similarly, for each SNL node, generate a loop to their FNL nodes, which are the third-nearest-layer (TNL) nodes of the initial node i and their indirect impact values (TNL value) towards the initial node i could be calculated by the formula below:

$$TNL(i) = \sum_{j \neq i}^{n} \sum_{k \neq i, k \neq j}^{n} \sum_{l \neq i, l \neq j, l \neq k}^{n} \{R(i,j) \times R(j,k) \times R(k,l)\}$$

$$i=1,2,\ldots,n \qquad (9)$$

4. Finally, the centrality of node i would be the summation of all the impacts of FNL, SNL and TNL.

The algorithm as described above is straight forward. However, more thoughts and work are needed when we realize it by using Matlab. Here is how it works: First of all, we build a n × n matrix in which each element presents the degree of connectedness between node i and j. In fact, this is a symmetric matrix which is called connectedness matrix whose elements are already known to us. Besides, we have to build another three zero n ×n matrices which are FNL matrix, SNL matrix and TNL matrix respectively. Each element in these three matrices represents the corresponding centrality values as the FNL, SNL and TNL nodes respectively. For example, element (i, j) in these three matrices may have different values which are not equal to zero and this indicates that as a FNL, SNL and TNL node, the same node j will have three impacts on the same node i. Therefore, the final impact which node j contributes to node i would be the summation of these three impacts. Secondly, we go through all the elements in the connectedness matrix and calculate all the values in FNL, SNL and TNL matrices in which each elements represents the FNL, SNL and TNL impacts a specific node j for another node i. Then the centrality matrix would be obtained by adding all these three matrices up since they have the same structure. Finally, if we sum all the values up in each row, an n dimension vector would be obtained in which each element represent the final centrality value for each node in the network.

3.2.4 Utilizing the Proposed Measure to Identify Central Nodes in a Sample Contact Network

By using both the traditional metrics and the proposed method, we obtain the results in Table 3. We can see that individual 4 has the largest values for each traditional measure. Individuals 6 and 7 exhibit the second highest degree centrality and meanwhile individual 7 exhibits the second highest closeness and betweenness centrality as well. All these results could indicate which individuals are the most "central" people in a contact network to some extent. However, as previously mentioned these metrics treat each individual in the network identically and assume the same intensity of contact among all individuals. In reality, this is not the case. In these instances, the traditional metrics may not adequately quantify the criticality of individuals within the network. However, using the centrality metrics proposed in this dissertation, we note that individuals 4, 6 and 7 have a higher centrality values than others which indicates that they the most central individuals in the network who should have a higher priority to get vaccinated.

Individual	Degree	Closeness	Betweenness	Betweenness*	Centrality	Centraliy*
1	0.20	0.0500	0	0	3.4640	0.6307
2	0.10	0.0435	0	0	1.3080	0.2382
3	0.10	0.0435	0	0	0.8920	0.1624
4	0.60	0.0714	43	1	5.1880	0.9446
5	0.30	0.0588	2	0.0465	5.4920	1
6	0.50	0.0588	17	0.3953	4.9190	0.8957
7	0.50	0.0667	36	0.8372	5.3190	0.9685
8	0.20	0.0435	0	0	3.2100	0.5845
9	0.20	0.0435	0	0	3.2100	0.5845
10	0.20	0.0455	0	0	3.4320	0.6249
11	0.30	0.0526	8	0.7273	4.3780	0.7972

Table 3. Metrics of a sample network

Consider another sample non-dynamic contact network model which contains 30 individuals represented in Figure 10. The number beside each arc indicates the

connection probability between two adjacent nodes. Using this sample we illustrate the utilization of the proposed centrality measure to identify “central” nodes in the network.

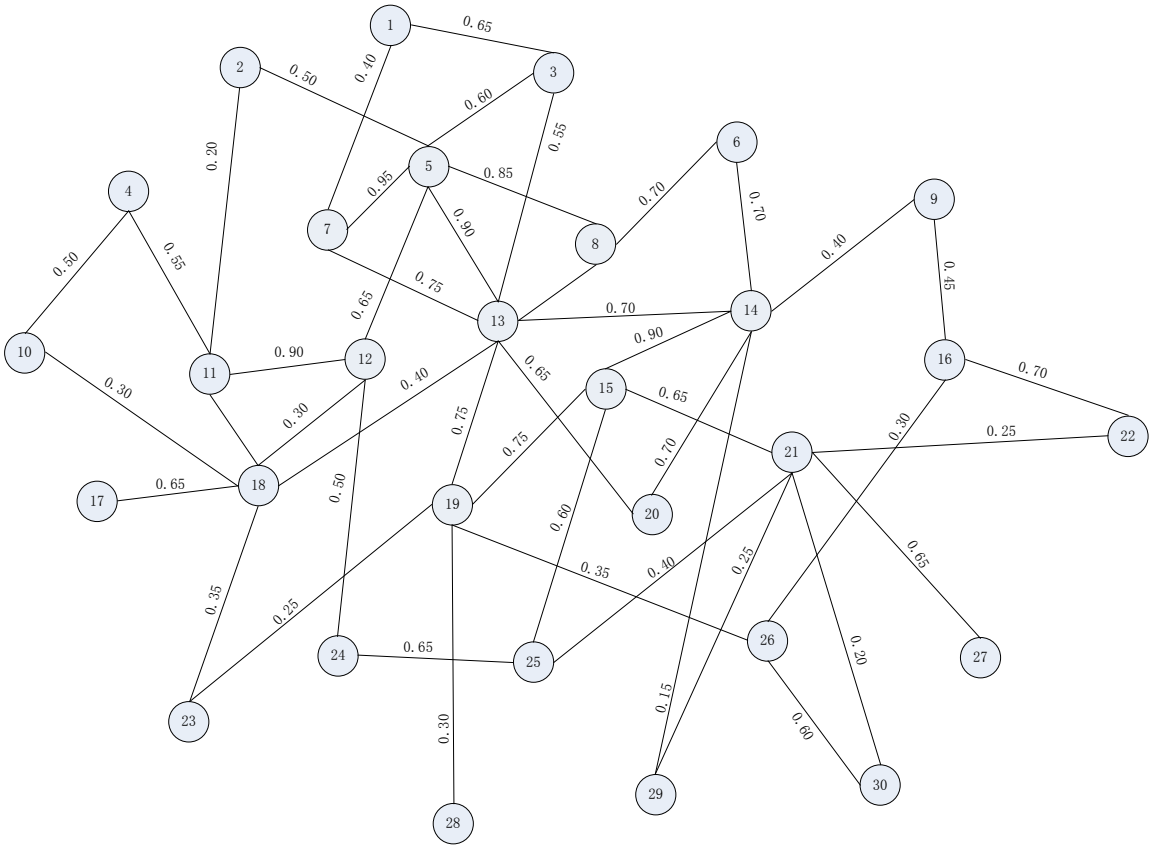


Figure 10. A 30-size sample population network

The centrality values of all nodes based on the proposed measure are shown in Table 4.

Node	Value	Node	Value	Node	Value
1	5.8410	11	12.1001	21	8.1659
2	7.4620	12	15.4282	22	2.8680
3	7.3070	13	26.0583	23	5.0380
4	4.6883	14	20.6846	24	6.8458
5	21.0422	15	15.5823	25	8.1482
6	12.5615	16	3.2845	26	4.5005
7	21.4805	17	5.3983	27	3.2793
8	21.1098	18	14.8474	28	2.4795
9	5.0700	19	17.4846	29	2.8969
10	3.8628	20	16.4306	30	2.7025

Table 4. Centrality values of a 30-size population network

From Table 4, we can see that nodes 5, 7, 8, 13 and 14 seem more central than other individuals. The Matlab program developed to automate the calculation is shown in the Appendix.

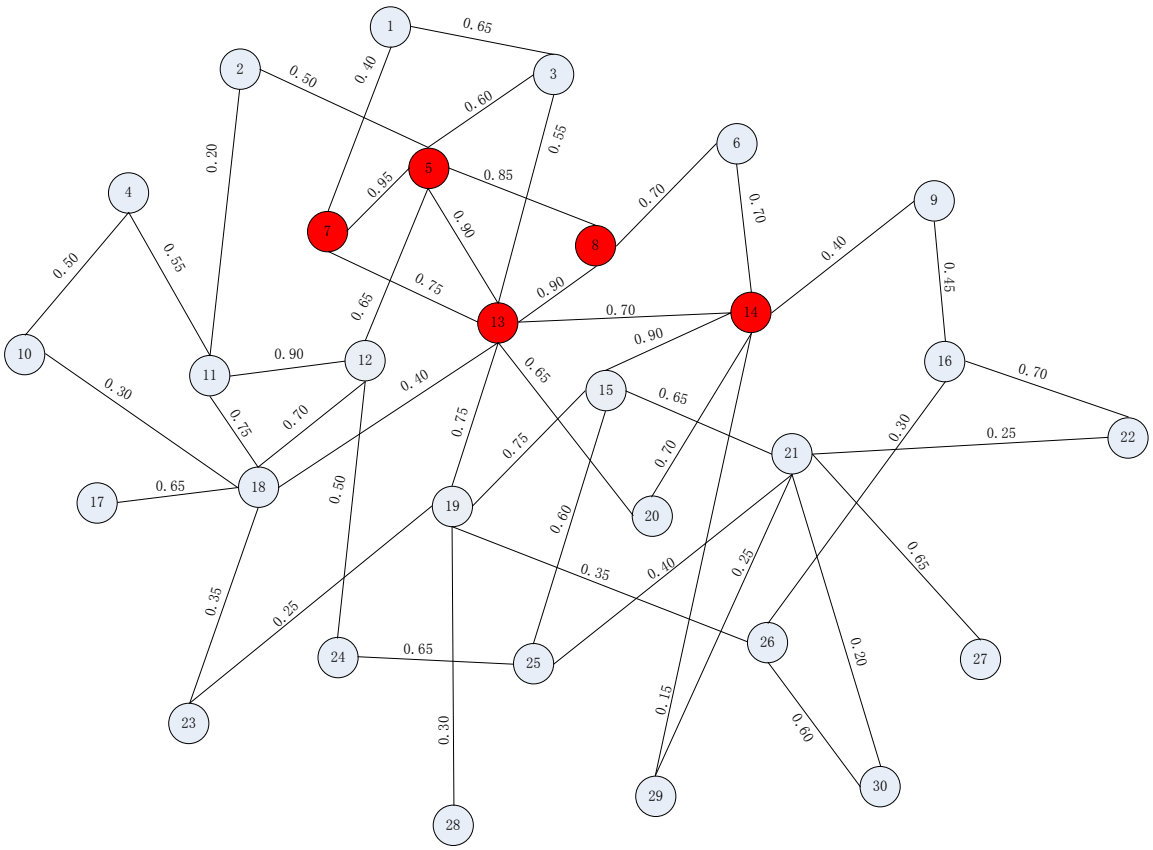


Figure 11. More central individuals in the sample network

Visually examining, the identified central nodes (red nodes) and their probabilities in the network provide face validity to the suggested measure of centrality.

3.3 Extension of Network Analysis

In fact, each node in the network could be colored according to their centrality values in order to classify different importance of the nodes like the social network below (Christakis, 2010). Table 5 demonstrates the relationships between range of the risk index and color and risk level.

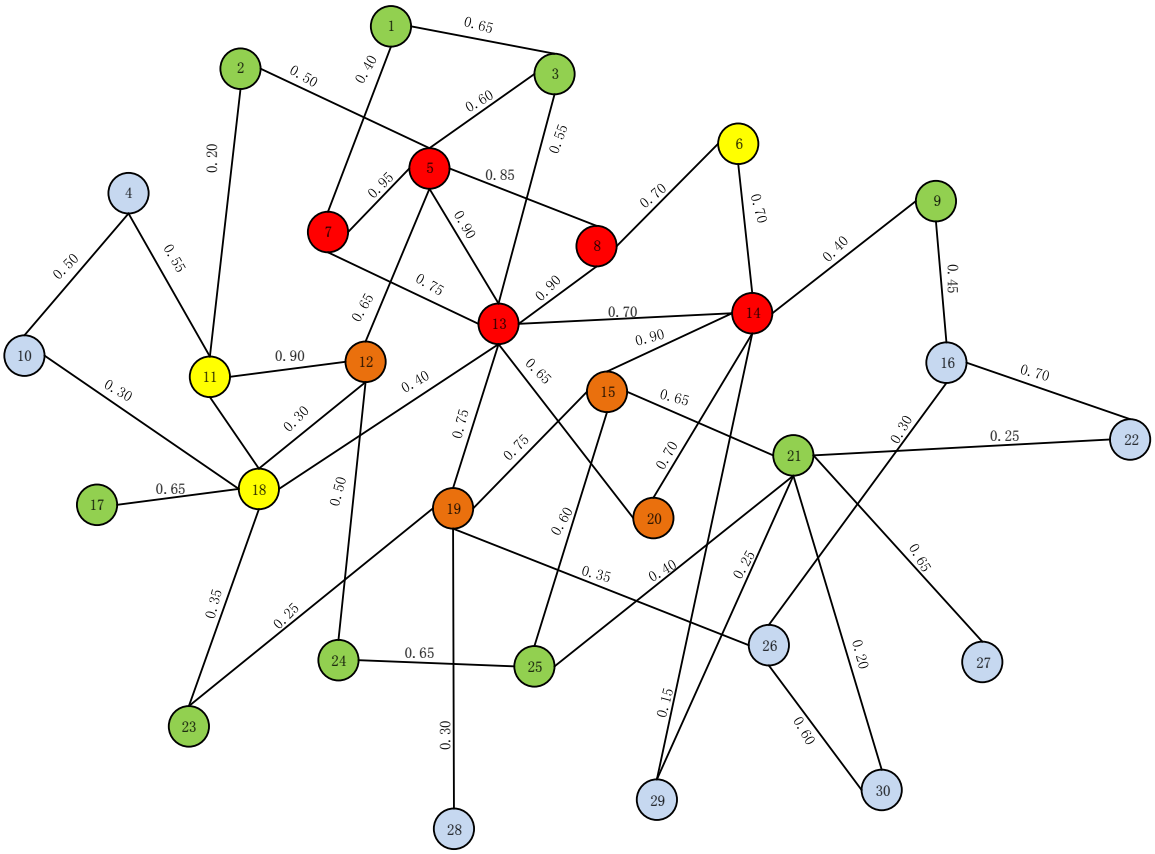


Figure 12. The colored 30-size sample population network

The social network visualization can support such intervention in numerous ways.

First of all, they can be used to identify clusters of connected individuals with similar disease susceptibility and health-relevant attributes. The clusters could be targeted for collective interventions. Secondly, they can be used to identify and target individuals for public heath interventions. Thirdly, the knowledge of the overall network structure may be crucial to the design of public health intervention strategies.

Range	Color	Risk Level
0.000 – 5.000	Light Blue	Extremely Low
5.001 – 10.000	Green	Low
10.001 – 15.000	Yellow	Medium
15.001 – 20.000	Orange	High
20.001 – above	Red	Extremely High

Table 5. The relationship between risk index and risk levels

3.4 Transmission Rate

Since the air-borne infectious disease could be transformed rapidly between individuals through physical contact, it is critical to do some analysis on the transmission rate of the infectious disease, especially for the emerging disease, which have newly appeared in a population or have existed but are rapidly increasing in geographic range, such as SARS and H1N1. Some potential factors precipitating disease emergence can be identified in virtually all cases. These include ecological, environmental, or demographic factors that place people at increased contact with a previously unfamiliar microbe or its natural host or promote dissemination (Morse, 1995).

In this dissertation, we will use the data mining methodologies and predictive modeling to perform the analysis on the potential factors of the transmission rate for a specific infectious disease, which is the disease parameter γ in the formula. And the potential risk factors might include the gender, age, the month to get infected, smoker or not, obesity, the amount of exercise per week, vegetarian or not, high blood pressure, diabetes, medical care rate and so forth. Then by building different models and

weighting different factors by the sample data, we could get a reasonable value for the disease parameter γ, which would be used in the formula to obtain the infection index $F_{ab.}$

3.5 Data Mining Methodology

It is widely recognized that the risk factors of getting infected includes both individual and disease characteristics. The variables considered in this analysis are the potential risk factors or the infected diseases and they do not affect the presence of each individual disease equally. Therefore, several data mining methods have been applied to get a comprehensive parameter by taking all the potential risk factors into account.

3.5.1 Introduction to Data Mining

In this dissertation, we will use the data mining methodologies to extract the useful information from our datasets. The reason we utilize data mining rather than statistical methods is that data mining is more practically oriented discipline than the statistics.

- Dataset must be prepared with appropriate preprocessing techniques in data mining and the preparation can have as much or even more influence on the quality of the final results than the selected technique.
- Data mining use flexible predictive techniques that are often based on strong algorithms (such as artificial neural network and decision trees).

- Data mining attempts to find not only general models base on the dataset but also local patterns in large dataset, which is especially useful when the number of dimensions are relatively large.

Specifically, data mining concentrates on data management and optimization of data searches (with a focus on problems of preprocessing, data cleaning, algorithms, and data structures). Statistics is more oriented toward formalisms for final model representation and score function formalization in the data space to perform inference (with a focus on problems of models and principles of statistical inference).

Also, the fitness and quality of data is another critical aspect in data analysis. When we are analyzing the data statistically, the datasets should be in a format that we can analyze. Unfortunately, it is very common that most datasets in the real world have poor quality and therefore we must consider the quality of the data as well as the applicability of the statistical models.

Finally, the main difference between data mining and statistics is the size of the datasets. In statistics, p-values and statistical significance are the primary measures of model fitness. However, in real world the datasets are generally so large that p-values have no meaning. Also the confidence width and effect size in such large datasets decreases to 0 as the dataset size increases. It is usual to have a regression model with every parameter statistically significant but with a correlation coefficient of almost 0. Therefore, we need other methods to measure the fitness of the model. In data mining,

the data sets are usually large enough to partition into three types: training, testing, and validation. The training data set is used to define the model, the testing data set is used in an iterative process to change the model if necessary to improve it, and the validation data set represents a final examination of the model. Depending upon the profit and loss requirements in the data, misclassification is used in supervised learning where there is a specific outcome variable (Cerrito, 2006).

3.5.2 Decision Tree

A decision tree serves as a tree-like graph to display decision models and their relevant potential consequences, among which are chance event outcomes, cost of resource and utility. As a tool to present an algorithm, decision trees are commonly used to help identify a strategy to achieve a goal in many fields, especially in decision analysis and operations research. Decision trees and decision rules are data-mining methodologies applied in many real world applications as a powerful solution to classification problems. A well-known tree-growing algorithm for generating decision trees is Quinlan's ID3 with an extended version called C4.5(Kantardzic, 2011).

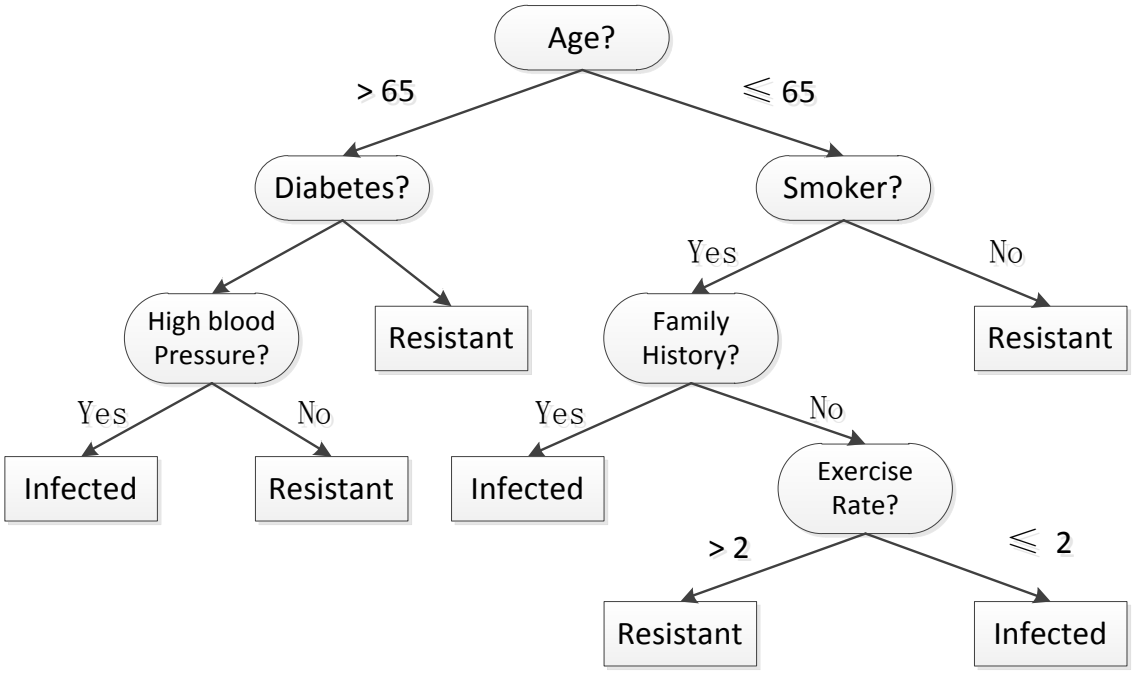


Figure 13. An example of a decision tree

The ID3 algorithm creates the root node of the tree first and searches all the training samples. And then an attribute would be selected to partition these samples and a branch is generated. Based on the value of this attribute, the corresponding subset of samples which have the attribute value specified by the branch is moved to the new child nodes separately. The algorithm will repeat all these steps recursively until all samples are at least in one class. Finally, a decision tree is created and each leaf in the decision tree represents a classification rule.

C4.5 algorithm is the extension of ID3 algorithm, which extends the domain of classification from categorical attributes to numeric ones (Kantardzic, 2011). Both ID3 and C4.5 algorithms are based on mining an information entropy measure applied to the samples at a node. Specifically, the C4.5 favors attributes which result in

partitioning the data into subsets with a low-class entropy, which indicates that the majority samples in it belong to a single class.

We will use C4.5 algorithm as the main decision tree methodology to classify all the potential factors of individual conditions which would have a contribution of disease spread with different weights. And a decision tree would be created and illustrate the classification rules of the infection. Finally, the analysis would be made for the decision tree results to figure out the different impacts from all the potential individual factors of the disease spread.

3.5.3 Artificial Neural Network

An artificial neural network (ANN), usually called neural network (NN), is a mathematical model or computational model that is inspired by the structure and/or functional aspects of biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Modern neural networks are non-linear statistical data modeling tools. They are usually used to model complex relationships between inputs and outputs or to find patterns in data.

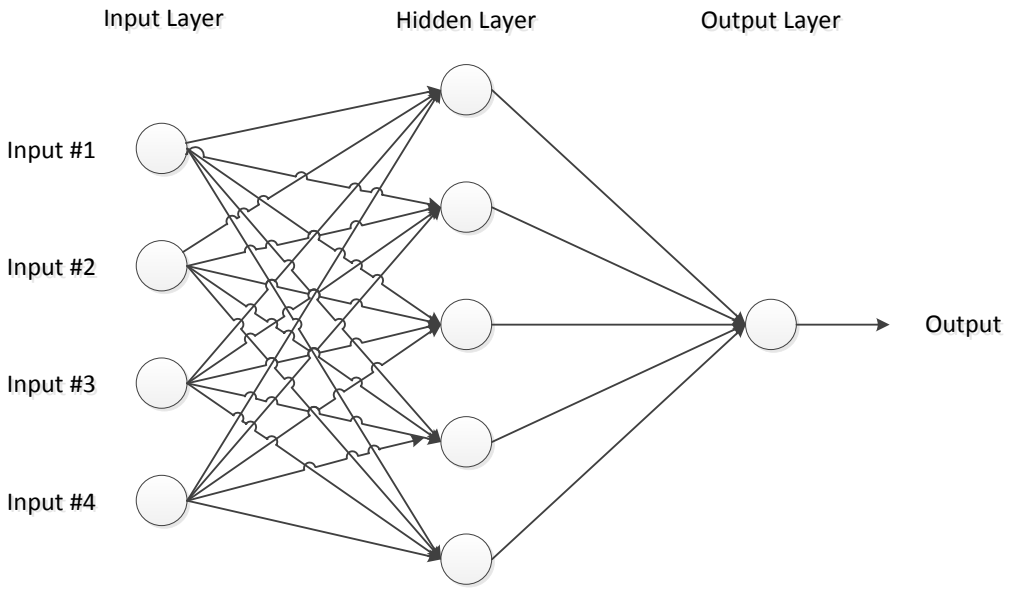


Figure 14. The Structure of artificial neural network

Multilayer feedforward networks are one of the most important and most popular classes of ANNs in real-world applications (Kantardzic, 2011). A multilayer perceptron (MLP) is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate output. It consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Each node is a neuron or a processing element with a nonlinear activation function in addition to the input node. MLP is the standard algorithm for any supervised learning pattern recognition process, which can also distinguish data that is not linearly separable.

Artificial neural networks act like black box and there is no definite equation or model and the model is not presented in concise format available for regression. Its accuracy is examined similar to the diagnostics of the regression curve including misclassification rate and the average error. Its complexity increases with the number of hidden layers and input variables increases. Each input variable is connected to each

variable in the hidden layer and each hidden variable is connected to each outcome variable. The hidden layers combine inputs and apply a function to predict outputs.

3.6 Logistic Regression

Regression Analysis is widely used to estimate the relationships among variables and predict the future situation. It is helpful to understand how the typical value of the dependent variable changes when the independent variable varies by using regression analysis. Regression analysis is also interesting theoretically because of elegant underlying mathematics and a well-developed statistical theory. Successful use of regression requires an appreciation of both the theory and the practical problems that typically arise when the technique is employed with real-world data (Montgomery, 1992). There are monadic regression analysis and multivariate regression analysis which depends on the numbers of input variables or independent variables; also, there are linear regression analysis and nonlinear regression in terms of the relationship between the independent variables and dependent variables. Besides, there is the logistic regression which would be used in this dissertation.

Logistic Regression has become, in many fields, the standard method of analysis in the situation that the outcome variable is discrete, taking on two or more possible values (Hosmer and Lemeshow, 1989). In epidemiology, we would like to determine the probabilities, which are bounded by 0 and 1, for a susceptible individual to get infected from a specific disease. But the linear functions are inherently unbounded. Therefore, the logistic regression would be applied to transform the probability such that it's no

longer bounded (Allison, 1999). To better understand the logistic regression, it's helpful to have an understanding of odds, which is the ratio of the expected numbers of times that an event will occur to the expected number of times it will not occur. If p is the probability of an event and O is the odd of the event, then

$$O = \frac{p}{1-p} = \frac{probability\ of\ event}{probability\ of\ no\ event} \tag{10}$$

Transforming the probability to an odds will removed the upper bound and also the lower will be removed if the logarithm of the odds is taken. A logit model will be obtained after setting the result to a liner function of the explanatory variables. For example, if there are k explanatory variables, the model is

$$log\left(\frac{p}{1-p}\right) = \alpha + \beta_1\chi_1 + \beta_2\chi_2 + \ldots + \beta_k\chi_k \tag{11}$$

where p is the probability and the expression on the left-hand side is usually referred to as the logit or log-odds. Either the natural logarithms or the base-10 logarithms could be used and the x's may be either interval-level variables or dummy variables in the ordinary regression model. We can solve the logit equation for p to obtain

$$p = \frac{exp\left(\alpha + \beta_1\chi_1 + \beta_2\chi_2 + \ldots + \beta_k\chi_k\right)}{1 + exp\left(\alpha + \beta_1\chi_1 + \beta_2\chi_2 + \ldots + \beta_k\chi_k\right)} \tag{12}$$

Exp(x) is the exponential function, equivalent to e^x and the equation has the desired property that no matter what values we substitute for the β's and the x's, p will always be a number between 0 and 1, which could be used as the probability of getting infected from a specific disease for a susceptible individual.

In addition, information Value is used to evaluate the overall predictive power of a characteristic, which is the characteristic's ability to separate between good and bad

records (susceptible individuals and infected individuals in our case). Information value is calculated as follows:

$$IV = \sum_{i=1}^{L} \left(DistrGood_i - DistrBad_i \right) \cdot \ln \left(\frac{DistrGood_i}{DistrBad_i} \right) \tag{13}$$

Here L is the number of attributes for the characteristic variable. In general an information value less than 0.02 is unpredictive, a value between 0.02 and 0.10 is weakly predictive, a value between 0.10 and 0.30 is moderately predictive, and a value greater than 0.30 is strongly predictive. Also, the Gini statistic is used as an alternative to the information value.

The weight of evidence (WOE) measures the strength of an attribute of a characteristic in differentiating susceptible and infected individuals. Weight of evidence is based on the proportion of susceptible individuals to infected individuals at each group level. For each group i of a characteristic WOE is calculated as follows:

$$WOE = ln \left(\frac{DistrGood_i}{DistrBad_i} \right) \tag{14}$$

Negative values indicate that a particular grouping is isolating a higher proportion of infected individuals than susceptible individuals. That is, negative WOE values are worse in the sense that individuals in that group present a greater infection risk. By default, missing values are assigned to their own group.

3.7 Model Performance Assessment

In statistics, we use the coefficient of determination R^2 to measure how well the regression line represents the data. First, let's look at its definition: suppose we have a data set with observed values y_i and each of them has an associated predicted value $\hat{y}_i$. The mean of the observed data

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{15}$$

Where n is the number of observations. The variability of the data set is measured through different sums of squares.

$$SST = SSR + SSE$$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 \tag{16}$$

where:

SST = total sum of squares
SSR = sum of squares due to regression
SSE = sum of squares due to error

We define the coefficient of determination $R^2 = SSR/SST$, and $0 \leq R^2 \leq 1$, which is useful since it gives the proportion of the variance of one variable that is predictable from the other variable. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of variation. The further the line is away from the points, the less it is able to explain.

A model realized through the data mining process using different inductive – learning techniques might be estimated using the standard error rate parameter as a measure of its performance (Kantardzic, 2011). An approximation of the true error rate which expressed by this value, can be computed by using a testing data through the data

mining techniques we introduced above. In terms of other parameters, such as speed, robustness and interpretability, the data mining models could also be compared. All of this might have an influence on both the verification of the validation of the final model. Here, we will use both confusion matrix and ROC curve to estimate the different data mining models.

The Confusion Matrix is commonly used to assess the prediction accuracy of a model, especially for classification models. It measures whether a model is confused or not, that is, whether the model is making mistakes in its prediction. . It is a specific table layout that visualization of the performance of an algorithm, typically a supervised learning one. Each column of the matrix represents the samples in the predicted models while each row is the instances in an actual class. The format of a confusion matrix for a two-class case with classes yes and no is shown in Table 6 (Kantardzic, 2011).

Actual Class / Predicted Class	Class 1 = Yes	Class 2 = No
Class 1 = Yes	A: True + (TP)	B: False + (FP)
Class 2 = No	C: False – (FN)	D: True – (TN)

Table 6. Confusion matrix for two-class classification model

Consider a two-class prediction problem in which the outcomes are labeled either as positive (p) or negative (n). Then measures represented in the confusion matrix include:

1) True Positive(TP): both the outcome from a prediction and the actual value are p;

2) False Positive(FP): the prediction outcome is p but the actual value is n;

3) True Negative(TN): both the prediction outcome and the actual value are n;

4) False Negative (FN): the prediction outcome is n while the actual value is p.

FP is also called Type I error and FN is called Type II error as well. And one of the most widely used metric for assessing the data mining methodologies is "Accuracy" defined as follows:

$$\text{Accuracy} = \frac{A + \text{D}}{A + B + C + \text{D}} = \frac{TP + TN}{TP + FP + FN + TN} \tag{17}$$

In addition, the Received Operating Characteristic (ROC) or simply ROC curve is also applied for assessing the classification model. The ROC can be represented equivalently by plotting both the true positive rate (TPR) and the false positive rate (FPR), which are the fraction of true positives out of the positives and the fraction of false positives out of the negatives respectively:

$$\text{TPR} = \frac{A}{A + C} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{B}{B + \text{D}} = \frac{FP}{FP + TN} \tag{18}$$

A ROC space depicts relative trade-offs between true positive (benefits) and false positive (costs), which is defined by FPR and TPR as x and y axes respectively. And TPR is equivalent with sensitivity and FPR is equal to 1- specificity, therefore the ROC curve is also the sensitivity and (1- specificity) plot. Each prediction result or instance of a confusion matrix represents one point in the ROC curve as shown in Figure 15.

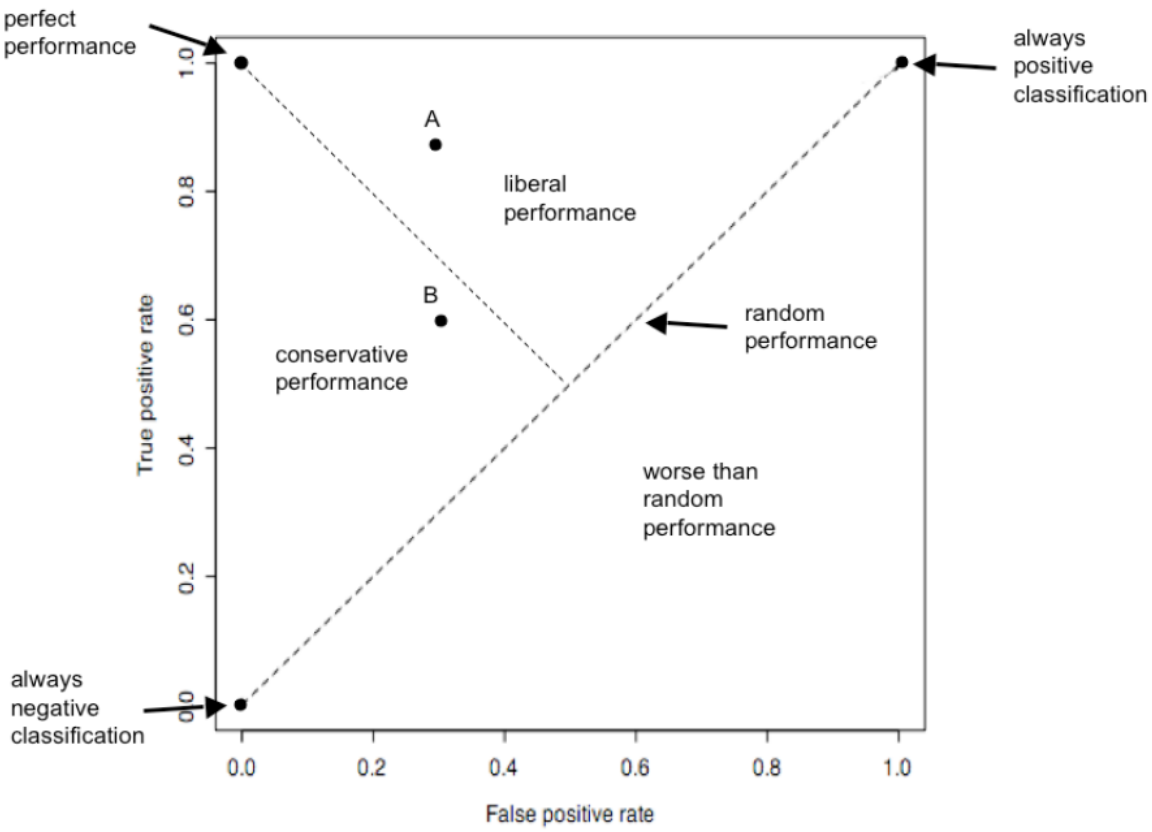


Figure 15. A ROC space

The top left corner is the perfect performance point and the closer to the top left corner, the better the performance is. So in Figure 15, the point A is better than the point B and ROC curves characterize the performance of a classification mode as a curve instead of a single point. The ROC curve is an appropriate tool to measure the success of subgroup discovery for classification models since subgroups could be discarded as insignificant if their TPR/FPR tradeoff is close to the diagonal which represents random performance. Conversely, those sufficiently far away from the diagonal are the significant subgroups (Xiaoyi, 2008). To summarize, the following classification measures are derived from the relationship in the confusion matrix, some of which represents the sensitivity and specificity in the ROC curve.

- Classification (Accuracy) Rate: (A + D) / (A + B + C + D)

- Misclassification Rate: (B + C) / (A + B + C + D)
- Sensitivity (True Positive Rate): A / (A + C)
- Specificity (True Negative Rate): D / (B + D)
- 1 – Specificity (False Positive Rate): B / (B + D)

Additionally, cumulative gains and lift charts can be also used to measure the model performance. Both charts consist of a lift curve and a baseline and the greater the area between the lift curve and the baseline, the better the model is. Lift is a measure of the effectiveness of a predictive model, which could be calculated as the ratio between the results obtained with and without the predictive model. Also, Lift is the ratio of the percentage of targets which is infected individuals in each decile to the percent of targets in the entire data set. Cumulative lift is the cumulative ratio of the percent of targets up to the decile of interest to the percent of targets in the entire data set. For lift and cumulative lift, the higher value in the lower deciles indicates a predictive model.

The Kolmogorov-Smirnov statistic is the maximum distance between the empirical distribution functions for the susceptible individuals and infected individuals. The difference is plotted, for all cutoffs, in the Kolmogorov-Smirnov Plot. The weakness of reporting only the maximum difference between the curves is that it provides only a measure of vertical separation at one cutoff value, but not overall cutoff values.

Also, the average square error is another fundamental statistical measure of the model performance in SAS. The squared difference between a predicted value and an actual value for the dependent variable is called the squared error and averaged over all cases, we obtain

$$Averaged\,square\,error = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2 \tag{19}$$

Where N is the number of cases or records or observations and for i^{th} case, y_i is the actual target value and $\hat{y}_i$ is the predicted target value. A model with a lower average square error is less biased, more accurate than a model with a higher average square error.

CHAPTER 4 ASSESSMENT OF DISEASE SPREAD BASED ON SPATIO-TEMPORAL INFORMATION

Although there are a variety of different tools to predict and detect the disease spread including the SIR model, contact network models and mathematical model, we suggest that it would be more effective and beneficial to predict the disease spread considering the intensity of exposure to the disease of susceptible individuals. This would involve the consideration of both space and time factors. Since diseases caused by either viruses or bacteria involve some type of contact, either direct (shaking hands) or through the atmosphere (e.g. coughing or sneezing) between the susceptible and infected individuals.

In this dissertation we use the concept of space-time paths to embellish the prediction of disease spread. As we know, the space-time path records the main activities and tracks during a certain period of time for an individual. Suppose there are two individuals and one is susceptible and the other is infected. By considering the space-time paths of the individuals, we could estimate exposure intensity and assess the likelihood for the susceptible individual to get infected.

In this case, both space and time would be taken into account which is unique in this dissertation as well as other parameters, such as the characteristics of the disease. The length of time for overlapping parts of the two-individual space-time paths could be a descriptive method to evaluate the chance for one individual to get infected from the other. More detail is on the procedure is provided in the next section.

4.1 Risk Analysis of Infection

Consider two individuals A & B. A is infected and B is susceptible. The locations of A & B at any point in time t, specified by geographic coordinates ($X_a(t)$, $Y_a(t)$) and ($X_b(t)$, $Y_b(t)$) respectively.

Therefore, their relatively proximity at any time t, can be represented by the Euclidean distance $D_{ab}(t)$ as indicated below:

$$D_{ab}(t) = \sqrt{\left(X_a(t) - X_b(t)\right)^2 + \left(Y_a(t) - Y_b(t)\right)^2} \tag{20}$$

Consider the start and end times that A & B are in the "infection range" at location i (i = 1, 2, …, n) as T_{si} and T_{ei}. where n presents the number of segments in common on the space-time paths for both A and B over a certain period of time.

In fact, for each location i, D_{ab} could be regarded as a constant.

Therefore, $D_{ab} = D_i = \sqrt{\left(X_{ia} - X_{ib}\right)^2 + \left(Y_{ia} - Y_{ib}\right)^2}$

We propose a function F_{ab} that incorporates individual and spatio-temporal factors to define the likelihood that B catches the infection from A, while at location i.

$$F_{ab} = \sum_{i}^{n} \int_{Ts_i}^{Te_i} \frac{\gamma_s \gamma_d \left(t - Ts_i\right)^p}{Di_{ab}{}^p + \varepsilon} \, dt \tag{21}$$

Where p is the time parameter and q is the distance parameter. p should be between 0 and 1; q should be greater than 1. Both p and q could be obtained from experimental and field studies. Additionally, ε is a very small positive number which prevents the denominator from being zero.

γ_s is a parameter that defines the susceptibility of a particular individual to the infectious disease. This would depend on a variety of factors, and is discussed further in the next section; γ_d is a disease parameter which represents the virulence of the disease and could also be estimated by R_0 in the compartmental SIR model approximately.

Let $\gamma = \gamma_s \times \gamma_d$ and γ is roughly equal to the force of infection in epidemiology, which is the rate at which susceptible individuals become infected by an infectious disease. The advantage of the method to calculate the force of infection is that data on the average age of infection is very easily obtainable from doctors' reports, even though they are not reporting all cases of the disease. It also can be used to compare the rate of transmission between different groups of the population for the same infectious disease or even between different infectious diseases. If the force of infection is denoted as λ, then we have

$$\lambda = \frac{number\ of\ new\ infections}{number\ of\ susceptible\ persons\ exposed \times average\ duration\ of\ exposure} \quad (22)$$

It is difficult to do such a calculation since not all new infections are reported but also to know how many susceptible were exposed. Therefore, we estimate the value of γ from two aspects, the characteristics of the infectious disease itself and the existing health status of the susceptible individuals.

However, in the real world it is unlikely that individuals could specify the locations where they had been in terms of exact geometric coordinates. It is conceivable however, that they could identify landmarks, such as buildings, museums, libraries, restaurants. where they had been. Therefore, the formula above could be approximately rewritten as follows:

$$F_{ab} = \sum_{i}^{n} \int_{Ts_i}^{Te_i} \frac{\gamma_s \gamma_d \left(t - Ts_i\right)^p}{R_i^q} \, dt \quad (23)$$

Where R_i represents radius of a circle that could encompass the areas of possible movement at common location i between two individuals.

Fab is referred to as infection index between two individuals A and B, who have a coexistence relationship at locations i. By normalizing F over all individuals (l,m € C) with coexistence relationships, we can obtain the relative transmission rate R_{ab} which defines the likelihood the disease could be transmitted between two individuals in Equation 18.

$$R_{ab} = \frac{F_{ab} - min\,(F(l,m))}{max\,(F(l,m)) - min\,(F(l,m))} \qquad (\forall(l, m \in C) \quad l \neq m) \qquad (24)$$

The following example illustrates this concept:

Consider two individuals and their activity information over two days in Table 7.

	Individual 1				Individual 2		
	Start Time	End Time	Activity & Place		Start Time	End Time	Activity & Place
1st Day	08:40	09:00	Walked to IE Dept		08:15	08:57	Drove to IE Dept
	10:02	10:08	Walked to UofL Book Store		10:05	10:12	Walked to UofL Library
	10:41	10:47	Walked to Library		10:55	10:58	Walked to Classroom A
	12:24	12:40	Walked to Qdoba for lunch		12:15	12:25	Walked to SAC for Lunch
	13:12	13:30	Returned to IE Dept		13:10	13:22	Returned to IE Dept
	17:10	17:32	Returned Home		16:30	16:47	Walked to SAC for Sports
					19:08	19:35	Drove Home
2nd Day	09:00	09:18	Walked to UofL Library		09:50	10:20	Drove to GE for Co-op
	12:30	12:47	SAC for Lunch		11:25	12:10	Left GE for KFC
	13:25	13:41	UofL Library		12:40	12:53	Drove back to GE
	16:40	17:02	Returned Home		15:14	15:38	Drove Home

	Common Place	Time	Duration	Radius
1st Day	IE Dept	09:00-10:02	62 mins	25
	UofL Library	10:47-10:55	8 mins	35
	IE Dept	13:30-16:30	180 mins	25

Table 7. Activity information of two individuals during two days

And by using ArcGIS software, we can generate the space-time paths of their activity information as shown in Figure 16. The red line represents the first individual's activity information and the blue one for the other individual. On the space-time path, there are several overlapping parts, which represents synchronous activities, where they are both are in close proximity. This, as noted previously, would factor into risk of getting infected.

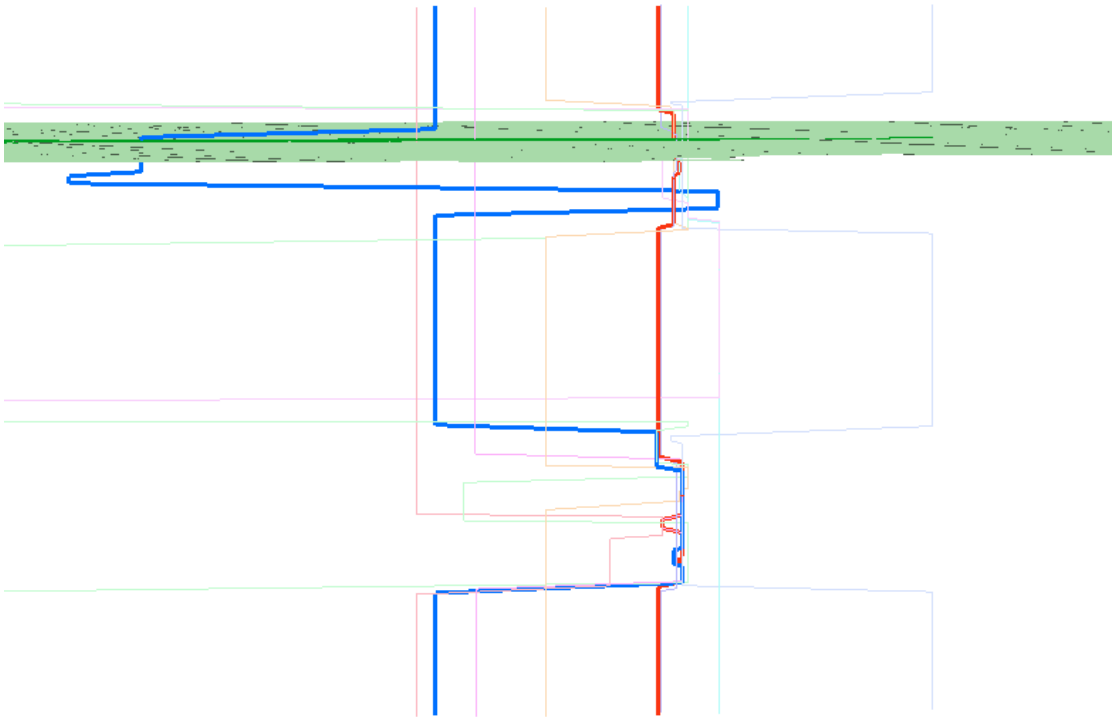


Figure 16. The Space-time paths of two individuals

Let $\gamma = \gamma_s \times \gamma_d$, so given different parameters such as γ, p and q, F_{ab} could have different values which are displayed in Table 8 and their data analyses are shown in Figure 17.

F_{ab}				
$\gamma = 0.2$	q\p	0.3	0.6	0.9
	1.5	1.3263	4.8136	18.4013
	2	0.2649	0.0622	3.6794
	2.5	0.0529	0.1924	0.7358
$\gamma = 0.4$	q\p	0.3	0.6	0.9
	1.5	2.6526	9.6272	36.8026
	2	0.5298	1.9244	7.3588
	2.5	0.1058	0.3848	1.4716
$\gamma = 0.6$	q\p	0.3	0.6	0.9
	1.5	3.9789	14.4408	55.2039
	2	0.7947	0.1866	11.0382
	2.5	0.1587	0.5772	2.2074
$\gamma = 0.8$	q\p	0.3	0.6	0.9
	1.5	5.3052	19.2544	73.6052
	2	1.0596	0.2488	14.7176

	2.5	0.2116	0.7696	2.9432
γ = 1.0	q\p	0.3	0.6	0.9
	1.5	6.6315	24.068	92.0065
	2	1.3245	0.311	18.397
	2.5	0.2645	0.962	3.679

Table 8. Different F_{ab} values with different parameters

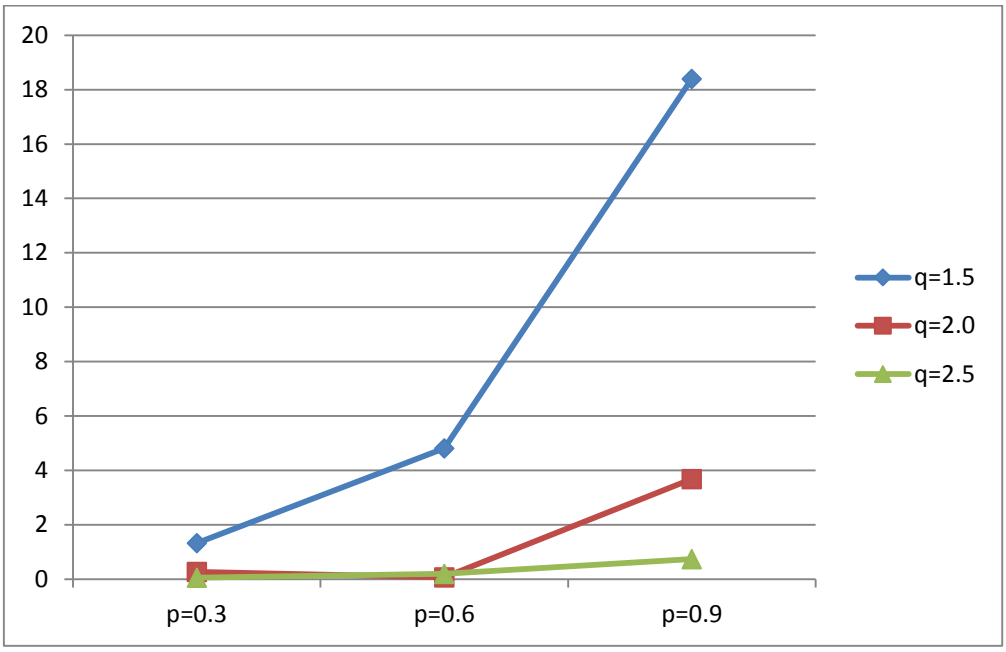


Figure 17. The Curve of different F_{ab} values with different parameters (γ = 0.2)

So from both Table 8 and Figure 17, we can see that for the same value of γ, as p increases, F_{ab} would be increased accordingly; Conversely, F_{ab} would decrease when q increases. For the same p and q, the risk would increase if γ increases. Intuitively this makes sense, since higher the infectivity of the disease, higher risk of getting infected.

4.2 Parameter γ_s and γ_d Computation

In the formula 11 discussed in section 4.1, the susceptible parameter γ_s could be related to multiply factors, especially the physical conditions and behaviors of the individuals.

Therefore, we will use both data mining methodologies and MIC to compute the approximate value of γ_s. And for γ_d, we will just normalize the R_0 values among common airborne diseases.

4.2.1 Datasets

Since the real world data sets are difficult to collect, we establish a reasonable dataset based on the academic statistical results from both journals and related organizational websites, which records the body physical conditions and the routine activities of 10000 different individuals with the format of comma-separated value (.CSV) files.

No.	Name	Data Type	Role	Range
1	ID	interval	ID	40001-50000
2	State	nominal	Rejected	US 50 States
3	Gender	binary	Input	0 or 1
4	Age	interval	Input	1-100
5	Month	interval	Rejected	1 to 12
6	Body Mass Index	Interval	Input	15-33
7	Smoker	binary	Input	0 or 1
8	Obesity	binary	Input	0 or 1
9	Diabetes	binary	Input	0 or 1
10	Asthma	binary	Input	0 or 1
11	Alcohol	binary	Input	0 or 1
12	Prescription Drugs	binary	Input	0 or 1
13	Illicit Drugs	binary	Input	0 or 1
14	Vegetarian	binary	Input	0 or 1
15	Exercise Rate	ordinal	Input	0 to 7
16	High Blood Pressure	binary	Input	0 or 1
17	Medical Care Rate	ordinal	Input	0 - 4
18	Pregnant	binary	Input	0 or 1
19	Family History	binary	Input	0 or 1
20	Allergy	binary	Input	0 or 1
21	Avg Working Hours	interval	Input	0 to 45
22	Infected	binary	Output	0 or 1

Table 9. Dataset variables

The variables in the dataset are explained as follows:

1) ID: Sample ID;

2) State: indicates which state each individual gets infected from;

3) Gender: 1 represents male and 0 represents female; Data was generated based on US 2010 Census Briefs: Male 49.2%, Female 50.8%;

4) Age: also based on US 2010 Census Briefs:

Age	Under 18 years	18 to 44 years	45 to 64 years	65 years and over
Percentage	24.00%	36.60%	26.40%	13%

Table 10. Age percentage of US population 2010

5) Month: month in which each individual got infected and might have some relationships with temperature and humidity;

6) Body Mass Index (BMI): Based on the US statistics in the year of 1999

Age	< 20	20-29	30-39	40-49	50-59	60+
BMI	15-28	20-28	20-30	22-30	22-32	22-33
BMI Median	22.72	25.05	25.77	25.94	26.51	26.70

Table 11. US BMI statistics by Age in 1999

7) Smoker: 1 is for Yes and 0 is for No. Data was established based on CDC statistics results in US.2010: 22% of adults aged 18-64 years and 9.5% of adults aged 65 years and older;

8) Obesity: 1 is for Yes and 0 is for No. Data was created based on CDC US. Obesity Trends: 9.5% for children and adolescents aged less than 19 years and 35.7% for adults aged 19 years and older;

9) Diabetes: 1 is for Yes and 0 is for No. Data was generated based on CDC statistics from 1980 through 2010: 12.3% for adults aged 45-65 years and 20.5% for adults aged 66 years and older;

10) Asthma: 1 is for Yes and 0 is for No. Data was established based on CDC statistical results in 2010: 9.4% for children aged less than 18 and 8.2% for adults aged 19 years and older.

11) Alcohol: 1 is for Yes and 0 is for No. 72% for adults aged 18-54 years and 59% for adults aged 55 years and older (Frank Newport, 2010)

12) Recent Prescription Drugs: 1 is for Yes and 0 is for No. 10% for children and adolescents aged less than 18 years, 30% for adults aged19-60 years and 30% for adults aged 61 years and older;

13) Illicit Drugs: 1 is for Yes and 0 is for No. Data was created based on the statistical results of National Institute on Drug Abuse (NIDA) in 2011: 2% for adults.

14) Vegetarian: 1 is for Yes and 0 is for No. 12% for children and adolescents aged less than 18 years, 7% for adults aged19-55 years and 14.4% for adults aged 56 years and older;

15) Exercise rate: How many times do individuals do exercise per week. 1-7 times

per week for the 23% adults aged 19-44 years old, 1-7 times per week for the 17% adults aged 45-64 years and 1-5 times per week for the 9.6% adults aged 65 years and older;

16) High Blood Pressure: 1 is for Yes and 0 is for No. Data was generated based on the statistical results by American Heart Association: 22% for the adults aged 20-54 years and 67% for the adults aged 55 years and older;

17) Medical Care Rate: how many times to get a physical examination per year. Never for 35% of the whole population, only once for 45% population, twice for 15% population, three times for 6% population and four times for 4% population;

18) Pregnant: 1 is for Yes and 0 is for No. 30% female adults aged 15-28 years;

19) Family History: 1 is for Yes and 0 is for No. 14% of the whole population;

20) Allergy: 1 is for Yes and 0 is for No. Approximate 10% of the whole US population have allergy based on CDC survey in 2008;

21) Avg Working Hours: average working hours per week. See Table 12 below:

Age	Unemployed or Retired	10-25 hours per week	33-45 hours per week
18-64	10%	30%	60%
65 and older	95%	3%	2%

Table 12. US． Average working hours by age

22) Infected: 1 is for Yes and 0 is for No. Seasonal flu & influenza rate: around 6.5% for the children and adolescents aged below 18 years old and 6.3% for the adults aged 18 years and older based on the statistical results by the American Lung Association in 2007.

4.2.2 Data Preparation

We will use the SAS Enterprise Miner 12.1 as the software tool to do the data mining process since the SAS Enterprise Miner streamlines the data mining process to create highly accurate predictive and descriptive models based on analysis of vast amounts of data.

In SAS Enterprise Miner, the data mining process has the following (SEMMA) steps:

1. Sample the data by creating one or more data sets. The sample should be large enough to contain significant information, yet small enough to process. This step includes the use of data preparation tools for data import, merge, append, and filter, as well as statistical sampling techniques.

2. Explore the data by searching for relationships, trends, and anomalies in order to gain understanding and ideas. This step includes the use of tools for statistical reporting and graphical exploration, variable selection methods, and variable clustering.

3. Modify the data by creating, selecting, and transforming the variables to focus the model selection process. This step includes the use of tools for defining

transformations, missing value handling, value recoding, and interactive binning.

4. Model the data by using the analytical tools to train a statistical or machine learning model to reliably predict a desired outcome. This step includes the use of techniques such as linear and logistic regression, decision trees, neural networks, partial least squares, LARS and LASSO, nearest neighbor, and importing models defined by other users or even outside SAS Enterprise Miner.

5. Assess the data by evaluating the usefulness and reliability of the findings from the data mining process. This step includes the use of tools for comparing models and computing new fit statistics, cutoff analysis, decision support, report generation, and score code management.

There are several models included in our data mining process.

The Input Data tool represents the data source that we choose for our data mining analysis and provides details (metadata) about the variables in the data source; while the File Import enables us to convert selected external flat files, spreadsheets, and database tables into a format that SAS Enterprise Miner recognizes as a data source.

The Data Partition tool enables us to partition data sets into training, test, and validation data sets. The training data set is used for preliminary model fitting. The validation data set is used to monitor and tune the model during estimation and is also used for model assessment. The test data set is an additional holdout data set that you can use for model

assessment. This tool uses simple random sampling, stratified random sampling, or cluster sampling to create partitioned data sets.

The Graph Explore tool is an advanced visualization tool that enables us to explore large volumes of data graphically to uncover patterns and trends and to reveal extreme values in the database. The tool creates a run-time sample of the input data source. We can use the Graph Explore node to interactively explore and analyze our data using graphs.

The MultiPlot tool is a visualization tool that enables us to explore large volumes of data graphically and it automatically creates bar charts and scatter plots for the input and target. Also, the code created by this tool can be used to create graphs in a batch environment.

The StatExplore tool is a multipurpose tool used to examine variable distributions and statistics in the data sets. The tool generates summarization statistics and we can use the StatExplore tool to select variables for analysis, for profiling clusters, and for predictive models; It can be also used to compute standard univariate distribution statistics, standard bivariate statistics by class target and class segment and correlation statistics for interval variables by interval input and target.

The Impute tool enables us to replace values for observations that have missing values. We can replace missing values for interval variables with the mean, median, midrange, mid-minimum spacing, or distribution-based replacement, or we can use a replacement M-estimator such as Tukey's biweight, Huber's, or Andrew's Wave. We can also estimate the replacement values for each interval input by using a tree-based imputation method. Missing values for class variables can be replaced with the most frequently occurring value, distribution-based replacement, tree-based imputation, or a constant.

The Transform Variables tool enables us to create new variables that are transformations of existing variables in our data. Transformations are useful if we want to improve the fit of a model to the data. For example, transformations can be used to stabilize variances, remove nonlinearity, improve additivity, and correct nonnormality in variables. The Transform Variables tool supports various transformation methods. The available methods depend on the type and the role of a variable.

The Control Point tool enables us to establish a control point to reduce the number of connections that are made in process flow diagrams. For example, suppose that three Input Data Source tools are to be connected to three modeling tools. If no Control Point tool is used, then nine connections are required to connect all of the Input Data Source tools to all of the modeling tools. However, if a Control Point tool is used, only six connections are required.

The Model Comparison tool provides a common framework for comparing models and predictions from any of the modeling tools. The comparison is based on the expected

and actual profits or losses that would result from implementing the model. The tool produces several charts that help to describe the usefulness of the model, such as lift charts and profit/loss charts.

The Score tool enables us to manage, edit, export, and execute scoring code that is generated from a trained model. Scoring is the generation of predicted values for a data set that might not contain a target variable. The Score tool generates and manages scoring formulas in the form of a single SAS DATA step, which can be used in most SAS environments even without the presence of SAS Enterprise Miner. The Score tool can also generate C score code and Java score code.

4.2.3 γ_s computation

As mentioned previously, γ_s is the parameter which describes the susceptibility for those who have not been infected yet. We believe that it is associated with people's demographic information and health status and therefore logistic regression and several data mining methodologies might be used to predict the likelihood of infection. In real world case, health care records in clinics or hospital would be very helpful for the infection prediction.

The estimation of transmission parameters has been problematic for diseases that rely predominantly on transmission of pathogens from person to person through small

infectious droplets (Wallinga et al. 2006). For example, age-specific transmission parameters determine how such respiratory agents will spread among different age groups in a human population. Estimating the values of these parameters is essential in planning an effective response to potentially devastating pandemics of influenza and in designing control strategies for diseases. In Wallinga's paper, the estimated age-specific transmission parameters suggested that school-aged children and young adults will experience the highest incidence of infection and will contribute most to further spread of infections during the initial phase of an emerging respiratory-spread epidemic in a completely susceptible population (Wallinga et al. 2006).

Similarly, from all the potential factors, we conduct regression models and several other data mining methodologies to determine which factors have important implications for controlling future outbreaks of respiratory-spread infectious agents. Specifically, these significant factors would be utilized to build a predictive model and estimate the susceptibility parameter γ_s. To accomplish this, it is necessary to choose a particular software system to carry out the computations. Although there are many good statistical packages for doing the logistic regression, SAS is certainly among the best in terms of the range of estimation methods, available features and options, efficiency and stability of the algorithms, and quality of the documentation.

4.2.4 γ_d computation

γ_d is a parameter associated with the characteristics of the infectious disease itself as well as the ability of the disease to infect susceptible individuals during the pandemics.

Therefore, we can use the basic reproduction rate R_0 in the compartmental SIR model to approximately estimate the value of γ_d. In epidemiology, the basic reproduction number R_0 describes the number of susceptible individuals one infected individual on average can infect over the course of the infectious period. Table 13 provides the R_0 values for a couple of common different infectious diseases. The initial R_0 value is a range and therefore the average R_0 is computed to obtain an exact value. By normalizing the average R_0 (divided by 20, this could vary in real world case), we can roughly get a value for γ_d, that would be used in the infection risk formula introduced previously.

Disease	Transmission	R_0	Average R_0	γ_d (/20)
Measles	Airborne	12 – 18	15	0.75
Pertussis	Airborne droplet	12 – 17	14.5	0.725
Rubella	Airborne droplet	5 – 7	6	0.3
Mumps	Airborne droplet	4 – 7	5.5	0.275
SARS	Airborne droplet	2 – 5	3.5	0.175
Influenza	Airborne droplet	2 – 3	2.5	0.125
Diphtheria	Saliva	6 – 7	6.5	0.325
Smallpox	Social contact	5 – 7	6	0.3
Polio	Fecal-oral route	5 – 7	6	0.3
HIV/AIDS	Sexual contact	2 – 5	3.5	0.175

Table 13. R_0 values for different diseases

4.3 Process Flowchart Summary

The essence of the decision support system for risk analysis, is the assessment of the four parameters (γ_s, γ_d, F_{ab}, R_{ij})defined previously and summarized below:

(1) Susceptible parameter γ_s and disease parameter γ_d: data mining tools such as Decision Trees, Artificial Neural Network, and Regression are used to compute the susceptible parameter γ_s and the disease parameter γ_d in the infection index formula.

(2) Infection Index F_{ab}: Provided with information from both susceptible individuals and infected individuals, the decision support system would use the spatio-temporal formula to calculate the infection index F_{ab} after the susceptible parameter γs and disease parameter γd are obtained empirically.

(3) Transmission Rate R_{ij}: By normalizing the infection index F_{ab}, the transmission rate Rij between two specific individuals would be computed in the decision support system.

(4) Risk Analysis: the centrality value for each individual in the network would be evaluated by the method proposed in this paper. Conceptually all individuals could then be ranked or color coded different colors according to their centrality values. As explained in the previous section, this information could be utilized in a decision support system to effectively and efficiently mitigate the spread of a virulent disease. Figure 18 displays the process flow chart of the decision support system.

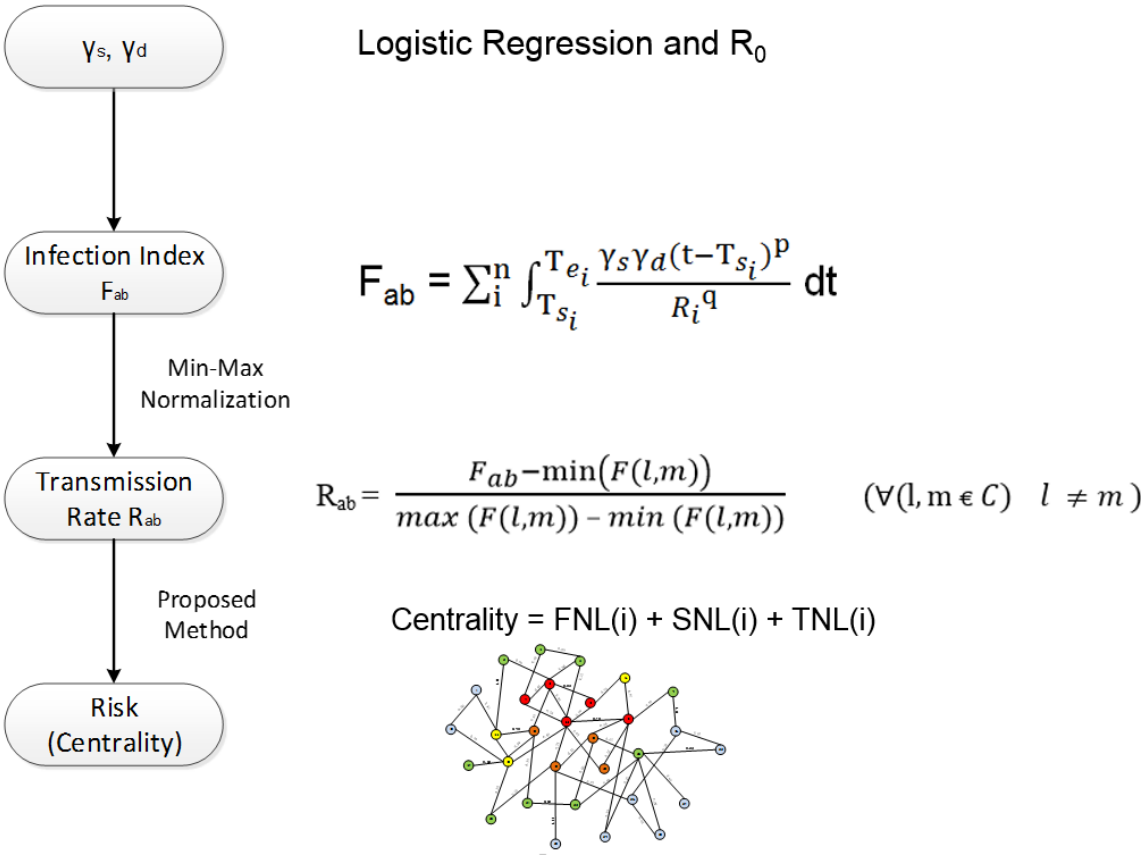


Figure 18. The working flowchart of contact network analysis with spatiotemporal information

Again, by normalizing the infection index F_{ab}, we could obtain the transmission rate between two individuals which could be used in the proposed centrality computation method discussed in the previous section. According to the centrality values, each individual would be classified into different color risk zones and given different priorities for the vaccination.

4.4 Decision Support System

Generally speaking, a decision support system is a computerized information system that supports business and organizational decision-making activities. It is intended to help decision makers compile useful information from raw data, documents, personal

knowledge and business models to identify and solve problems and make decisions. While in the field of healthcare, clinical decision-support systems (CDSSs) are computer programs that are designed to provide expert support for health professionals making clinical decisions (Musen et al. 2001). It is an interactive system used to give clinical advice for the patients based on the existing patient data.

The core of the spatio-temporal analysis is the decision support system since it compares the information provided by both the susceptible people and infected people and gives constructive suggestions to both the public health organizations and the general population. All these procedures discussed previously are operated by the decision support system in Figure 19 which has the main processes.

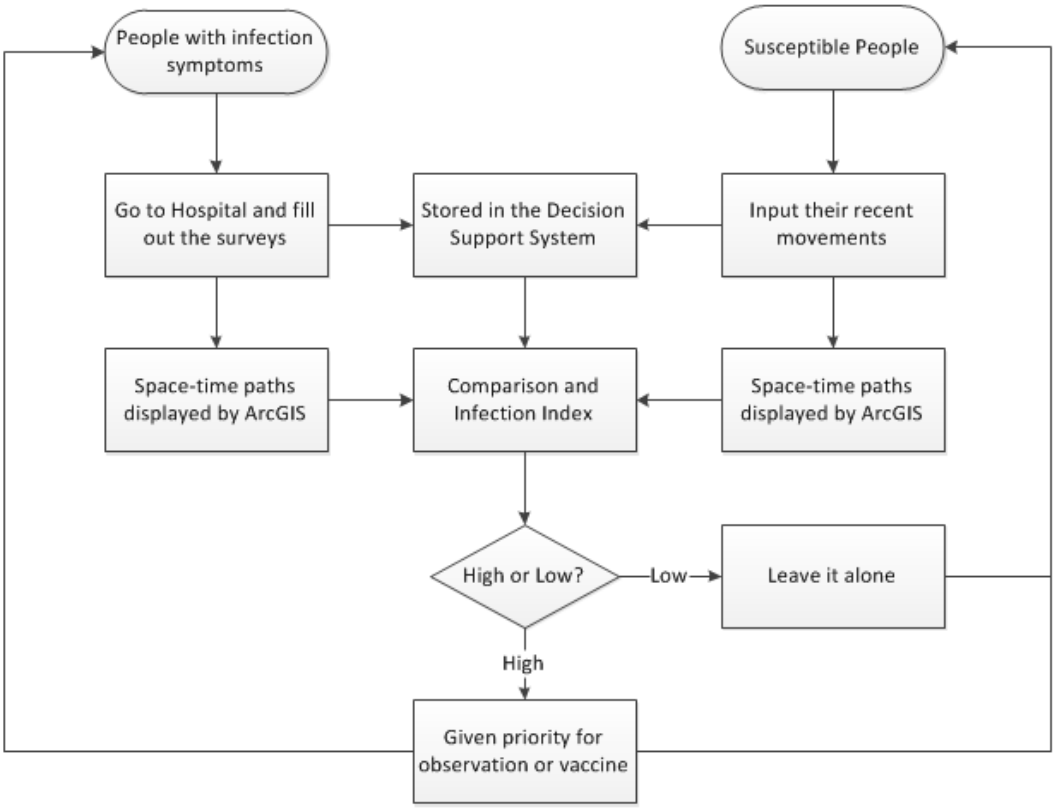


Figure 19. Decision support system

(1) The people who seem to have the infection symptoms go to the hospitals spontaneously and then they have to fill out the surveys provided by the hospital. The surveys contain some questions about the information of where they have been to recently and when.

(2) Then the geographic information of these infected individuals could be obtained by space-time paths, which would be displayed in ArcGIS environment.

(3) For those who care more about their risk of getting infected and mostly susceptible or not detected to be infected yet, they can input their information about their recent tracks into the decision support system. Then they will get a feedback or report about their infection risk based on the infection index.

(4) If the risk is high, they would be given priority for observation or vaccines. Otherwise, they could do the routine activities as normal but still need to keep an eye on the spatial-temporal trend of the disease spread.

4.5 Comprehensive Application

As a matter of fact, not only people with symptoms could be used for detecting the potential pandemics, some other potential "sensors" could also have this functionality. For example, the "Louisville Connectors" could be considered as one set of sensors for the Louisville metropolitan area. The Louisville connectors are a diverse group of 128 individuals ranging in age from 28 to 71, from 5,500 nominations submitted by people through Louisville and Southern Indiana and they could be considered as "central individuals" to monitor. Individuals identified by sources, such as, hospital laboratories, and "First Watch" with symptoms of virulent infections can also answer questionnaires.

Additionally, The Early Aberration Reporting System (EARS) of the Centers for Disease Control and Prevention (CDC) allows the analysis of public health surveillance data using available aberration detection methods.

In a similar fashion, infection spread within a hospital or treatment facility could be mitigated by tracking spatial-temporal movement and contact of infected patients, especially those with long hospital stays. This approach could also be extended to equipment used on infected patients, which hospitals typically do not track. One way to accomplish this would be through the use of RFID tags that track the movement of patients, hospital staff and equipment and their spatial-temporal interactions with an infected person. This would enable the defining a risk level for all hospital residents who are susceptible to infection, which would allow efficient prophylactic measures to be appropriately scheduled and adopted.

CHAPTER 5 EXPERIMENTAL EVALUATION AND ANALYSIS

5.1 Experimental Evaluation

5.1.1 Explore dataset and descriptive statistics

First of all, we need to understand the data, draw data graphs and perform basic data summaries including means, standard deviation etc. The exploration of data is relatively straightforward in SAS since they can be obtained by using a couple of SAS components. Here we are mainly using the StatExplore node and Multiplot node to realize the data exploration.

The results of StatExplore node include the data summary statistics for both class (category) variables and interval variables.

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Alcohol	INPUT	2	0	0	52.04	1	47.96
TRAIN	Allergy	INPUT	2	0	0	90.08	1	9.92
TRAIN	Asthma	INPUT	2	0	0	91.22	1	8.78
TRAIN	Diabetes	INPUT	2	0	0	94.25	1	5.75
TRAIN	Family_His	INPUT	2	0	0	86.64	1	13.36
TRAIN	Gender	INPUT	2	0	0	50.81	1	49.19
TRAIN	HighBloodPressure	INPUT	2	0	0	73.39	1	26.61
TRAIN	Illicit_Drugs	INPUT	2	0	0	98.36	1	1.64
TRAIN	Obesity	INPUT	2	0	0	72.07	1	27.93
TRAIN	Pregnant	INPUT	2	0	0	97.41	1	2.59
TRAIN	Recent_Perscription_Drugs	INPUT	2	0	0	75.71	1	24.29
TRAIN	Smoker	INPUT	2	0	0	85.96	1	14.04
TRAIN	Vegetarian	INPUT	2	0	0	90.42	1	9.58
TRAIN	Infected	TARGET	2	0	0	93.73	1	6.27

Table 14. Class variable summary statistics

From the class variable table, we can see that there are totally 13 input variables and 1 target variables, most of which are binary variables. There are no missing values for the training set which is good for predictive modeling since our data is generated according to the pubic statistics from authority agencies, such as US census and CDC. But as a matter of fact, the raw data could be very messy in the real world and we need to perform data preprocessing before the predictive modeling, such as data cleaning, data reduction and outlier detection. According to the class variable table, almost half of the whole sample population have an experience of drinking alcohol. Only a small percentage have allergy, asthma, diabetes and also a small percentage are pregnant, smokers and vegetarians. Also, currently 6.27 % of the sample population have already been infected to have the influenza, which indicates that the number of infected people could be very large if it is true for the whole population.

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Age	INPUT	38.1263	24.59634	10000	0	1	36	99	0.497655	-0.48757
BMI	INPUT	24.792	3.859019	10000	0	15	25	33	-0.29484	-0.26492
Exercise_Rate	INPUT	0.5241	1.47419	10000	0	0	0	7	2.886691	7.361191
Working_Hours	INPUT	17.8347	17.63538	10000	0	0	18	45	0.205036	-1.66753

Table 15. Interval variable summary statistics

There are 4 interval variables, which are all input variables also with no missing values. From the interval variable table, we can see that the sample population have an age range between 1 and 99, and their average age is 38. Also, the average BMI for the sample population is approximately 24.8, which is within the normal range. On average, people do not do exercise that much, less than once per week. The average working hours per week is 17.8 hours and some people work more than 40 hours per week.

5.1.2 Analyze the sample dataset

In order to explore the data well, we need to know about the domain and therefore it is essential to discuss the sample dataset into more details. The sample dataset contains all the demographic information and health condition status information of individuals with a total of 20 variables and 10000 observations. Our goal is to determine which factors among the potential demographic and health condition information have a significant impact on the likelihood of infection for individuals and what that infection risk is for one susceptible individual. The variable list for the sample dataset is given in Table 16.

Variable Name	Variable Description
Age	People's Age
Alcohol	Check if an individual has an experience of drinking alcohol
Allergy	Check if an individual has allergy
Asthma	Check if an individual has asthma
BMI	People's BMI (Weight over Height)
Diabetes	Check if an individual has diabetes
Exercise_Rate	Describes how often an individual does exercises
Famaily_His	Check if an individual has a family history infection
Gender	People's gender
HighBloodPressure	Check if an individual has high blood pressure
Illicit Drugs	Check if an individual has an experience of illicit drugs
Infected	Check if an individual has been already infected
Obesity	Check if an individual has obesity
Pregnant	Check if a female individual is pregnant
Recent_Perscription_Drugs	Check if an individual has an experience of prescription drugs recently
Smoker	Check if an individual is a smoker
Vegetarian	Check if an individual is a vegetarian
Working_Hours	How many hours an individual works per week

Table 16. Variable names and variable description

Most variables are identified as binary variables and only Age, BMI, Exercise_Rate and Working_Hours are identified as interval variables. In order to further investigate these variables, we will use StatExplore and MultiPlot nodes in the SAS Enterprise Miner.

Also, SAS/Insight is the main module of the explorative data analysis, which could be reached by typing “insight” in the command box of SAS toolbar or selecting system menu “Solutions – Analysis – Interactive Data Analysis”.

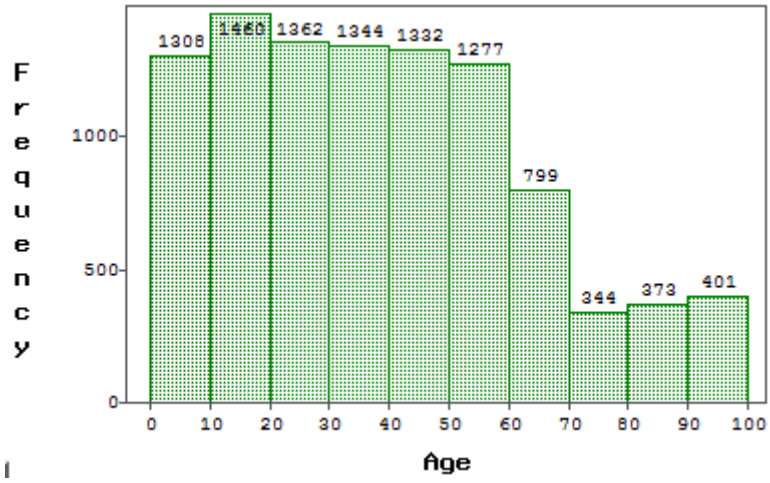


Figure 20. Distribution of Age

Data could be displayed through diagrams and sometimes we also need to extract data summarization information to describe the characteristics of the entire dataset. These abstract and summary data could be collected from the raw dataset after summarization, which utilizes relatively small amount of variables and metrics to represent the whole data information. Meanwhile, these extracted information computed from the sample data is referred as to the sample statistics. Since different sample dataset could be obtained from the population through different ways, the sample statistics varies according to the different sample dataset even for the same population. Therefore, the sample statistics is mostly uncertain but it is known and it could be measured from several perspectives, such as converge trend, dispersion, distribution and data shape.

1. Age

Moments			
N	10000	Sum Weights	10000
Mean	38.1263	Sum Observations	381263
Std Deviation	24.5963421	Variance	604.980046
Skewness	0.49765464	Kurtosis	-0.4875661
Uncorrected SS	20585343	Corrected SS	6049195.48
Coeff Variation	64.512796	Std Error Mean	0.24596342

Basic Statistical Measures			
Location		Variability	
Mean	38.12630	Std Deviation	24.59634
Median	36.00000	Variance	604.98005
Mode	14.00000	Range	98.00000
		Interquartile Range	37.00000

Analysis Variable : Age Age								
Minimum	Maximum	Median	Lower Quartile	Upper Quartile	Range	Coeff of Variation	Skewness	Kurtosis
1.0000000	99.0000000	36.0000000	18.0000000	55.0000000	98.0000000	64.5127960	0.4976546	-0.4875661

Table 17. Statistical results for the variable Age

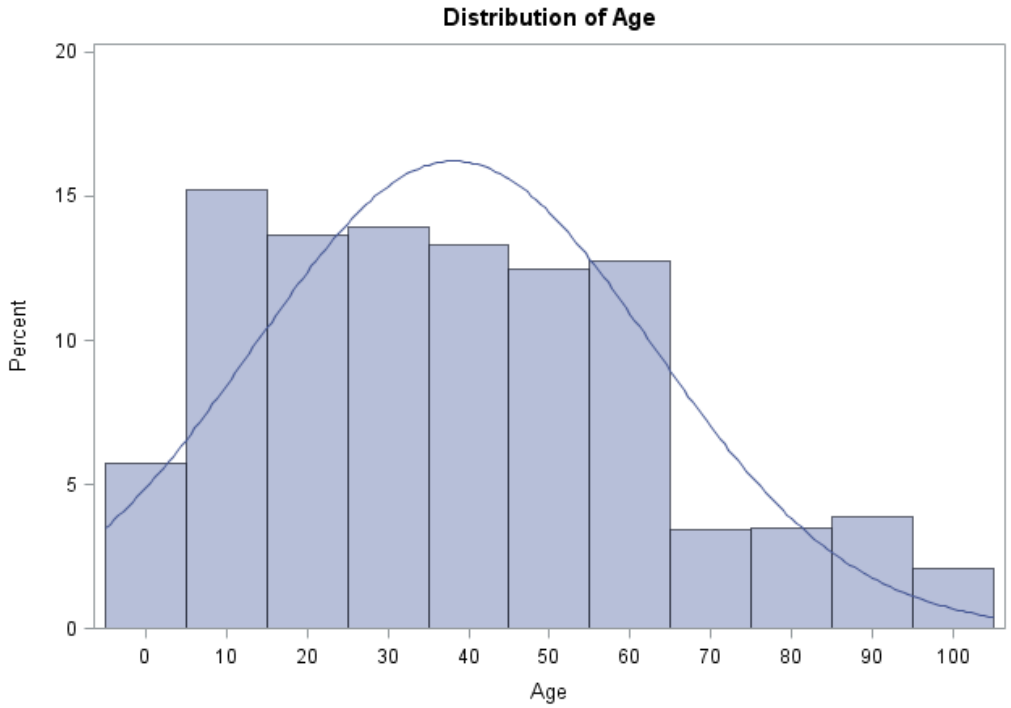


Figure 21. The distribution and trend of variable age

The Age variables statistics table indicates that there are 10000 individuals in the sample dataset at the age ranging from 1 to 99 and their average age is approximately 38 with a standard deviation of 24.6, which leads to a high coefficient of variation 64.5%. Also, the histogram of age distribution illustrates that the majority in the sample have an age between 10 and 60 and the minority are children and senior citizens. It is right skewed since the skewness is 0.498 > 0 and the peak value is lower than standard normal distribution because the Kurtosis is -0.487 < 0.

2. BMI

Moments			
N	10000	Sum Weights	10000
Mean	24.792	Sum Observations	247920
Std Deviation	3.85901868	Variance	14.8920252
Skewness	-0.2948391	Kurtosis	-0.2649187
Uncorrected SS	6295338	Corrected SS	148905.36
Coeff Variation	15.5655804	Std Error Mean	0.03859019

Basic Statistical Measures			
Location		Variability	
Mean	24.79200	Std Deviation	3.85902
Median	25.00000	Variance	14.89203
Mode	24.00000	Range	18.00000
		Interquartile Range	6.00000

Analysis Variable : BMI BMI								
Minimum	Maximum	Median	Lower Quartile	Upper Quartile	Range	Coeff of Variation	Skewness	Kurtosis
15.0000000	33.0000000	25.0000000	22.0000000	28.0000000	18.0000000	15.5655804	-0.2948391	-0.2649187

Table 18. Statistical results for the variable BMI

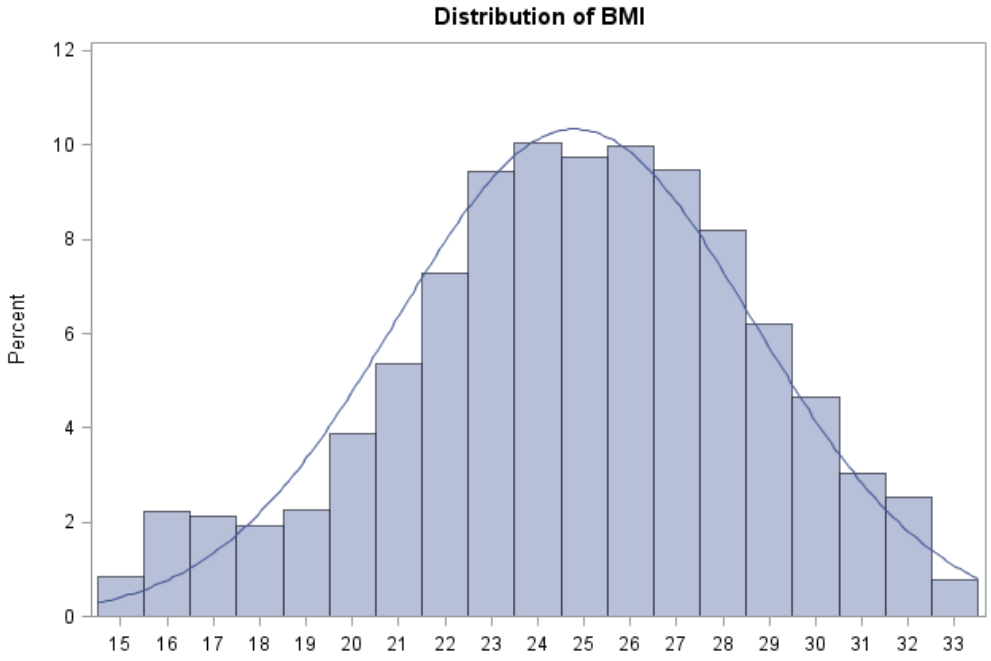


Figure 22. Distribution of BMI

3. Exercise Rate

Moments			
N	10000	Sum Weights	10000
Mean	0.5241	Sum Observations	5241
Std Deviation	1.47419012	Variance	2.17323651
Skewness	2.88669096	Kurtosis	7.36119069
Uncorrected SS	24477	Corrected SS	21730.1919
Coeff Variation	281.280313	Std Error Mean	0.0147419

Basic Statistical Measures			
Location		Variability	
Mean	0.524100	Std Deviation	1.47419
Median	0.000000	Variance	2.17324
Mode	0.000000	Range	7.00000
		Interquartile Range	0

Analysis Variable : Exercise_Rate Exercise Rate								
Minimum	Maximum	Median	Lower Quartile	Upper Quartile	Range	Coeff of Variation	Skewness	Kurtosis
1.0000000	7.0000000	5.0000000	3.0000000	6.0000000	6.0000000	34.6802864	-0.3161132	-0.8504092

Table 19. Statistical results for the variable Exercise_Rate

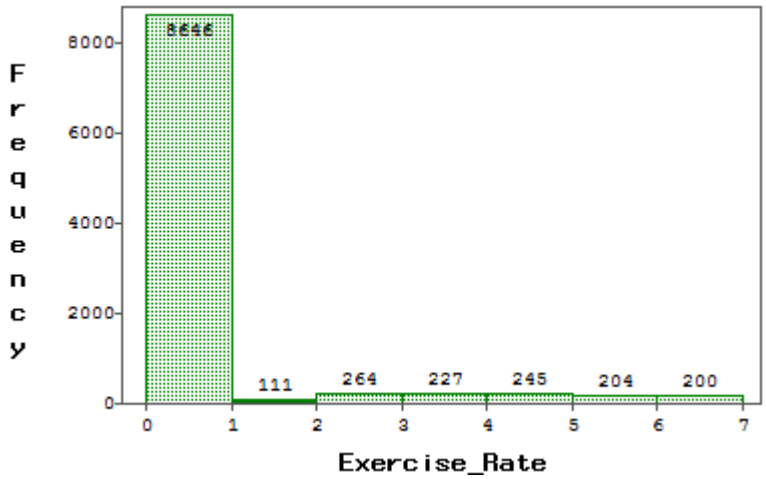


Figure 23. Distribution of exercise rate

4. Average Working Hours

Moments			
N	10000	Sum Weights	10000
Mean	17.8347	Sum Observations	178347
Std Deviation	17.6353757	Variance	311.006477
Skewness	0.20503586	Kurtosis	-1.6675255
Uncorrected SS	6290519	Corrected SS	3109753.76
Coeff Variation	98.8823794	Std Error Mean	0.17635376

Basic Statistical Measures			
Location		Variability	
Mean	17.83470	Std Deviation	17.63538
Median	18.00000	Variance	311.00648
Mode	0.00000	Range	45.00000
		Interquartile Range	37.00000

Table 20. Statistical results for the variable Average Working Hours

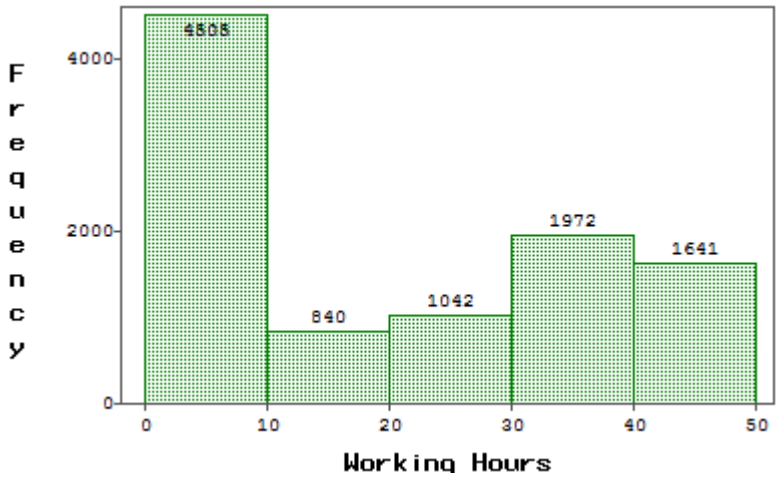


Figure 24. Distribution of Working Hours

5. All other Category Variables

Variable	Value	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Gender	0	5081	50.81	5081	50.81
	1	4919	49.19	10000	100.00
Smoker	0	8596	85.96	8596	85.96
	1	1404	14.04	10000	100.00
Obesity	0	7207	72.07	7207	72.07
	1	2793	27.93	10000	100.00
Diabetes	0	9425	94.25	9425	94.25
	1	575	5.75	10000	100.00
Asthma	0	9122	91.22	9122	91.22
	1	878	8.78	10000	100.00
Alcohol	0	5204	52.04	5204	52.04
	1	4796	47.96	10000	100.00
Recent_ Prescription_Drugs	0	7571	75.71	7571	75.71
	1	2429	24.29	10000	100.00
Illicit_Drugs	0	9836	98.36	9836	98.36
	1	164	1.64	10000	100.00
Vegetarian	0	9042	90.42	9042	90.42
	1	958	9.58	10000	100.00
High_Blood_Pressure	0	7339	73.39	7339	73.39
	1	2661	26.61	10000	100.00
Pregnant	0	9741	97.41	9741	97.41
	1	259	2.59	10000	100.00
Family_His	0	8664	86.64	8664	86.64
	1	1336	13.36	10000	100.00
Allergy	0	9008	90.08	9008	90.08
	1	992	9.92	10000	100.00
Infected	0	9373	93.73	9373	93.73
	1	627	6.27	10000	100.00

Table 21. All other categorical variables

The MEANS procedure provides data summarization tools to compute descriptive statistics for variables across all observations and within groups of observations.

In data mining, we try to find patterns in the data and it is always possible to make it. However, pattern recognition with validation could be a reason that data mining as a method was often disparaged in statistical training. Therefore, it is strongly recommended that the raw data is partitioned routinely into three datasets: training, validation and testing. Specifically, the training dataset is used to develop the model and the validation datasets iteratively ensure that the developed model fits a fresh dataset. Once the model is completed, the testing datasets make a final comparison.

For a given output variable (or target variable), the accuracy of the final developed model is initially judged by misclassification rate and misclassification occurs when the predicted output value is not equal to the actual output value. Also, there are many different models that can be used and compared to investigate the data instead of choosing just one model to define a specific p-value. These assessment methods have been developed to make these comparisons using the training, validation and testing methodology.

Another important component of data mining in SAS is the ability to score the data. The predicted value is related to the actual value through scoring and the closeness of one to the other can be examined by using other statistical techniques. Also, scoring is particularly important when examining the likelihood of getting infected for an individual before the pandemics. In this case, scoring assigns a level of infection risk

to a susceptible individual such that the pandemic response team could control the spread of infectious disease effectively.

5.1.3 Data Mining Process

According to the SEMMA principle discussed previously, the brief steps of the data mining process are as follows:

1) Open the SAS Enterprise Miner WorkStation 7.1 and in the Welcome to Enterprise Miner Window, create a New Project and specify the SAS Server Directory in which SAS data sets and other files that are generated by the project will be stored.

2) Create a Diagram with the same name of the project. Create the File Import Node under Sample by dragging it into the Text workstation. Select this Import Node and in the Property Panel click on the ellipses that represent Import File to import the original training data set. And right click on the File Import Node and select run to make sure that the original data is successfully imported.

3) Create both the StatExplore and Multiplot Nodes under Explore also by dragging it into the Text workstation and connect it to the previous nodes. Select the StatExpore node and click on the value of Interval Variables and select Yes from the drop-down menu that appears. And then Select the MultiPlot node, in the Properties Panel, click on the value of Type of Charts and select Both from the drop-down menu that appears. Finally right-click the Multiplot node, and select Run from the resulting menu. And we could see the Results in the Train Graphs window which displays a bar chart and a scatter plot for each variable.

4) Create Data Partition node under Sample also by dragging it into the Text workstation and connect it to the previous nodes. Select the Data Partition node, and click on the value of Training and enter 70.0, similarly click on the value of Validation and enter 30.0, finally click on the value of Test, and enter 0.0. Run the Data Partition node and the results are as follows:

Partition Summary

Type	Data Set	Number of Observations
DATA	EMWS4.Ids_DATA	10000
TRAIN	EMWS4.Part_TRAIN	6998
VALIDATE	EMWS4.Part_VALIDATE	3002

Table 22. Partition summary

5) Create the Replacement node under Modify, drag it into the Diagram Workspace and connect it to the previous nodes. And in the Properties Panel set the Default Limits Method as Standard Deviations from the Mean in order to reduce the missing values.

6) Create the Decision Tree node under Model and drag it into the Diagram Workspace and connect it to the previous nodes. Right-Click the Decision Tree node and rename the node as Interactive Decision Tree. In the Properties Panel, click on the ellipses that represent the value of Interactive. Select the root node and then from the Action menu select Split Node. The Split Node window appears that lists the candidate splitting rules ranked by logworth (-Log(p)). Since the Smoker variable has the highest logworth, it is first split and the tree now has two additional nodes. Repeat this process for each newly generated node until the logworth value for variable equals to 0. Figure 25 displays the

Leaf Statistics results and the structure of the Decision Tree is illustrated in Appendix 4.

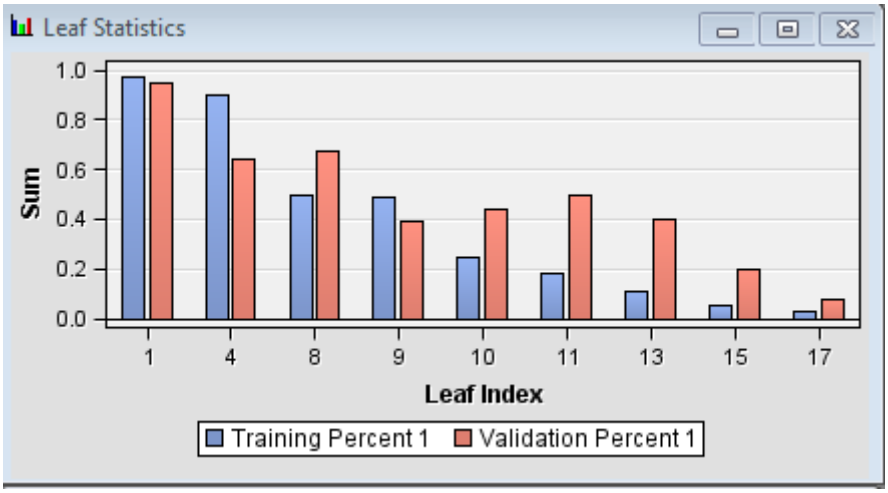


Figure 25. The Leaf Statistics

7) Create the Impute node under Modify and drag them into the Diagram Workspace and connect them to the previous nodes. For class variable, click on the value of Default Input Method and select Tree Surrogate from the drop-down menu that appears; for interval variables, click on the value of Default Input Method and select Median from the drop-down menu that appears. In our case, the values of missing interval variables are replaced by the median of the nonmissing values. This statistic is less sensitive to extreme values than the mean or midrange and is therefore useful for imputation of missing values from skewed distributions.

8) Create both the Variable Selection node and AutoNeural nodes and drag them into the Diagram Workspace and connect them to the previous nodes. Click on the value of the model option Architecture and select Cascade from the drop-down menu that appears. This action causes SAS Enterprise Miner 7.1 to train

only cascade network models. Click on the value of the model option Train Action and select Search. This action causes SAS Enterprise Miner 7.1 to perform a search to find the best of the candidate network models.

Classification Table

Data Role=TRAIN Target Variable=VAR5001

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
-1	-1	80.7692	35.0975	126	17.5243
1	-1	19.2308	8.3333	30	4.1725
-1	1	41.3854	64.9025	233	32.4061
1	1	58.6146	91.6667	330	45.8971

Data Role=VALIDATE Target Variable=VAR5001

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
-1	-1	65.6566	26.9710	65	13.5135
1	-1	34.3434	14.1667	34	7.0686
-1	1	46.0733	73.0290	176	36.5904
1	1	53.9267	85.8333	206	42.8274

Figure 26. Results of AutoNeural model

9) Create the Regression node under Model and drag it into the Diagram Workspace and connect it to the previous nodes. Click on the Selection Model property in the Model Selection subgroup and select Stepwise from the drop-down menu that appears. This specification causes SAS Enterprise Miner to use stepwise variable selection to build the logistic regression model. Make all the other settings by default and run the model.

10) Create two Neural Network nodes under Model by dragging them into the Text workstation and connect them to the previous nodes. Rename the nodes as MLP and GLIM respectively. In the Properties Panel, click on the ellipses that represent the value of Network. Select the root node and then from the Action menu select Split Node. Change the Architecture property as Multilayer Perceptron and Generalized Linear Model respectively for MLP and GLIM nodes from the Network menu that appears. This selection enables the network to have connections directly between the inputs and the outputs in addition to connections via the hidden units. Click on the value of Number of Hidden Units and enter 5. This case trains a multilayer perceptron neural network with five units on the hidden layer.

11) Create the DMNeural node under Model and drag it into the Diagram Workspace and connect it to the previous nodes. Make all the setting by default and run the model.

12) Create both the Control Point node under Utility and Model Comparison node under Assess and drag them into the Diagram Workspace and connect them to the previous nodes in order to compare models and select a champion model, which according to an evaluation criterion performs best in the validation data. Control Point nodes enable you to better organize your process flow diagram. These nodes do not perform calculations; they simply pass data from preceding nodes to subsequent nodes. Finally run the Model Comparison node and display the results. The Results indicate that the AutoNeural Network is the champion model.

13) Stop Criteria: Could be specified in each model and if the criteria are met, the process will automatically stop. Or we can set a limited running time and as long as the time is reached, the data mining process would also come to an end.

The whole data mining process is displayed in Figure27:

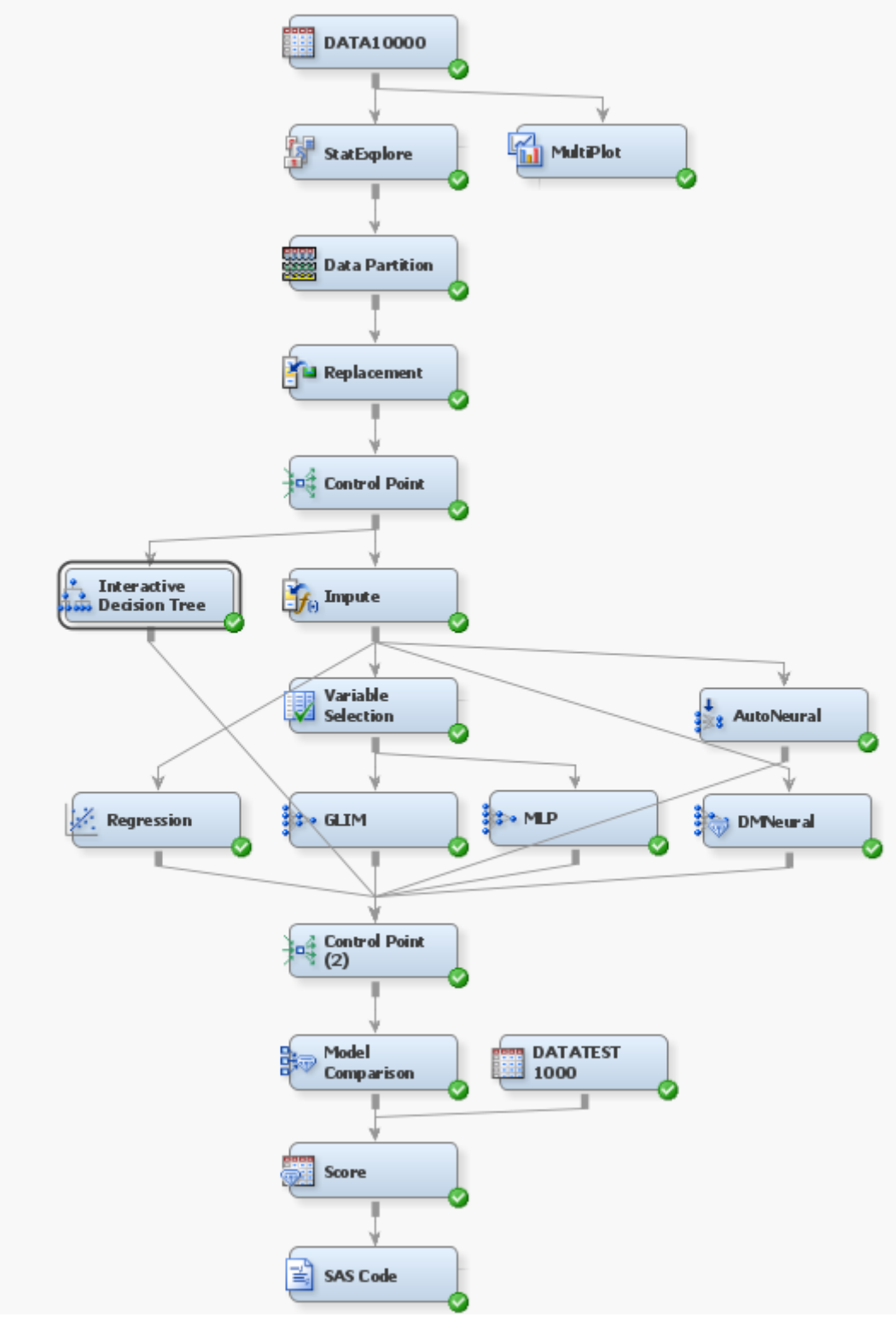


Figure 27. Process flow diagram of data mining process

5.1.4 Data Visualization

Data visualization is the presentation of data in a pictorial or graphical format. For centuries, people have depended on visual representations such as charts and maps to understand information more easily and quickly. As more and more data is collected and analyzed, decision makers at all levels welcome data visualization software that enables us to see analytical results presented visually, find relevance among the millions of variables, communicate concepts and hypotheses to others, and even predict the future. Because of the way the human brain processes information, it is faster for people to grasp the meaning of many data points when they are displayed in charts and graphs rather than poring over piles of spreadsheets or reading pages and pages of reports. Therefore, we will utilize several data visualization tools to explore our data and give us a direct way to understand how our data looks like.

Bar charts are most commonly used for comparing the quantities of different categories or groups. Values of a category are represented using the bars, and they can be configured with either vertical or horizontal bars with the length or height of each bar representing the value. When values are distinct enough that differences in the bars can be detected by the human eye, we can use a simple bar chart. However, when the values (bars) are very close together or there are large numbers of values (bars) that need to be displayed, it becomes more difficult to compare the bars to each other. Figure 28 demonstrates the number of infected individuals by their age group.

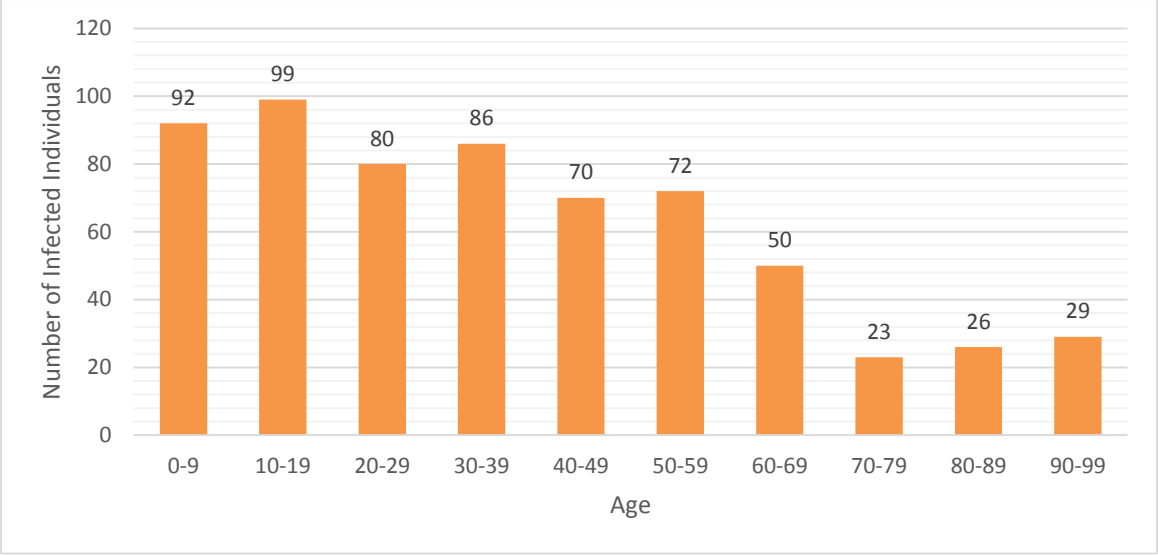


Figure 28. Number of infected individuals by age

A line graph, or line chart, shows the relationship of one variable to another. They are most often used to track changes or trends over time. Line charts are also useful when comparing multiple items over the same time period. The stacking lines are used to compare the trend or individual values for several variables. Bar charts and line graphs are more appropriate for continuous data rather than the categorical data.

Pie charts are most effective when there are limited components and when text and percentages are included to describe the content. There is much debate around the value of pie charts, which are used to compare the parts of a whole. However, they can be difficult to interpret because the human eye has a hard time estimating areas and comparing visual angles. Another challenge with using a pie chart for analysis is that it is difficult to compare slices of the pie that are similar in size but not located next to each other. If you do use pie charts, they are most effective when there are limited components and when text and percentages are included to describe the content. Figure

29 illustrates the average weekly working hours and their percentage of the whole sample population.

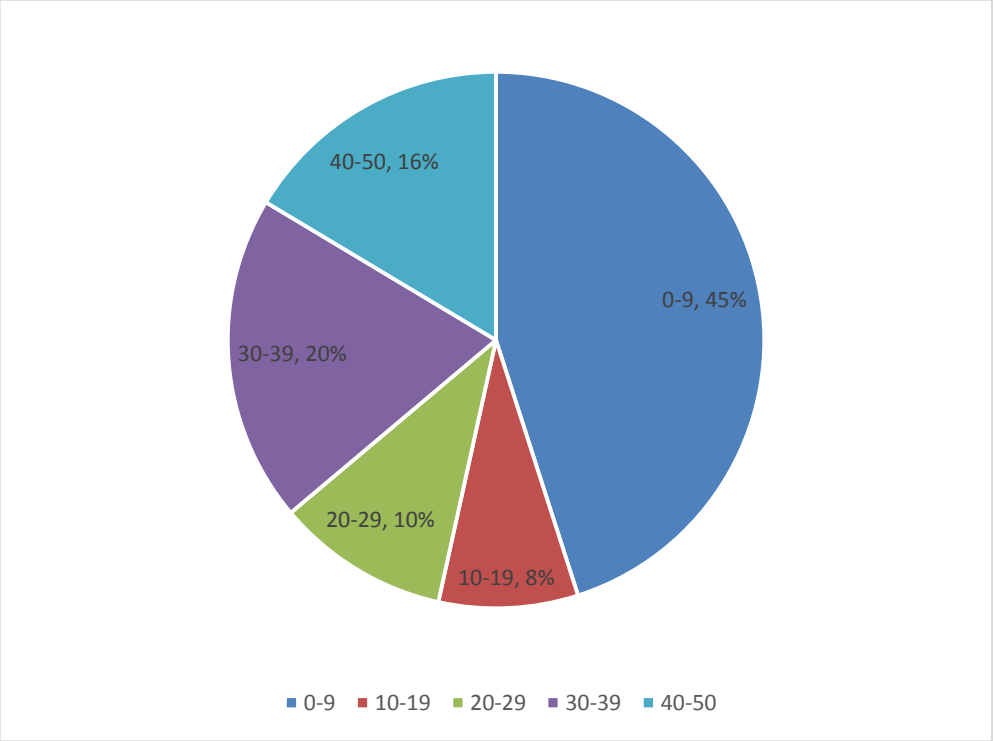


Figure 29. Pie chart of population percentage of average weekly working hours

A bubble plot is a variation of a scatter plot in which the markers are replaced with bubbles. In a bubble plot, each bubble represents an observation. The location of the bubble represents the value for two measured axes; the size of the bubble represents the value for a third measure. These plots are useful for data sets with dozens to hundreds of values or when the values differ by several orders of magnitude. We can also use a bubble plot if we want specific values to be represented by different bubble sizes. Animated bubble plots are a good way to display changing data over time.

Scatter plots are useful for examining the relationship, or correlations, between X and Y variables. Variables are said to be correlated if they have a dependency on, or are somehow influenced by, each other. For example, “profit” is often related to “revenue”

and the relationship that exists might be that as revenue increases profit also increases (a positive correlation). A scatter plot is a good way to visualize these relationships in data. In a scatter plot, you can also apply statistical analysis with correlation and regression. Correlation identifies the degree of statistical correlation between the variables in the plot. Regression plots a model of the relationship between the variables in the plot.

A box plot is a graphical display of five statistics (the minimum, lower quartile, median, upper quartile and maximum) that summarize the distribution of a set of data. The lower quartile (25th percentile) is represented by the lower edge of the box, and the upper quartile (75th percentile) is represented by the upper edge of the box. The median (50th percentile) is represented by a central line that divides the box into sections. Extreme values are represented by whiskers that extend out from the edges of the box. Usually, these display well when using big data. Often, box plots are used to understand the outliers in the data.

Moreover, server other data visualization tools might be also useful. For example, Matrix Graph allows us to see more variables simultaneously and shows a series of scatter plots that examine the variables two at a time; Lattice Graph can be used with a variety of different plots.

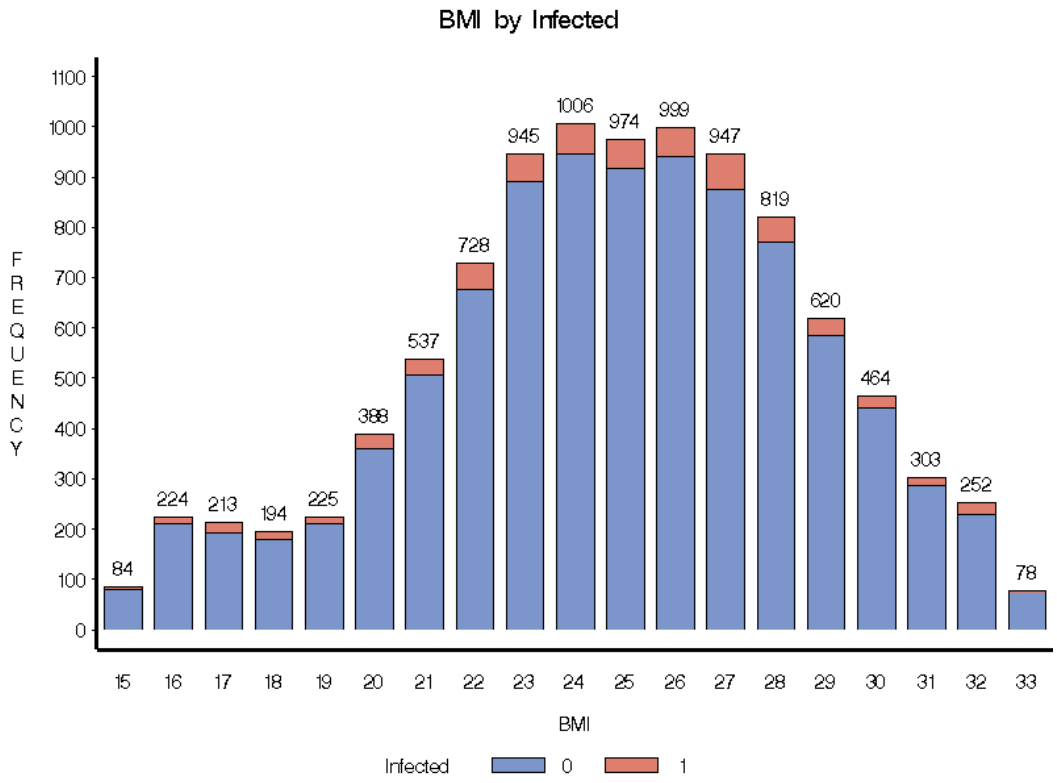


Figure 30. Distribution of BMI by Infected

5.2 Result Analysis

5.2.1 Predictive modeling

Predictive modeling is an extension of statistical linear models. Since typically the datasets are so large that a statistical p-value has no meaning, other measures are used to judge the quality of the model, especially misclassification rates. Predictive modeling is considered supervised learning because you can use the outcome to judge the accuracy. As a form of predictive modeling, classification is an important part of data mining because it defines groups within the population. Also, it could help us to predict which individuals are at high risk for infectious diseases when pandemic happens.

There are many different classification methods to classify data in SAS Enterprise Miner, such as neural network, decision trees and regression analysis. And we can determine the best model for classification by comparing the results of different classification methods. Specifically speaking, we can compare the rates of correct classification and select the model with the highest rate. Unfortunately, accuracy tends to be inflated when data are used to define the model. For example, it is possible to define a predictive model which is 100% accurate on a training set but 0% accurate on a validation set. Therefore, validation is also essential (Cerrito, 2006).

The classification methods used for data mining are similar to those used for statistical inference. However, data mining uses many different models and then compares the results on a testing set while statistical inference tends to examine a single model and measure its effectiveness by the p-value and by adherence to model assumptions. But

Data mining focuses less on model assumptions and more on the model's ability to actually predict outcomes. Assumptions are not as important as outcomes.

5.2.2 Regression Analysis

In SAS enterprise miner, we will use the regression node for linear or logistic regression, which depends on the type of target variable. If the target variable is nominal, the regression node will use logistic regression; if the target variable is an interval variable or ordinal with more than 7 categories, the linear regression will be applied automatically by the regression node. Since our outcome variable is a binary variable, which has two potential values 0 and 1, the regression node will conduct logistic regression.

Additionally, covariance and correlation between variables should be checked before the regression analysis. Both the correlation and grouping steps provide valuable information on the data at hand, and are more than just statistical exercises. In SAS, we use a type of principal component analysis to identify group of variables that are highly correlated. If some variables are highly correlated, basically clustering could be conducted and choose the best variable or a linear combination of those variables in the same cluster as the final input variable to represent that cluster. Appendix 6 illustrates the covariance and correlation matrices, from which it can be seen that there is no high correlation or covariance between variables. Therefore, the clustering of variables is unnecessary at this point.

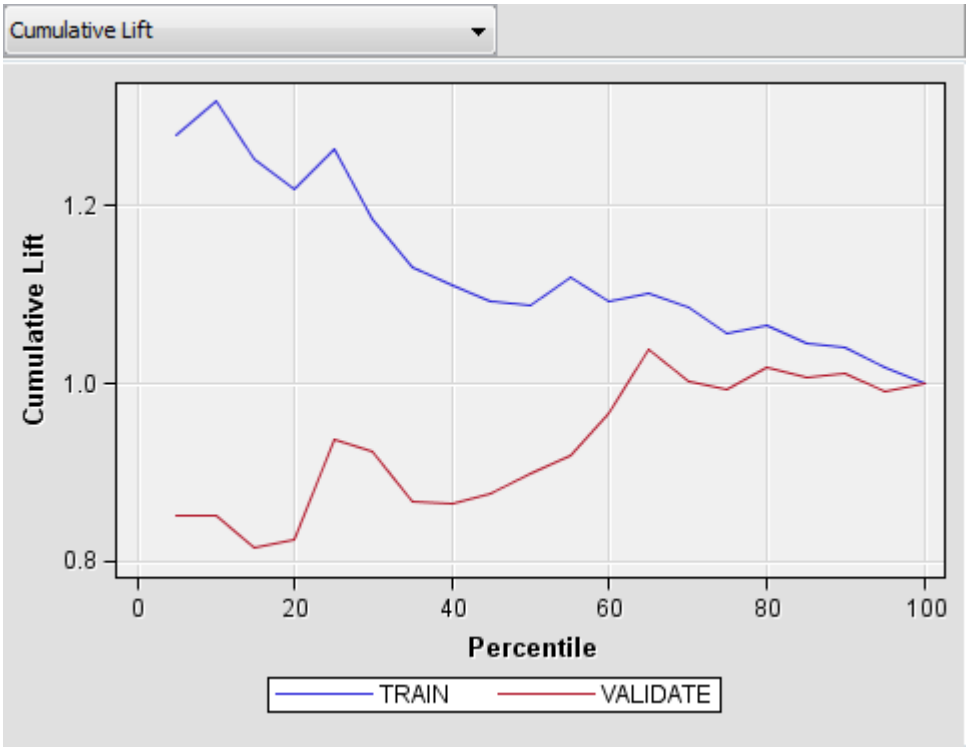


Figure 31. Cumulative life curve for the regression model

The curve above gives the cumulative lift for the regression model. Cases in both training and validation data are ranked based on decreasing predicted target values. A fraction or decile of the ranked data is selected, which corresponds to the horizontal axis of the chart. The ratio, (proportion of cases with the primary outcome, in our case "infected" in the selected fraction) to the proportion of cases with the primary outcome overall, is defined as cumulative lift, which corresponds to the vertical axis. High values of cumulative lift suggest that the model is doing a good job separating the infected and the susceptible and generally if the depth increases, the lift will inversely decrease.

From the Figure 31, we can see that the lift is decreasing for the training set while the lift is increasing for the validation set both from the perspective of entire tendency. Specifically, for the training dataset there are mainly three peaks where the lift first increases and then decrease at a depth of approximately 10%, 25% and 55%

respectively. Similarly, there are two peaks for the validation dataset and lift increases to the peaks at 25% and 65%, then decreases beyond that. Sometimes the training data could inflate results and therefore it is better to examine the lift for the validation dataset.

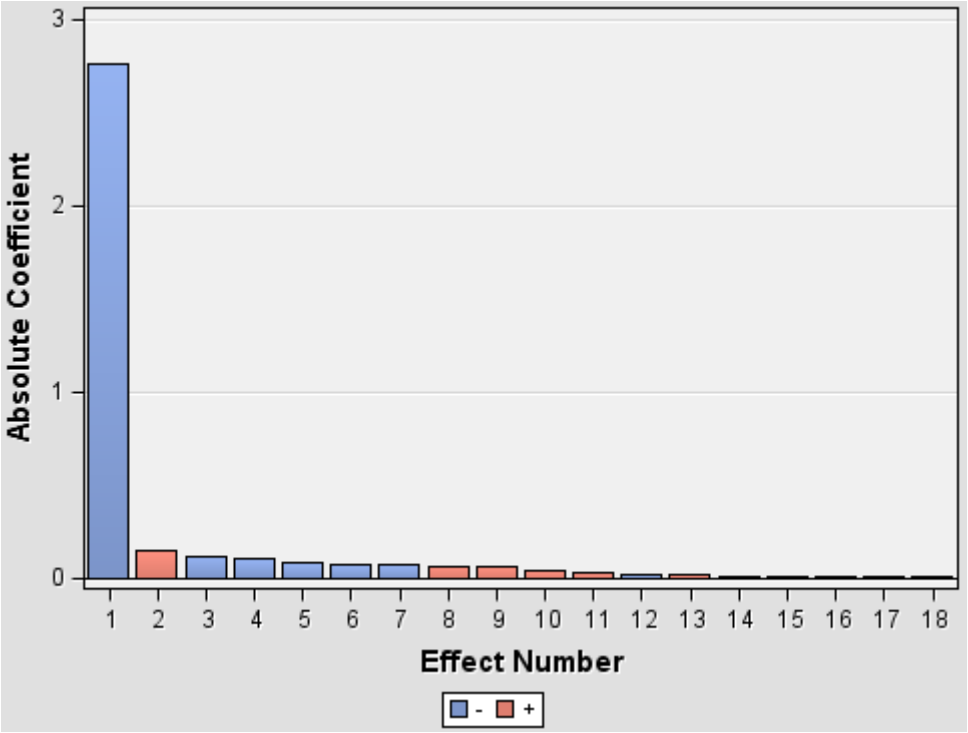


Figure 32. Effects plot of the regression model

The effects plot illustrates the r^2 value for each variable and in this case, we can see that the first four factors, which are intercept, smoker, illicit_drugs and obesity, account for most of the cumulative r^2 and the remaining factors only contributes a small part.

Also, the results indicate that the misclassification rate on the initial training dataset is 6.25%, which is the best indication of the regression accuracy. And the misclassification rates for validation dataset is slightly increased to 6.27%z. In the real world case, we have to also consider the actual group size to determine whether the

prediction is valid and also the model accuracy should be compared to random chance. If the misclassification rate is close to or even higher than random choice rate, it will indicate that the model built is not a good fit. But in our case, since the misclassification for both training and validation datasets are relatively small, we can draw the conclusion that the model we developed is accurate.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	-2.7590	0.6184	19.91	<.0001		0.063
Alcohol	0	1	0.0631	0.0764	0.68	0.4086		1.065
Allergy	0	1	0.0624	0.1152	0.29	0.5884		1.064
Asthma	0	1	-0.0684	0.1077	0.40	0.5255		0.934
Diabetes	0	1	-0.0985	0.1358	0.53	0.4681		0.906
Family_His	0	1	0.0299	0.0993	0.09	0.7633		1.030
Gender	0	1	-0.00128	0.0665	0.00	0.9846		0.999
HighBloodPressure	0	1	0.0387	0.0863	0.20	0.6537		1.039
Illicit_Drugs	0	1	-0.1173	0.2367	0.25	0.6202		0.889
Obesity	0	1	-0.0763	0.0732	1.09	0.2971		0.927
Pregnant	0	1	0.00706	0.2029	0.00	0.9723		1.007
REP_Age		1	0.00115	0.00378	0.09	0.7620	0.0156	1.001
REP_BMI		1	0.00232	0.0208	0.01	0.9110	0.00492	1.002
REP_Exercise_Rate		1	-0.0688	0.0582	1.40	0.2369	-0.0485	0.934
REP_Working_Hours		1	0.00298	0.00422	0.50	0.4801	0.0290	1.003
Recent_Perscription_Drugs	0	1	-0.0161	0.0781	0.04	0.8365		0.984
Smoker	0	1	0.1446	0.1056	1.88	0.1707		1.156
Vegetarian	0	1	0.0120	0.1114	0.01	0.9139		1.012

Table 23. Analysis of Maximum Likelihood Estimates

Table 23 displays the analysis results of maximum likelihood estimates. The estimate column of results shows the weight or contribution to the linear regression equation for each variable. If we take into consideration the first four factors with p-value < 0.30 including intercept, which have the most significant impact on the outcome variable, the linear regression equation after logit function is as follows:

$$logit\left(\frac{p}{1-p}\right) = -2.759 + 0.1446 \times Smoker - 0.0688 \times Exercise\_Rate - 0.0763 \times Obesity \qquad (25)$$

Therefore, we can obtain the probability of getting infected p for an individual.

$$p = \frac{e^{-2.759 + 0.1446 \times Smoker - 0.0688 \times Exercise\_Rate - 0.0763 \times Obesity}}{1 + e^{-2.759 + 0.1446 \times Smoker - 0.0688 \times Exercise\_Rate - 0.0763 \times Obesity}} \quad (26)$$

Odds Ratio Estimates

Effect		Point Estimate
Alcohol	0 vs 1	1.135
Allergy	0 vs 1	1.133
Asthma	0 vs 1	0.872
Diabetes	0 vs 1	0.821
Family_His	0 vs 1	1.062
Gender	0 vs 1	0.997
HighBloodPressure	0 vs 1	1.081
Illicit_Drugs	0 vs 1	0.791
Obesity	0 vs 1	0.858
Pregnant	0 vs 1	1.014
REP_Age		1.001
REP_BMI		1.002
REP_Exercise_Rate		0.934
REP_Working_Hours		1.003
Recent_Perscription_Drugs	0 vs 1	0.968
Smoker	0 vs 1	1.335
Vegetarian	0 vs 1	1.024

Table 24. Odds Ratio Estimates

Table 24 illustrates the odds ratio for each variable, which can be obtained from the Wald Chi-Square column in SAS results. If an odds ratio is greater than 1, as the input variable increases, the output variable will also increase. But when the odds ratio is less than 1, the output variable increases from 0 to 1 as the input variable decreases. The output variable in our experiment is the infected variables, where 1 represents infected and 0 represents not.

Figure 33. Classification chart for regression

From the classification chart for the validation dataset, the model appears correctly classify the individuals based on the input values.

5.2.3 Neural Network

1. Multilayer Perceptrons (MLP)

There are mainly four types of neural network methodologies in SAS Enterprise Miner, the MLP, GLIM, AutoNeural and DMNeural. MLP, the multi-layer perceptron is the default model, which is the most popular form of neural network architecture and a perceptron is a classifier that maps an input to an output. The GLIM represents the generalized linear interactive modeling used in PROC GENMOD in SAS/STAT software, which has no hidden layers. AutoNeural acts as an automated tool to find optimal configurations for a neural network. DMNeural uses the bucketed principal components as input to predict a binary target variable and it also overcomes several

problems of the common neural networks for data mining purposes including nonlinear estimation problem and computing time problem. We can compare these models to see the impact on the results.

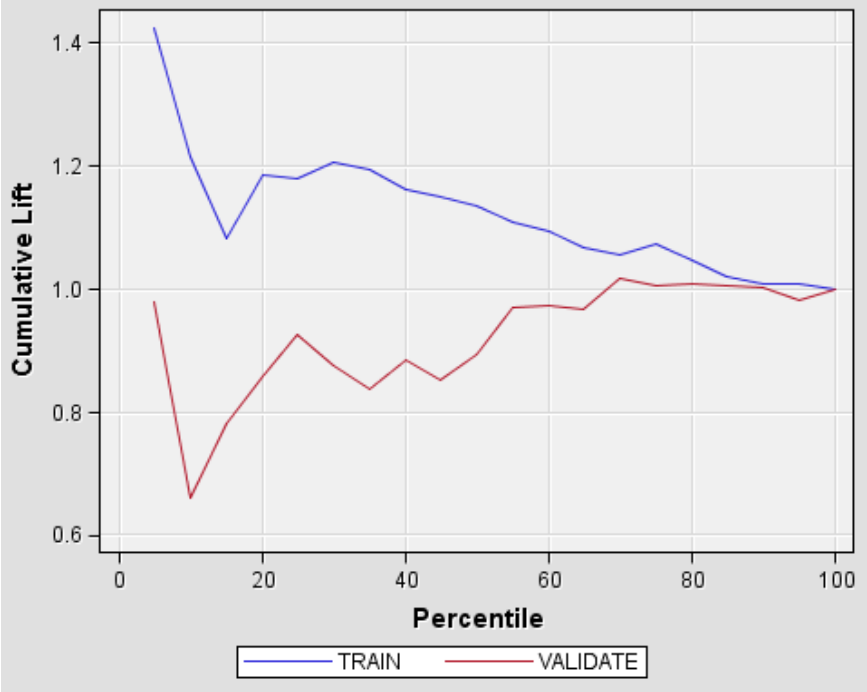


Figure 34. Cumulative lift curve for neural network

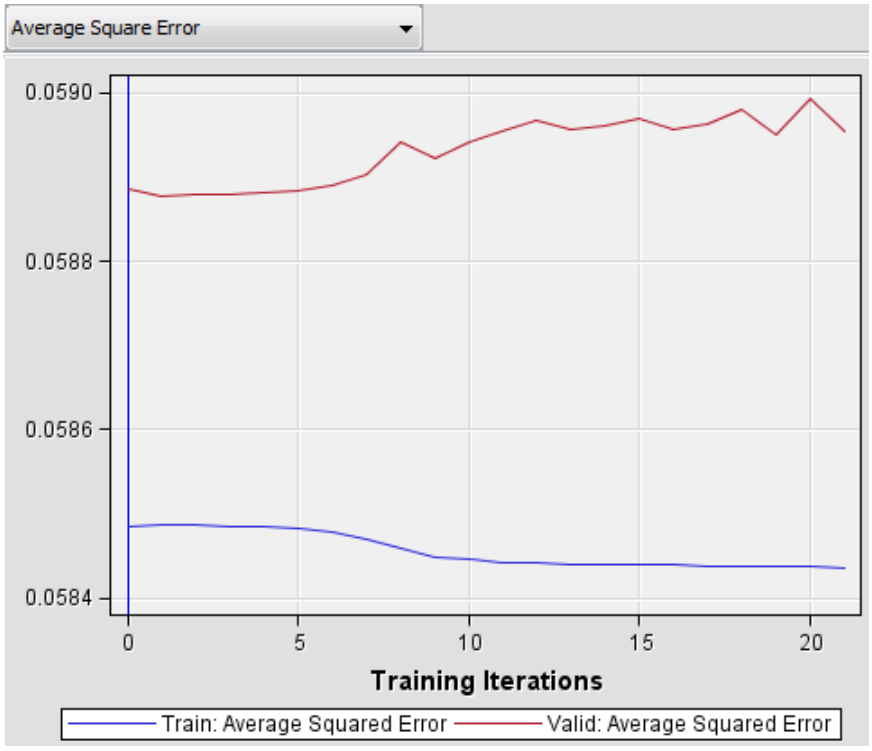


Figure 35. Average square error for training and validation sets

The cumulative lift curve for neural network looks similar to that for regression and just a little bit lower compared to the regression cumulative lift. Since neural network represents an iterative process, the results provide a graph of the rate of convergence to a final model. Additionally, the training set converges after 20 iterations while the validation set seems not to converge at all.

Fit Statistics

Target	Fit Statistics	Statistics Label	Train	Validation
Infected	_DFT_	Total Degrees of Freedom	3998	.
Infected	_DFE_	Degrees of Freedom for Error	3985	.
Infected	_DFM_	Model Degrees of Freedom	13	.
Infected	_NW_	Number of Estimated Weights	13	.
Infected	_AIC_	Akaike's Information Criterion	1893.653	.
Infected	_SBC_	Schwarz's Bayesian Criterion	1975.469	.
Infected	_ASE_	Average Squared Error	0.058565	0.05875
Infected	_MAX_	Maximum Absolute Error	0.943624	0.943624
Infected	_DIV_	Divisor for ASE	7996	6000
Infected	_NOBS_	Sum of Frequencies	3998	3000
Infected	_RASE_	Root Average Squared Error	0.242001	0.242384
Infected	_SSE_	Sum of Squared Errors	468.2829	352.4986
Infected	_SUMW_	Sum of Case Weights Times Freq	7996	6000
Infected	_FPE_	Final Prediction Error	0.058947	.
Infected	_MSE_	Mean Squared Error	0.058756	0.05875
Infected	_RFPE_	Root Final Prediction Error	0.24279	.
Infected	_RMSE_	Root Mean Squared Error	0.242396	0.242384
Infected	_AVERR_	Average Error Function	0.233573	0.23432
Infected	_ERR_	Error Function	1867.653	1405.918
Infected	_MISC_	Misclassification Rate	0.062531	0.062667
Infected	_WRONG_	Number of Wrong Classifications	250	188
Infected	_PROF_	Total Profit for Infected	3748	2812
Infected	_APROF_	Average Profit for Infected	0.937469	0.937333

Table 25. Fit statistics for neural network

The misclassification rates for both the training and validation sets are almost the same compared to those for regression. And the average error for neural network is also almost the same compared to regression. The classification chart in Figure 36 appears

the same rates as the regression. Therefore, at this point in the analysis, no model has been identified as a better one for the classification.

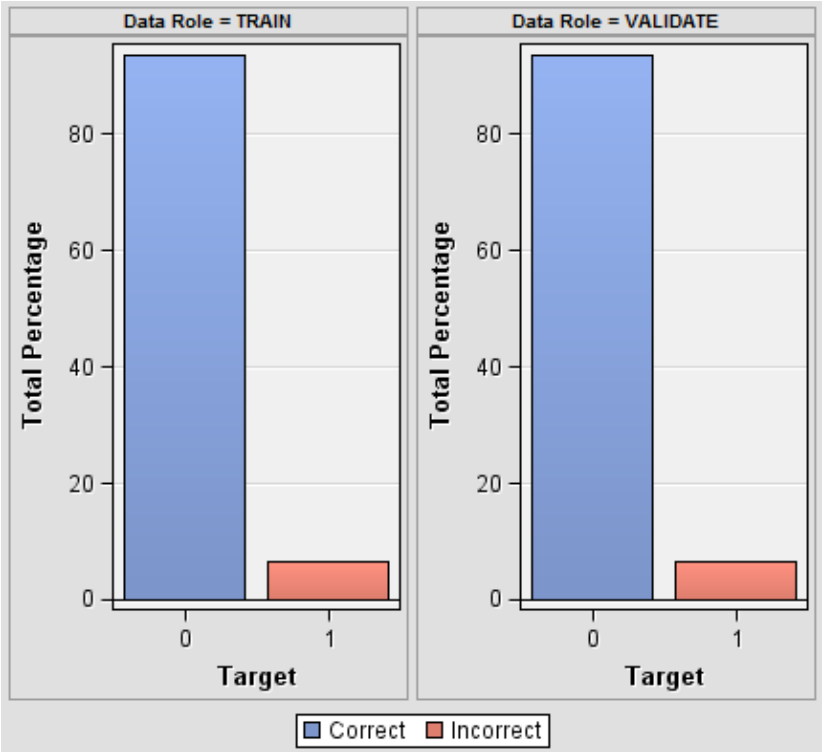
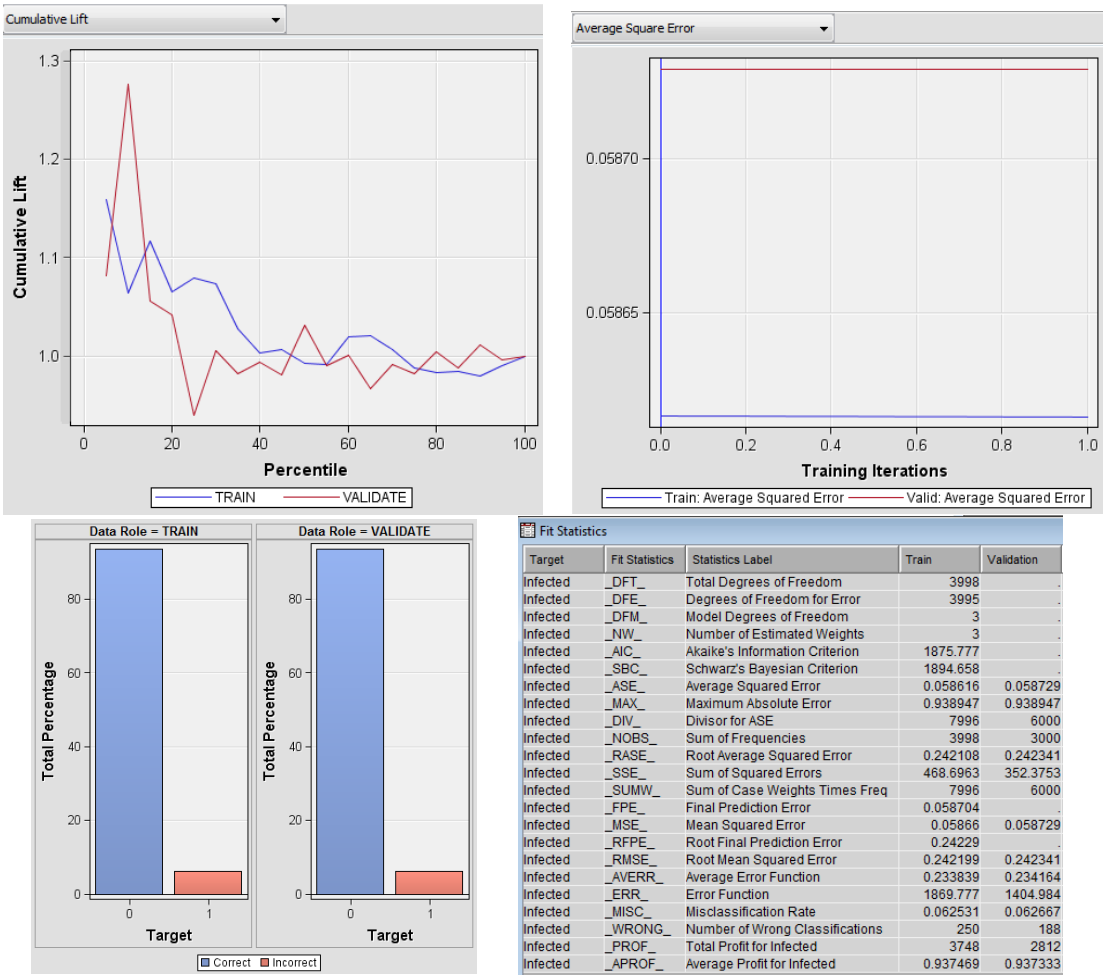


Figure 36. Classification chart of infected variable

2. Generalized Linear Model (GLIM)

Since there is not the best model so far, we will try to build another neural network model. Let's change the architecture to GLIM and set the training techniques to quasi-Newton. Then we can obtain the results as follows.



Fit Statistics

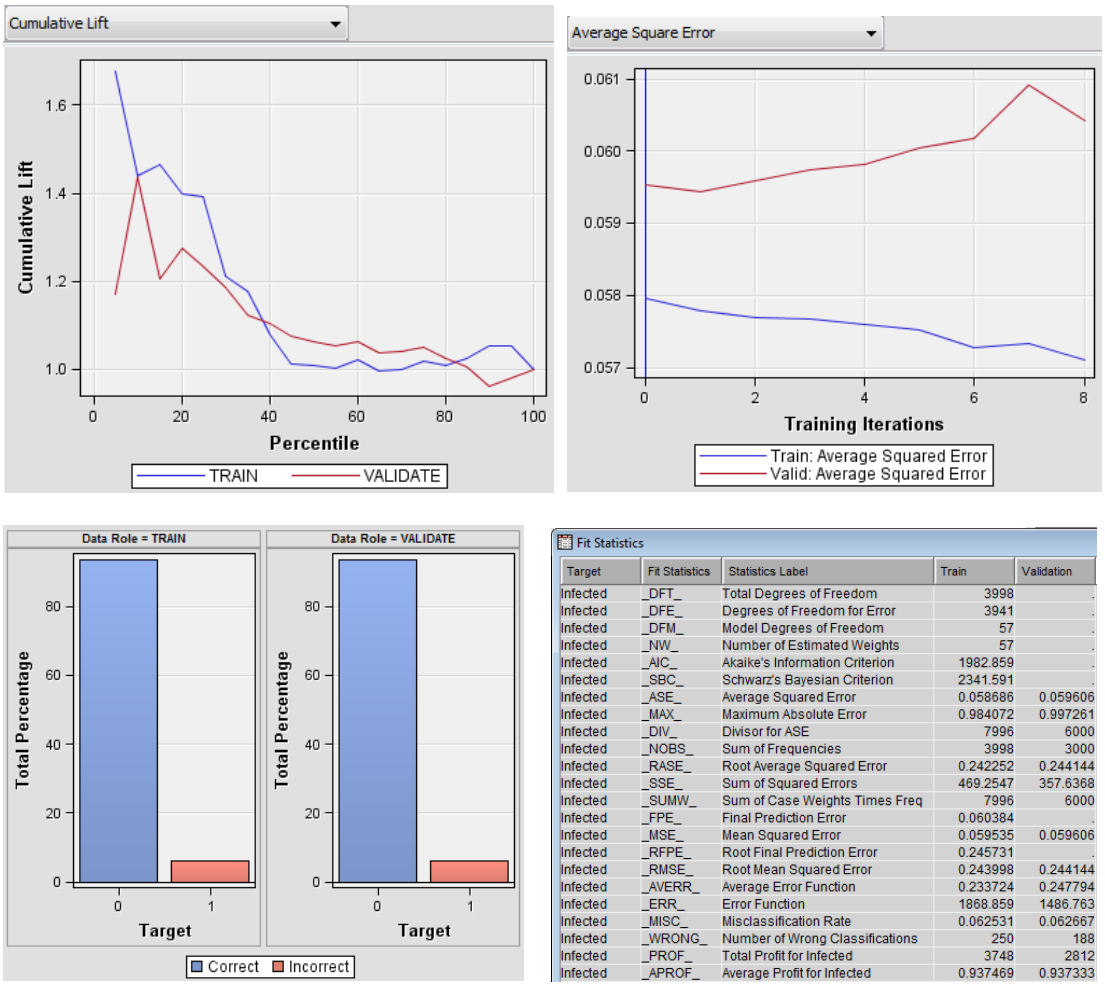
Target	Fit Statistics	Statistics Label	Train	Validation
Infected	_DFT_	Total Degrees of Freedom	3998	.
Infected	_DFE_	Degrees of Freedom for Error	3995	.
Infected	_DFM_	Model Degrees of Freedom	3	.
Infected	_NW_	Number of Estimated Weights	3	.
Infected	_AIC_	Akaike's Information Criterion	1875.777	.
Infected	_SBC_	Schwarz's Bayesian Criterion	1894.658	.
Infected	_ASE_	Average Squared Error	0.058616	0.058729
Infected	_MAX_	Maximum Absolute Error	0.938947	0.938947
Infected	_DIV_	Divisor for ASE	7996	6000
Infected	_NOBS_	Sum of Frequencies	3998	3000
Infected	_RASE_	Root Average Squared Error	0.242108	0.242341
Infected	_SSE_	Sum of Squared Errors	468.6963	352.3753
Infected	_SUMW_	Sum of Case Weights Times Freq	7996	6000
Infected	_FPE_	Final Prediction Error	0.058704	.
Infected	_MSE_	Mean Squared Error	0.05866	0.058729
Infected	_RFPE_	Root Final Prediction Error	0.24229	.
Infected	_RMSE_	Root Mean Squared Error	0.242199	0.242341
Infected	_AVERR_	Average Error Function	0.233839	0.234164
Infected	_ERR_	Error Function	1869.777	1404.984
Infected	_MISC_	Misclassification Rate	0.062531	0.062667
Infected	_WRONG_	Number of Wrong Classifications	250	188
Infected	_PROF_	Total Profit for Infected	3748	2812
Infected	_APROF_	Average Profit for Infected	0.937469	0.937333

Figure 37. GLIM model results

In the cumulative lift curve, there is a big rise and fall on both training and validation sets but the whole tendency is decreasing for both. There is still no convergence in the validation set. Also, the misclassification rate for training and validation sets are 6.25% and 6.27% respectively and there is almost no change in misclassification compared to the results in the previous models. However, inversely the average error is slightly higher. Therefore, it still does not improve the results a lot by changing the architecture to GLIM and unfortunately it is still unreliable to differentiate between events and nonevent in both training and validation sets by using the new model.

3. AutoNeural

AutoNeural is another type of neural network technique in SAS Enterprise Miner and the Neural Network node is required to change manually while the AutoNeural node changes the default setting automatically. Also it adds hidden nodes one at a time and we can define the maximum number of iterations with an adjustable setting. The default setting defines a single-layer neural network and specifies how hidden layers are added to the model. We make the activation function include the logistic function, which a slight difference from the default setting. The Figure 38 illustrates the results of AutoNeural Network analysis.



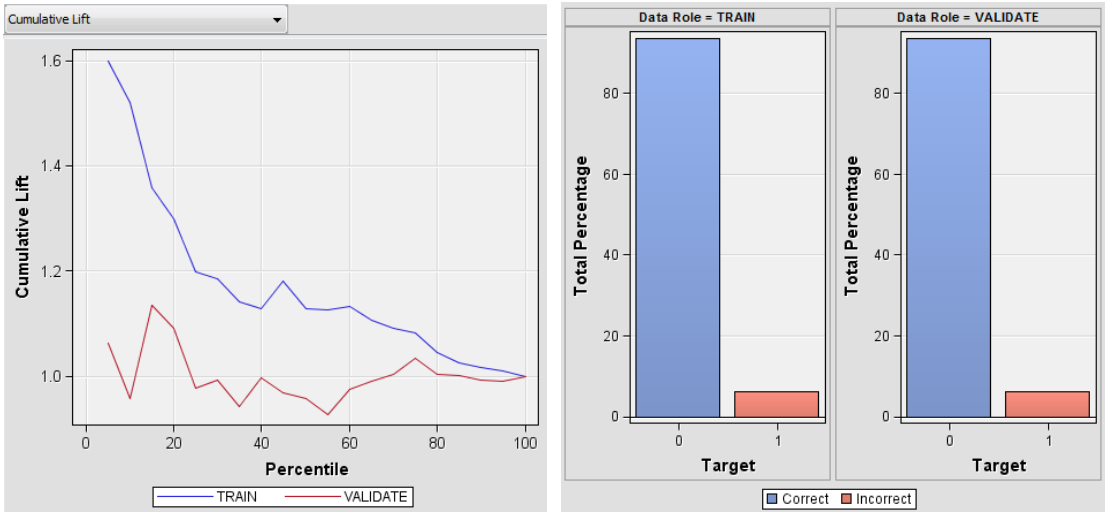
Target	Fit Statistics	Statistics Label	Train	Validation
Infected	_DFT_	Total Degrees of Freedom	3998	.
Infected	_DFE_	Degrees of Freedom for Error	3941	.
Infected	_DFM_	Model Degrees of Freedom	57	.
Infected	_NW_	Number of Estimated Weights	57	.
Infected	_AIC_	Akaike's Information Criterion	1982.859	.
Infected	_SBC_	Schwarz's Bayesian Criterion	2341.591	.
Infected	_ASE_	Average Squared Error	0.058686	0.059606
Infected	_MAX_	Maximum Absolute Error	0.984072	0.997261
Infected	_DIV_	Divisor for ASE	7996	6000
Infected	_NOBS_	Sum of Frequencies	3998	3000
Infected	_RASE_	Root Average Squared Error	0.242252	0.244144
Infected	_SSE_	Sum of Squared Errors	469.2547	357.6368
Infected	_SUMW_	Sum of Case Weights Times Freq	7996	6000
Infected	_FPE_	Final Prediction Error	0.060384	.
Infected	_MSE_	Mean Squared Error	0.059535	0.059606
Infected	_RFPE_	Root Final Prediction Error	0.245731	.
Infected	_RMSE_	Root Mean Squared Error	0.243998	0.244144
Infected	_AVERR_	Average Error Function	0.233724	0.247794
Infected	_ERR_	Error Function	1868.859	1486.763
Infected	_MISC_	Misclassification Rate	0.062531	0.062667
Infected	_WRONG_	Number of Wrong Classifications	250	188
Infected	_PROF_	Total Profit for Infected	3748	2812
Infected	_APROF_	Average Profit for Infected	0.937469	0.937333

Figure 38. AutoNeural model results

From the cumulative lift curve, the score rankings for both training and validation sets are decreasing rapidly. Also there are no significant changes on misclassification rates and the average error is even slightly higher, which indicates that the AutoNeural is not a good fit for the data modeling.

4. DMNeural

The DMNeural node offers another method for neural network analysis. It is focusing on nonlinear estimation and could simultaneously reduce the computation time. Also, it begins with principle component analysis which determines which input variables have a significant impact on the outcome variables. Then a small set of principal components is selected for further modeling.



Fit Statistics

Target	Fit Statistics	Statistics Label	Train	Validation
Infected	_ASE_	Average Squared Error	0.058415	0.058959
Infected	_DIV_	Divisor for ASE	7996	6000
Infected	_MAX_	Maximum Absolute Error	0.977752	0.980316
Infected	_NOBS_	Sum of Frequencies	3998	3000
Infected	_RASE_	Root Average Squared Error	0.241691	0.242815
Infected	_SSE_	Sum of Squared Errors	467.0834	353.7549
Infected	_DISF_	Frequency of Classified Cases	3998	3000
Infected	_MISC_	Misclassification Rate	0.062531	0.062667
Infected	_WRONG_	Number of Wrong Classifications	250	188
Infected	_PROF_	Total Profit	3748	2812
Infected	_APROF_	Average Profit	0.937469	0.937333
Infected	_ERR_	Error Function	1859.545	.
Infected	_AVERR_	Average Error Function	0.232559	.
Infected	_DFT_	Total Degrees of Freedom	3998	.
Infected	_DFE_	Degrees of Freedom for Error	3979	.
Infected	_MSE_	Mean Squared Error	0.058694	.
Infected	_RMSE_	Root Mean Squared Error	0.242268	.
Infected	_NW_	Number of Weights	19	.
Infected	_FPE_	Final Prediction Error	0.058973	.
Infected	_RFPE_	Root Final Prediction Error	0.242843	.
Infected	_AIC_	Akaike's Information Criterion	1897.545	.
Infected	_SBC_	Schwarz's Bayesian Criterion	2017.123	.

Figure 39. DMNeural model results

The cumulative lift curve does not look accurate and also the misclassification rate for training and validation sets stay the same. Also, the average error is still high and all these indicate the DMNeural model is not the best.

5.2.4 Decision Tree

Decision tree is another different approach for classification, which develops a series of if-then rules. Each rule assigns one sample to at least one branch of the tree. The initial node is the root of the decision tree, which includes the entire dataset. The final nodes are the leaves of the tree, which are the predictive values for a sample.

The imputation is not required in decision trees since missing value could be used in creating if-then rules. For the splitting rules, interval criterion is set by default while Gini index and Entropy are selected for nominal criterion and ordinal criterion respectively. Also, the significant level is 0.2 by default. After running the program, the

decision tree is not generated successfully as a result. Therefore, we would create an interactive tree as an alternative in order to prune a decision tree instead of generating the decision tree automatically. Every time splitting a node, the candidate variables to split are ranked by logworth (-Log(p-value from Chi-square or F test)) and the variable with maximum logworth value are selected. Also, we can manually specify the split point for the rule. For example, the exercise rate variable is selected on the second layer with the split point 3 rather than 2.5. In our case, we use logworth=0 as the stopping criterion when adding generations.

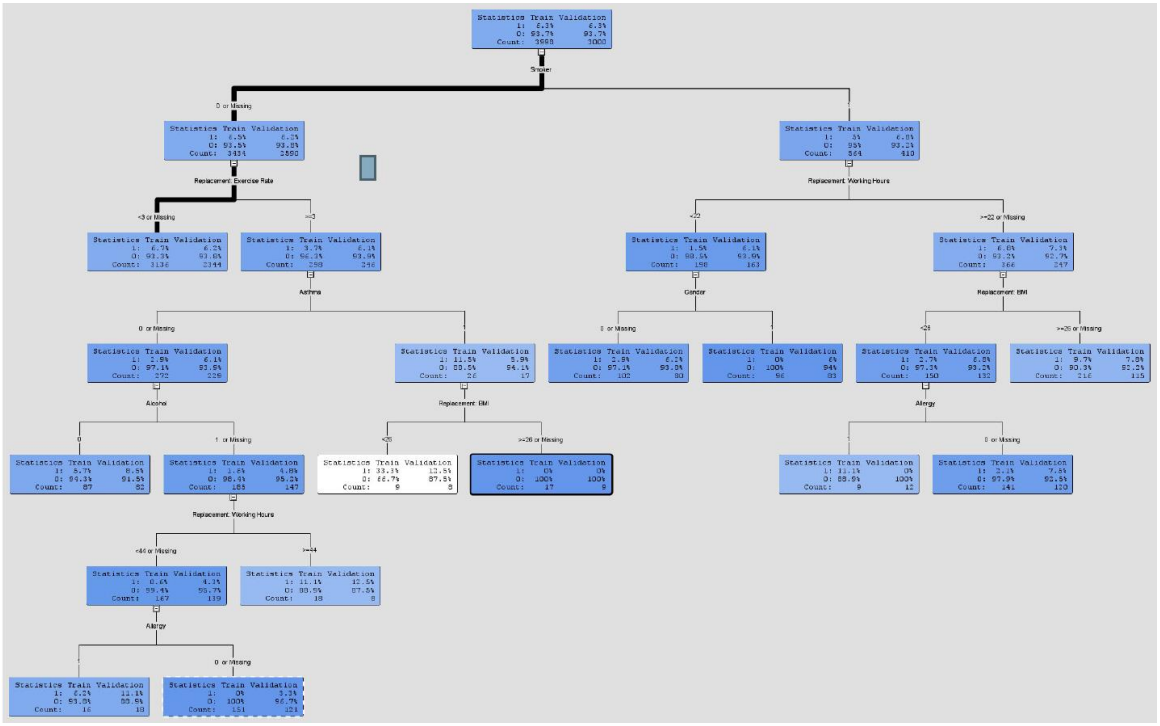


Figure 40. Decision tree result structure

From the structure of the interactive decision tree, there are 6 levels and 12 leaves, which indicates that for each individual, he or she could be classified into one of these 12 classes. Their English rules are shown in Appendix 6. The decision tree demonstrates the predicted infected probability at each node in the English rules. Also for the leaf node, the percentage of susceptible and infected individuals is used as a value of

parameter γ_s in Formula 22, which is mainly from the training datasets. For example, for Node ID 29 shown in Figure 41, if one individual is a smoker, works more than 22 hours per week on average, has a BMI of lower than 26 and also has allergy, the probability for him or her to get infected would be 0.11, which is also used as γ_s value in Formula 22.

```
*------------------------------------------------------------*
 Node = 29
*------------------------------------------------------------*
if Smoker IS ONE OF: 1
AND Replacement: Working Hours >= 22 or MISSING
AND Replacement: BMI < 26
AND Allergy IS ONE OF: 1
then
 Tree Node Identifier   = 29
 Number of Observations = 9
 Predicted: Infected=1  = 0.11
 Predicted: Infected=0  = 0.89
```

Figure 41. English rules for Node 29

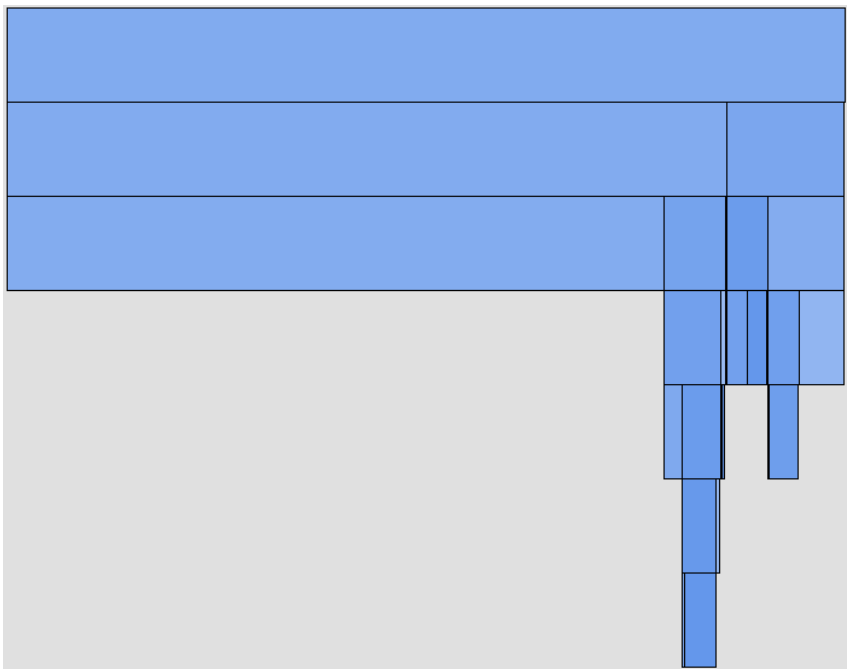
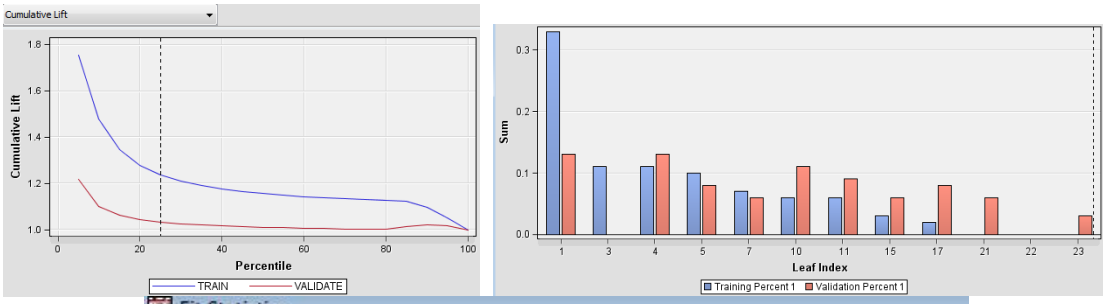


Figure 42. Decision tree map

The decision tree map also illustrates the percentage of susceptible individuals and infected individuals at each node in the tree.



Fit Statistics

Target	Fit Statistics	Statistics Label	Train	Validation
Infected	_NOBS_	Sum of Frequencies	3998	3000
Infected	_MISC_	Misclassification Rate	0.062531	0.062667
Infected	_MAX_	Maximum Absolute Error	0.978723	1
Infected	_SSE_	Sum of Squared Errors	463.8527	354.8323
Infected	_ASE_	Average Squared Error	0.058011	0.059139
Infected	_RASE_	Root Average Squared Error	0.240854	0.243185
Infected	_DIV_	Divisor for ASE	7996	6000
Infected	_DFT_	Total Degrees of Freedom	3998	.
Infected	_APROF_	Average Profit for Infected	0.937469	0.937333
Infected	_PROF_	Total Profit for Infected	3748	2812

Figure 43. Cumulative lift, leaf index and fit statistics for interactive decision tree

As we know, if the cumulative lift is much more above 1.0 for both training and validation sets, there would be a reasonably accurate mode. However, from our result the cumulative life is only a little more than 1.0, therefore the interactive decision tree we built only provides a slightly better prediction than the chance. For each leaf of the decision tree, the leaf index demonstrates the bar chart of predicted infection probability descent by training sets. It graphically compares the results for both training and validation sets and also indicates how well the decision rules separate infected individuals from the susceptible. In our case, the results look acceptable since there is no significant difference between the training and validation datasets for each leaf node. Also, for the misclassification, they are still the same as the neural network model.

5.3 Model Comparison

So far, we have been using logistic regression model, four neural network models and decision tree model to train our datasets and build the predictive models. Then we can use model comparison module in SAS enterprise miner to compare these models and determine which one could predict the outcome the best. The fit statistics we used in the previous sections could also make a more direct comparison between models and additionally the receiver operating curve (ROC) and the cumulative lift functions are provided as the visual representations to compare different predictive models.

As discussed previously, there are a couple of metrics which we can use to assess the model performance. In SAS Enterprise Miner, for decision prediction the Model Comparison tool rates the model performance based on accuracy or misclassification rate, profit or loss, and by the Kolmogorov-Smirnov (KS) statistic, which describes the ability of the model to separate the infected individuals and non-infected individuals. For ranking predictions, the Model Comparison tool gives two closely related measures of model fit. The ROC index is similar to concordance, which equals the percent of concordant cases plus one-half times the percent tied cases. The Gini coefficient (for binary prediction) equals 2 × (ROC Index − 0.5). For estimate predictions, the Model Comparison tool provides two performance statistics. Average squared error was used to tune many of the models fit in earlier chapters. The Schwarz's Bayesian Criterion (SBC) is a penalized likelihood statistic. The likelihood statistic was used to estimate regression and neural network model parameters and can be thought of as a weighted average squared error. SBC is provided only for regression and neural network models and is calculated only on training data. To summarize, we get the table displaying the relationship of choice of fit statistics and prediction of interest.

Prediction Type	Validation Fit Statistic	Direction
Decisions	Misclassification	smallest
	Average Profit/Loss	Largest/smallest
	Kolmogorov-Smirnov Statistics	largest
Rankings	ROC Index (concordance)	largest
	Gini Coefficient	largest
Estimates	Average Squared Error	smallest
	Schwarz's Bayesian Criterion	smallest
	Log-Likelihood	largest

Table 26. Choice of fit statistics and prediction of interest

The Model Comparison node in SAS Enterprise Miner gives us both statistical and data mining measures and enables us to compare the performance of different models using various benchmarking criteria. Several comparative criteria including K-S value, ROC Index, Gini coefficient, Gain, Lift, cumulative lift and their corresponding results are selected in Table 27.

Fit Statistics		Decision Tree	Regression	MLP	GLIM	DMNeural	AutoNeural
K-S Value	Train	0	0.069	0.022	0.004	0.076	0.088
	Validation	0	0.019	0.021	0.008	0.014	0.059
ROC Index	Train	0.5	0.546	0.52	0.504	0.554	0.537
	Validation	0.5	0.484	0.505	0.499	0.495	0.528
Gini Coefficient	Train	0	0.093	0.04	0.007	0.107	0.073
	Validation	0	-0.033	0.009	-0.002	-0.011	0.057
Gain	Train	2E-13	31.93	15.94	6.45	51.92	43.93
	Validation	2.66E-13	14.89	7.98	27.66	4.26	43.62
Lift	Train	1	1.36	0.95	0.97	1.44	1.20
	Validation	1	0.85	0.63	1.47	0.85	1.70
Cumulative Lift	Train	1	1.32	1.16	1.06	1.52	1.44
	Validation	1	0.85	0.92	1.28	0.96	1.44

Table 27. Fit Statistics for different models

From the fit statistics table, it appears that Regression, DMNeural, AutoNeural are better than other predictive modeling methods based on different statistical metrics. So these three models are currently our champion model candidates. If we look into the details more carefully, in most cases, as a matter of fact the AutoNeural model is our champion model at this point no matter for Training sample or validation sample. However, let's have a look at the performance of their ROC curve and cumulative life function before the final conclusion is drawn.

Fit statistics provides us with a more direct way to compare different models while the receiver operating curve (ROC) and the cumulative life function can also be used to compare models. As discussed in previous section, the ROC curve maps the sensitivity on the Y axis against 1-specificity on the X axis and it could indicate the overall accuracy of the model. Sensitivity is also the true positive rate (TPR), which is the proportion of target values that are predicted as a value of 1 and are actually equal to 1. However, the specificity is the true negative rate (TNR), which is the proportion of observation that is predicted as a value of 0 and is actually equal to 0. The outcome has two potential values, 0 and 1, which represents susceptible individuals and infected individuals respectively in our case. Moreover, if the more the area under the curve is close to 1, the more accurate the model for the training set is. It also indicates that the closer the ROC curve lies to the top left concern, the more accurate the predictive model is.

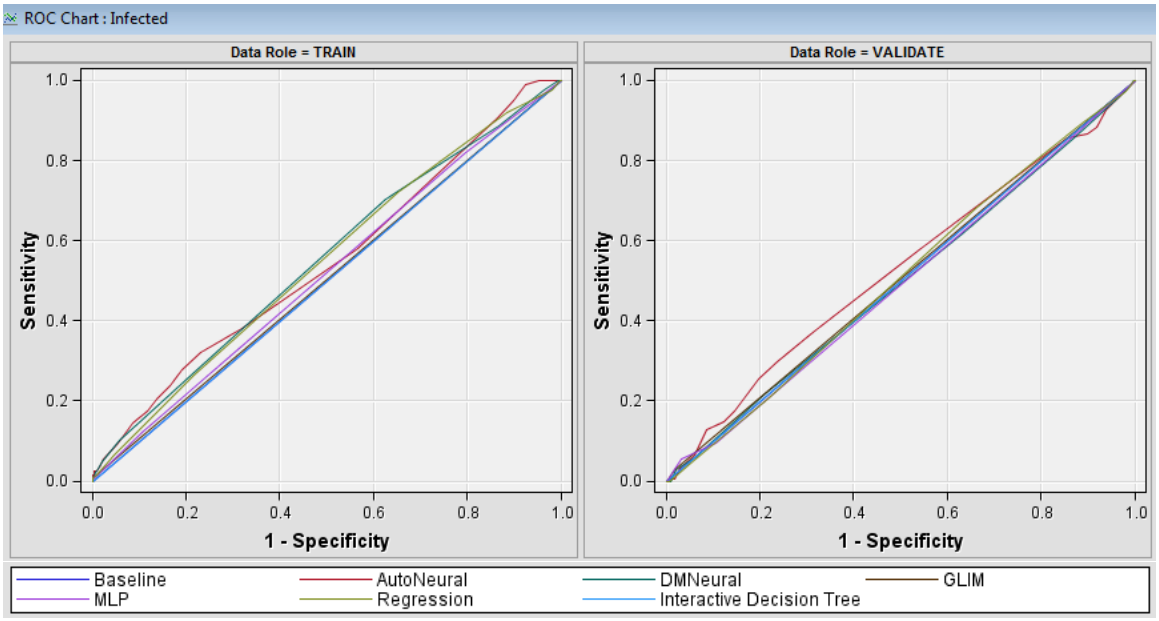


Figure 44. ROC Curve for Different Models

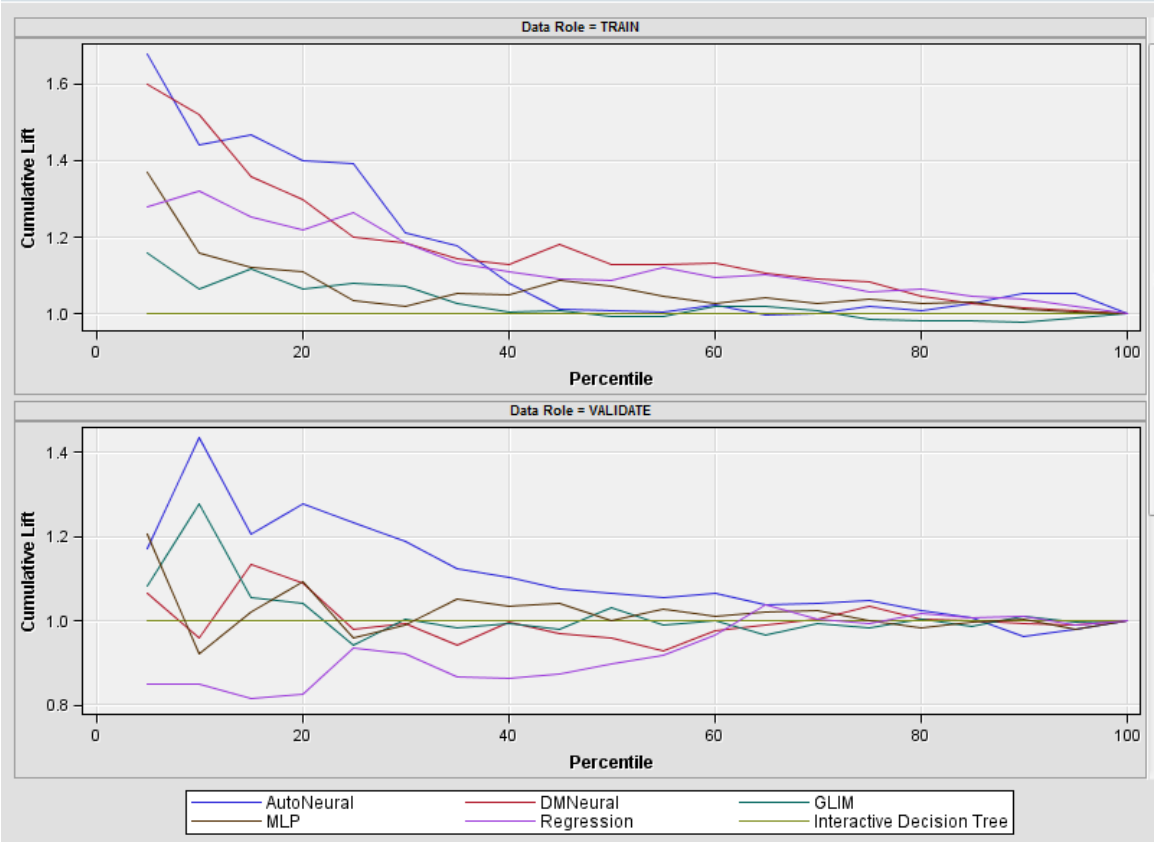


Figure 45. Cumulative lift charts for different models

From both the ROC curve and cumulative lift function, AutoNeural model also does the best job among all the models. Therefore, according to the comparison results of both fit statistics, ROC curve and cumulative lift functions, we can draw the conclusion for our case that the AutoNeural model is the best model, which has a better performance over other models.

But in real-world cases, it's better to choose logistic regression as the champion model if there is no big difference between regression model and neural network model. Unlike regression model and decision trees, neural networks do not present an easily-understandable model and it is more of a "black box" that delivers results without an explanation of how the results were derived. Thus, it is difficult or impossible to explain how decisions were made. In other words, if a challenge is made to a neural network, it is very difficult to explain and justify to non-technical people how decisions were made. Also, it is difficult to incorporate a neural network model into a computer system without using a dedicated "interpreter" to the model. So if the goal is to produce a program that can be distributed with an embedded predictive model, generally it is necessary to send some additional module or information for the neural network interpretation.

5.4 Model Implementation

After we train and compare the predictive models, AutoNeural model is selected to represent the relationship between inputs and the target. The model must be put into use and score a completely new dataset in the model. In SAS enterprise miner, the Score node can help us to model implementation and generate predicted values for an

infection risk. In our case, we have a score data source, which contains all the demographic and health status information of 1000 susceptible individuals. And the variable identification in the score data set is identical to the train and validation datasets.

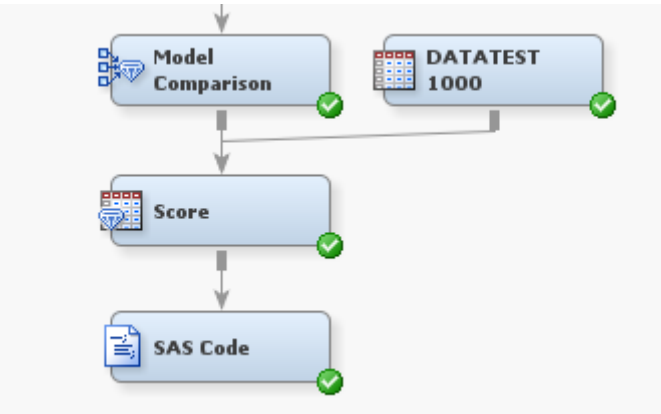


Figure 46. Scoring process with the best model

Figure 46 shows the scoring process and the score node created predicted values using the best model that is AutoNeural model in our case. Consequently, the predicted value for each susceptible individual would be used as γ_s, that describes the health condition and the susceptibility of an individual. Also, the SAS Code node is used to export the scored data and the statistical results and distribution of the scored data are illustrated in Table 28 and Figure 47.

Predicted Value γ_s	Number	Percentage
0.002-0.022	76	7.60%
0.022-0.042	58	5.80%
0.042-0.062	578	57.80%
0.062-0.082	107	10.70%
0.082-0.102	40	4.00%
0.102-0.122	63	6.30%
0.122-0.142	53	5.30%
0.142-0.162	19	1.90%
0.162-0.182	2	0.20%
0.202-0.222	3	0.30%
0.222-0.242	1	0.10%
Grand Total	**1000**	**100%**

Table 28. Number of susceptible individuals for different predicted value ranges

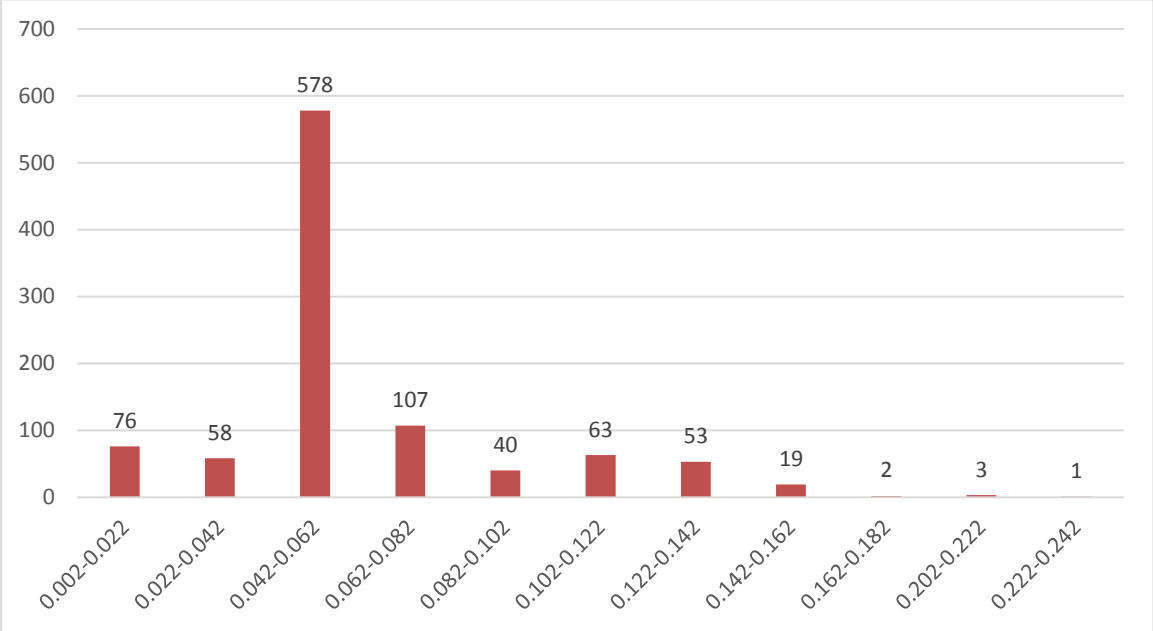


Figure 47. Distribution of number of susceptible individuals by predicted γ_s value

From both Table 28 and Figure 47, we can see that the predicted value of γ_s by using AutoNeural model is less than 0.25 and more than 50% of susceptible individuals have their predicted γ_s value between 0.042 and 0.062, that describes their health status and susceptibility. However, this is only part of their overall infection risk and other parts such as spatiotemporal information and the disease characteristics itself are also critical. Moreover, after the infection risk of each susceptible individual is estimated, their relationship with other people including both the infected and the susceptible via contact network plays an essential role in prioritizing the public health strategies such as vaccine distribution and resource allocation, which is of great help to decision makers in public health organizations.

CHAPTER 6 CONCLUSIONS

In recent year, the infectious diseases have emerged from some part of the world and rapidly spread around the globe. In the last decade alone disease, such as SARS in 2003 and H1N1 in 2009, have spread globally, and have received needed attention from the public as well as the public health agencies. It is essential to understand, predict and control the spread of the disease. However, as discussed previously, the mathematical techniques used to understand, forecast and control the spread of infectious disease is not effective and sometimes lags the actual spread of disease, and hence is of limited value for proactive actions to mitigate the spread.

The approach that has been developed and suggested in this dissertation is built on existing methods from diverse fields such as contact network modeling, graph theory, space-time path development and risk analysis. Some “sensors”, which in fact are the sample group of individuals, or web-based tools, or even survey methodology, and comprehensive application of the approach, are also introduced in this dissertation to mitigate the disease spread. Additionally, this approach provides a prototype of the infection risk estimation and could be practically implemented to better control the spread of infectious disease. In reality, a large amount of people’s medical records or clinical trial from hospital and clinics could be of great help to this approach. Also, the application of GIS tracking feature on our communication devices such as cellphone

would also be very helpful to collect people's GIS information, such as the daily movements and activities.

Last but not least, as a module of decision support system, this approach would be working close with other modules in the system. On one hand, the disease spread module collects relevant information from different sources, such as hospital, clinic, mobile devices, survey and even other modules in the system, to estimate the infection risk of susceptible individuals; On the other hand, the disease spread module also provides useful information for other modules in the decision support system, including vaccine distribution module, patient distribution module, resource allocation module and ambulance distribution module etc. All the parts in the system have to work effectively and efficiently make maximum efforts in order to reduce the loss and cost during the pandemics.

CHAPTER 7 FUTURE RESEARCH

Future research should aim to extend and validate the application of infection index and the likelihood of infection to large networks, such as a residential community, the population in a city, in a county or even in a state. Also, besides the study of disease spread through synchronous physical presence of susceptible and infected individuals, future research could be extended to the situation of asynchronous physical presence.

Additionally, the infection index itself could be expanded to a large potential factors including both individual and disease factors, such as the medical treatment records, clinical trials, population, climate, censors, etc. Also, in the infection index formula, the time parameter p and the distance parameter q should be estimated based on the statistical analysis on the historical data.

Finally, feedback of values of infection index and the likelihood of infection to the general public based on interaction with a web based tool would be a valuable contribution to society. The decision support system is implemented through this web-based tool and provides a platform to collect information and make a wise decision during the pandemics.

REFERENCES

Allison PD. Logistic regression using the SAS system: theory and application. Cary, N.C.: SAS Institute, 1999:179-97. Algeo T. J., Lyons T. W., Paleoceanography 21, PA1016 (2006).

Amaral L.A.N. and Ottino J. M., 2004, "Complex networks - augmenting the framework for the study of complex systems," Eur. Phys. J. B. 38, 147–162.

Aschwanden C. Spatial Simulation Model for Infectious Viral Diseases with Focus on SARS and the Common Flu. in Proceedings of the 37th Hawaii International Conference on System Sciences, 1-5, 2004.

Ball F., Mollison D., Scalia-Tomba G., 1997 Epidemics with two levels of mixing Ann. Appl. Prob., 7 (1997), p. 46.

Bansal S., Grenfell B., and Meyers L.A. When individual behavior matters: Homogeneous and network models in epidemiology. Journal of the Royal Society Interface 4:879–891, 2007.

Bansal S., Pourbohloul B., Hupert N., Grenfell B., and Meyers L.A. The shifting demographic landscape of influenza. PLoS One, in press, 2010.

Bansal S., Pourbohloul B., and Meyers L. (2006) A comparative analysis of influenza vaccination programs. PLoS Med 3: e387.

Bansal S., Pourbohloul B., and Meyers L.A. Comparative analysis of influenza vaccination programs. PLoS Medicine 3:e387, 2006.

Bernard L., and Kruger T. Integration of GIS and spatio-temporal simulation models: interoperable components for different simulation strategies. Transaction in GIS. 4(3). June, 2000.

Boccaletti S., Latora V., Moreno Y., Chavez M., and Hwang D.U. Complex networks: Structure and dynamics. Phys. Rep., vol. 424, p.175 , 2006.

Carneiro H.A., Mylonakis E (2009) Google trends: a web-based tool for real-timesurveillance of disease outbreaks. Clinical Infect Dis 49: 1557–1564.

Cassa C.A., Iancu K., Olson K.L., Mandl K.D. A software tool for creating simulated outbreaks to benchmark surveillance systems. BMC Medical Informatics and Decision Making 2005, 5:22.

Centers for Disease Control and Prevention. 1994 Fact Book 7 (CDC, Atlanta, 1994)

Cerrito P.B., Introduction to data mining using SAS enterprise miner. SAS press, SAS Institute Inc., Cary, NC, USA (2006)

Christakis N.A., Fowler J.H. (2010). Social Network Sensors for Early Detection of Contagious Outbreaks. PLoS ONE 5(9): e12948. doi:10.1371/journal.pone.0012948.

Christakis N.A., Fowler J.H. The spread of obesity in a large social network over 32 years. N Engl J Med 2007;357: 370–379.

Christakis, N.A., and Fowler J.H. 2009. Social network visualization in epidemiology. Norsk Epidemiologi 19(1): 5-16.

Christakis NA, Fowler JH. Social network sensors for early detection of contagious outbreaks. PloS one 2010; 5:e12948.

Clayton R. N., Mayeda T. K., Geochim. Cosmochim. Acta 60, 1999 (1996).

Cohen, M. L. Changing patterns of infectious disease. Nature 406, 762–767 (2000).

Cohen R., Havlin S., and ben Avraham D. Efficient immunization strategies for computer networks and populations. Physical Review Letters, 91:247901, 2003.

Craft M.E., Volz E., Packer C., and Meyers L.A. (2009). Distinguishing epidemic waves from disease spillover in a wildlife population. Proc. R. Soc. Lond., B, Biol. Sci., 276, 1777–1785.

Davey V.J., Glass R.J., Min H.J., Beyeler W.E., and Glass L.M. (2008). Effective, robust design of community mitigation for pandemic influenza: A systematic examination of proposed US guidance.

DellValle S.Y., Hyman J.M., Hethcote H.W., and Eubank S.G. Mixing patterns between age groups in social networks. Social Networks. 2007, 29:539-554.

Dimitrov N., Goll S., Meyers L.A., Pourbohloul B., Hupert N. Optimizing tactics for use of the U.S. antiviral strategic national stockpile for pandemic (H1N1) influenza, 2009. PLoS Currents: Influenza, 2009; RRN1127.

Dimitrov N.B. and Meyers L.A. Mathematical approaches to infectious disease prediction and control. Tutorials in Operations Research, INFORMS 2010, 1-25, 2009.

Eubank S., Guclu H., Kumar A., Marathe M.V., Srinivasan A., Toroczkai Z., and Wang N. Modeling disease outbreaks in realistic urban social networks, Nature 429 (2004). pp.180-184.

Farnsworth M.L., and Ward M.P. Identifying spatio-temporal patterns of transboundary disease spread: examples using avian influenza H5N1 outbreak, Veterinary Research 40 (2009), p. 20.

Ferguson N.M., Garnett G.P. 2000 More realistic models of sexually transmitted disease transmission dynamics: sexual partnership networks, pair models, and moment closure Sex Transm. Dis., 27 (2000), p. 600.

Freeman L.C. (1979). Centrality in Social Networks: Conceptual Clarification, Social Networks, 1, 215-239.

Golledge, R. and R. Stimson., Spatial Behavior: A Geographic Perspective. New York: The Guilford Press, 1997.

Hägerstrand, T. (1970). What about people in regional science? Papers of the Regional Science Association, 24, 1-12.

Hamill J.T., 2006, Analysis of Layered Social Networks, Dissertation. Air Force Institute of Technology.

Han, J., & Kamber, M. (2011). Data mining : concepts and techniques (3rd ed.). Burlington, MA: Elsevier.

Hethcote H. 2000 Mathematics of infectious diseases SIAM Rev., 42 (2000), p. 599.

Hosmer DW Jr, Lemeshow S. Applied logistic regression. New York: John Wiley, 1989:118-24.

Kantardzic, M. (2011). Data mining : concepts, models, methods, and algorithms (2nd ed.). Hoboken, N.J.: John Wiley : IEEE Press.

Keeling M.J., Danon L., Vernon M.C., and House T.A. Individual identity and movement networks for disease metapopulations. Proceedings of the National Academy of Sciences of the United States of America, vol. 107, no. 19, pp. 8866–8870, 2010.

Keeling M.J., Woolhouse M.E., May R.M., Davies G., Grenfell B.T. 2003 Modeling vaccination strategies against foot-and-mouth disease Nature, 421 (2003), pp. 136 – 142.

Khazen N., Hutton D.W., Garber A.M., Hupert N., Owens D.K. (2009). Effectiveness and cost-effectiveness of vaccination against pandemic influenza (H1N1). Ann Intern Med 151: 829–839.

Kilbourne, Edwin D. "Influenza pandemics of the 20th century." Emerging infectious diseases 12.1 (2006): 9.

Kretzschmar M., Duynhoven Y.T. van, Sverijnen A.J. 1996 Modeling prevention strategies for gonorrhea and Chlamydia using stochastic network simulations Am. J. Epidemiol., 144 (1996), pp. 306 – 317

Kwan M.P. Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: A methodological exploration with a large data set. Transportation Research Part C 2000, 8:185-203.

Leskovec J., Krause A., Guestrin C., Faloutsos C., VanBriesen J., and Glance N. Cost-effective outbreak detection in networks, in Proc. of ACM KDD, 2007.

Lin J., Jane C., and Yuan J. (1995). On Reliability Evaluation of a Capacitated-Flow Network in Terms of Minimal Pathsets, Networks, 25, 131-138.

Lloyd A.L., May R.M. 2001 Epidemiology. How viruses spread among computers and people Science, 292 (2001), p. 1316.

Longini, I. M. Jr, Koopman, J. S., Haber, M. & Cotsonis, G. A. Statistical inference for infectious diseases. Risk-specific household and community transmission parameters. Am. J. Epidemiol. 128, 845–859 (1988).

MacNab Y.C., and Dean C.B., 2002. Spatial-temporal modelling of rates for the construction of disease maps. Stat. Med. 21, 347–358.

Mangili, A. and Gendreau, M.A. (2005) Transmission of infectious diseases during commercial air travel, Lancet, 365, 989–996.

Mark D., Egenhofer M., Bian L., Rogerson P., and Vena J. Spatio-temporal GIS Analysis for Environmental Health, in: 2nd International Workshop on Geography and Medicine (GEOMED'), Paris, France, 1999.

Meyers L.A., Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. Bulletin of the American Mathematical Society 44:63–86, 2007.

Meyers L.A., Newman M.E.J., and Pourbohloul B. Predicting epidemics on directed contact networks. Journal of Theoretical Biology 240:400–418, 2006.

Meyers L.A., Pourbohloul B., Newman M.E.J., Skowronski D.M., and Brunham R.C. Network theory and SARS: Predicting outbreak diversity. Journal of Theoretical Biology 232:71 – 81, 2005.

Meyers L.A., Newman M.E.J., Martin M., and Schrag S. Applying network theory to epidemics: Control measures for Mycoplasma pneumoniae outbreaks. Emerging Infectious Diseases 9:204–210, 2003.

Miller, H. (2003). Travel Chances and Social Exclusion. Resource paper in 10th International Conference on Travel Behavior Research, Lucerne, Switzerland, 10-14th August, 2003.

Mokbel M.F., Ghanem T.M., and Aref W.G. Spatio-temporal Access Methods. IEEE Data Engineering Bulletin, 26(2), 2003.

Montgomery, D.C. and Peck, E.A. *Introduction to Linear Regression Analysis*. John Wiley and Sons, Inc., 2nd edition, 1992.

Morris M. 1995 Data driven network models for the spread of disease D. Mollison (Ed.), Epidemic Models: Their Structure and Relation to Data, Cambridge University Press, Cambridge (1995), pp. 302 – 322

Morris M., Kretzschmar M. 1997 Modeling prevention strategies for gonorrhea and Chlamydia using stochastic network simulations AIDS, 11 (1997), pp. 641 – 648.

Morse, S. S. Factors in the emergence of infectious diseases. Emerging Infect. Dis. 1, 7–15 (1995).

Mossong J., Hens N., Jit M., Beutels P., Auranen K., Mikolajczyk R., Massari M., Salmaso S., Tomba G.S., Wallinga J., Heijne J., Sadkowska-Todys M, Rosinska M., and Edmunds W.J. Social contacts and mixing patterns relevant to the spread of infectious diseases, PLoS Med 5 (2008), p. e74.

Murry, C.J.L & Lopez, A.D. 1996 "The global burden of disease: a comprehensive assessment of mortality and disability from disease" Geneva, Switzerland: World Health Organization.

Musen MA, Shahar Y, Shortliffe EH. Clinical decisionsupport systems. In: Shortliffe EH, Perreault LE, Wiederhold G, Fagan LM, eds. Medical informatics: computer applications in health care and biomedicine. 2nd ed. New York: Springer-Verlag, 2001:573-609.

Newman M.E.J. The spread of epidemic disease on networks. Phys. Rev. E, 66(016128), 2002. cond-mat/0205009.

O'Malley A.J., Christakis N.A. (2011). Longitudinal Analysis of Large Social Networks: Estimating the Effect of Health Traits on Changes in Friendship Ties. Statistics in Medicine 30: 950–964.

Ostfeld, R.S., Glass, G.E. and Keesing, F. (2005). Spatial epidemiology: an emerging (or re-emerging) discipline. Trends in Ecology and Evolution, 20, 328–336.

Parker, Richard. 2002. The Global HIV/AIDS Pandemic, Structural Inequalities, and the Politics of International Health. American Journal of Public Health 92 (3): 343-46.

Perez, L., and Dragicevic, S. An agent-based approach for modeling dynamics of contagious disease spread. International Journal of Health Geographics, 8:50, 2009.

Pourbohloul B., Meyers L.A., Skowronski D.M., Krajden M., Patrick D.M., and Brunham R.C. Modeling control strategies of respiratory pathogens, Emerg. Infect. Dis. 11 (2005), pp. 1249–1256.

Reshef D. N., Reshef Y. A., Finucane H. K., Grossman S. R., McVean G., Turnbaugh P. J., Lander E. S., Mitzenmacher M., Sabeti P. C. Detecting novel associations in large datasets. Science, 334: 1518-1524, (2011)

Rosenquist, J.N., Fowler, J.H. and Christakis, N.A. (2011). Social network determinants of depression. Molecular Psychiatry. 16(3): 273-281.

Rossion B., Delvenne J.F., Debatisse D., Goffaux V., Bruyer R., Crommelinck M., and Guérit J.M. Spatio-temporal localization of the face inversion effect: an event-related potentials study, Biol. Psychol. 50 (1999), pp. 173–189.

Ryan KJ; Ray CG (editors) (2004). Sherris Medical Microbiology(4th ed.). McGraw Hill. ISBN 0-8385-8529-9.

Salathe M., Kazandjieva M., Lee J.W., Levis P., Feldman M.W., and Jones J.H. A high-resolution human contact network for infectious disease transmission, Proc. Natl. Acad. Sci. U.S.A. 107 (2010), pp. 22020–22025.

Sander L.M., Warren C.P., Sokolov I.M., Simon C.P., Koopman J. 2002 Percolation on heterogeneous networks as a model for epidemics Math. Biosci., 180 (2002), pp. 293 – 305.

Sattenspiel L., Simon C.P. 1988 The spread and persistence of infectious diseases in structured populations Math. Biosci., 90 (1988), p. 341

Schneider K., Rainwater C., and Pohl E.A. Investigating Actor Importance in a Multi-State Social Network, Proceedings of the 2011 Industrial Engineering Research Conference.

Shaw S-L, Bombom L., and Yu H. 2008 A space-time GIS approach to exploring large individual-based spatiotemporal datasets. Transactions in GIS 12: 425–41.

Shaw, S-L and D. Wang., Handling Disaggregate Spatiotemporal Travel Data in GIS. GeoInformatica, 4(2), 161-178, 2000.

Shaw S. L., and Yu H. A GIS-based Time-geographic Approach of Studying Individual Activities and Interactions in A Hybrid Physicalvirtual Space. Journal of Transport Geography. 2009, vol.17, pp.141-149.

Soubeyrand S., Held L., Hohle M., and Sache I. Modelling the spread in space and time of an airborne plant disease, Journal of the Royal Statistical Society C 57 (2008), 253-272.

Sparks R., Carter C., Graham P.L., Muscatello D., Churches T., Kaldor J., Turner R., Zheng W., and Ryan L. Understanding sources of variation in syndromic surveillance for early warning of natural or intentional disease outbreaks, IIE Transactions 42 (9) (2010), pp. 613–631.

Tao Y., Papadias D., and Sun J. The TPR*-Tree: An Optimized Spatio-temporal Access Method for Predictive Queries. In VLDB, 2003.

Theophilides C.N., Ahearn S.C., Grady S., and Merlino M. Identifying West Nile virus risk areas: the dynamic continuous-area-space-time system, Am. J. Epidemiol. 157 (2003), pp. 843–854.

Volz E. SIR dynamics in structured populations with heterogeneous connectivity. Journal of Mathematical Biology 56:293–310, 2007.

Volz E., and Meyers L.A. Susceptible-infected-recovered epidemics in dynamic contact networks. Proceeding of the Royal Society B 274:2925–2933, 2007.

Volz E., and Meyers L.A. Epidemic thresholds in dynamic contact networks. Journal of the Royal Society Interface 6:233 – 241, 2009.

Wallinga, J., Teunis, P., & Kretzschmar, M. (2006). Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. American journal of epidemiology, 164(10), 936-944.

Watkins R.E., Eagleson S., Beckett S., Garner G., Veenendaal B., Wright G., Plant A.J. Using GIS to create synthetic disease outbreaks. BMC Med Inform Decis Mak 2007, 7:4.

Watts, Duncan (2003). Six degrees: the science of a connected age. London: William Heinemann. ISBN 0-393-04142-5.

World Health Organization Statistical Information Systems, World Health Organization Statistical Information Systems (WHOSIS) (2009);

Yu H. (2006). Spatio-temporal GIS design for exploring interactions of human activities. Cartography and Geographic Information Science 33: 3–19.

Yu H. and Shaw S. L, 2008. Exploring potential human activities in physical and virtual spaces: a spatio-temporal GIS approach. International Journal of Geographical Information Science, 22 (4) (2008), pp. 409 – 430.

Yuan M. (1996). Temporal GIS and spatio-temporal modeling. In: International Conference/Workshop Integrating GIS and Environmental Modeling, January.

APPENDIX

1. Matlab codes for the individual centrality model

```
function V = getCentrality

A = [0, 0, 0.65, 0, 0, 0, 0.40, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0;
     0, 0, 0, 0, 0.50, 0, 0, 0, 0, 0, 0.3, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0;
     0.65, 0, 0, 0, 0.60, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0;
     0, 0, 0, 0, 0, 0, 0, 0, 0, 0.50, 0.55, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0;
     0, 0.50, 0.60, 0, 0, 0, 0, 0.85, 0, 0, 0, 0.65, 0.90, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0;
     0, 0, 0, 0, 0, 0, 0, 0.70, 0, 0, 0, 0, 0, 0.70, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0;
     0.40, 0, 0, 0, 0.95, 0, 0, 0, 0, 0, 0, 0, 0.75, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0;
     0, 0, 0, 0, 0.85, 0.70, 0, 0, 0, 0, 0, 0, 0.90, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0;
     0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.40, 0, 0.45, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0;
     0, 0, 0, 0.50, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0.30, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0;
     0, 0.30, 0, 0.55, 0, 0, 0, 0, 0, 0, 0, 0, 0.90, 0, 0, 0, 0,
0, 0.75, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0;
     0, 0, 0, 0, 0.65, 0, 0, 0, 0, 0, 0.90, 0, 0, 0, 0, 0, 0,
0.70, 0, 0, 0, 0, 0, 0.50, 0, 0, 0, 0, 0, 0;
     0, 0, 0.55, 0, 0.90, 0, 0.75, 0.90, 0, 0, 0, 0, 0, 0.70,
0, 0, 0, 0.40, 0.75, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0;
     0, 0, 0, 0, 0, 0.70, 0, 0, 0.40, 0, 0, 0, 0.70, 0, 0.90,
0, 0, 0, 0, 0.70, 0, 0, 0, 0, 0, 0, 0, 0, 0.15, 0;
     0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.90, 0, 0, 0, 0,
0.75, 0, 0.65, 0, 0, 0, 0.60, 0, 0, 0, 0, 0;
     0, 0, 0, 0, 0, 0, 0, 0, 0.45, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0.70, 0, 0, 0, 0.30, 0, 0, 0, 0;
```

```
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.65,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0;
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0.30, 0.75, 0.70, 0.40, 0, 0,
0, 0.65, 0, 0, 0, 0, 0, 0.35, 0, 0, 0, 0, 0, 0, 0;
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.75, 0, 0.75, 0, 0,
0, 0, 0, 0, 0, 0.25, 0, 0, 0.35, 0, 0.30, 0, 0;
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.65, 0.70, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0;
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.65, 0, 0, 0,
0, 0, 0, 0.25, 0, 0, 0.40, 0, 0.65, 0, 0.25, 0.20;
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.70, 0, 0,
0, 0, 0.25, 0, 0, 0, 0, 0, 0, 0, 0, 0;
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.35,
0.25, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0;
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.50, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0.65, 0, 0, 0, 0, 0;
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.60, 0, 0, 0,
0, 0, 0.40, 0, 0, 0.65, 0, 0, 0, 0, 0, 0;
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.30, 0, 0,
0.35, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.60;
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0.65, 0, 0, 0, 0, 0, 0, 0, 0, 0;
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0.30, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0;
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.15, 0, 0, 0, 0,
0, 0, 0.25, 0, 0, 0, 0, 0, 0, 0, 0, 0;
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0.20, 0, 0, 0, 0, 0.60, 0, 0, 0, 0]

    n = size(A);

    MV = zeros(n);
    MV1 = zeros(n);
    MV2 = zeros(n);
    MV3 = zeros(n);


    V = zeros(1,n); % record the value of each node

    %--- The codes below is to find 3-layer nodes related to each
node.
    for i = 1:n
        for j = 1 : n
```

```
            if A(i, j) ~=0 & i ~= j
                MV1(i, j) = A(i, j);
                for k = 1: n
                   if A(j, k) ~=0 & j ~=k & i~= k
                       MV2(i, k) = MV2(i, k) + A(i, j)* A(j, k);
                       for m = 1:n
                           if A(k, m)~=0 & k~= m & i~= m & j~=m
                               MV3(i, m) = MV3(i, m) + A(i, j)*
A(j, k)* A(k, m);
                           end
                       end
                   end
                end

             end
        end
    end

    MV = MV1 + MV2 + MV3;

    for i = 1: n
        for j = 1: n
          V(i) = V(i) + MV(i,j);
        end
    end

%----------------------------------------
```

2. Matlab codes for the risk analysis model

```
function getInfect

syms x;

A = [ 540, 602, 1 ;
      608, 641, 2 ;
      647, 744, 3 ;
      760, 792, 4 ;
      810, 1030, 1];

B = [ 537, 605, 1 ;
      612, 655, 3 ;
      658, 735, 6 ;
      745, 790, 5 ;
      802, 990, 1 ;
      1007, 1148, 5];

R = [1, 25;
     2, 5;
     3, 35;
     4, 8;
     5, 20;
     6, 3;
     7, 100]

s = size(A, 1);    % get the number of rows in Matrix A
t = size(B, 1);    % get the number of rows in Matrix B

num = 0 ; % define the number of rows in the result Matrix C

for i = 1:s
  for j = 1:t

      if A(i, 3) == B(j, 3)
          if B(j, 1) <= A(i, 1)
              if B(j, 2) <= A(i, 2) & B(j, 2)>=A(i, 1)
                  num = num + 1;
                  C(num, 1) = A(i, 1);
                  C(num, 2) = B(j, 2);
                  C(num, 3) = A(i, 3);
                  C(num, 4) = R(A(i, 3), 2);
```

```
                elseif B(j, 2) >= A(i, 2)
                    num = num + 1;
                    C(num, 1) = A(i, 1);
                    C(num, 2) = A(i, 2);
                    C(num, 3) = A(i, 3);
                    C(num, 4) = R(A(i, 3), 2);
                end
            elseif B(j, 1) >= A(i, 1) & B(j, 1) <= A(i, 2)
                    if B(j, 2) >= A(i, 2)
                    num = num + 1;
                    C(num, 1) = B(j, 1);
                    C(num, 2) = A(i, 2);
                    C(num, 3) = A(i, 3);
                    C(num, 4) = R(A(i, 3), 2);

                 elseif B(j, 2) <= A(i, 2)
                    num = num + 1;
                    C(num, 1) = B(j, 1);
                    C(num, 2) = B(j, 2);
                    C(num, 3) = A(i, 3);
                    C(num, 4) = R(A(i, 3), 2);
                 end
            end

        end
    end

end

    m = size(C, 1);
    m1 = size(C, 1);
    m2 = size(C, 1);
    Eps = 0.00001;  % Epsilon is a very small positive number
    Gam = 0.2;      % Gamma is a medical parameter about the
feature of the disease
    f = 0.0;
    f1 = 0.0;
    f2 = 0.0;
    p = 0.30;
    q = 1.5;
    q1 = 2;
    q2 = 2.5;

    for k = 1: m
```

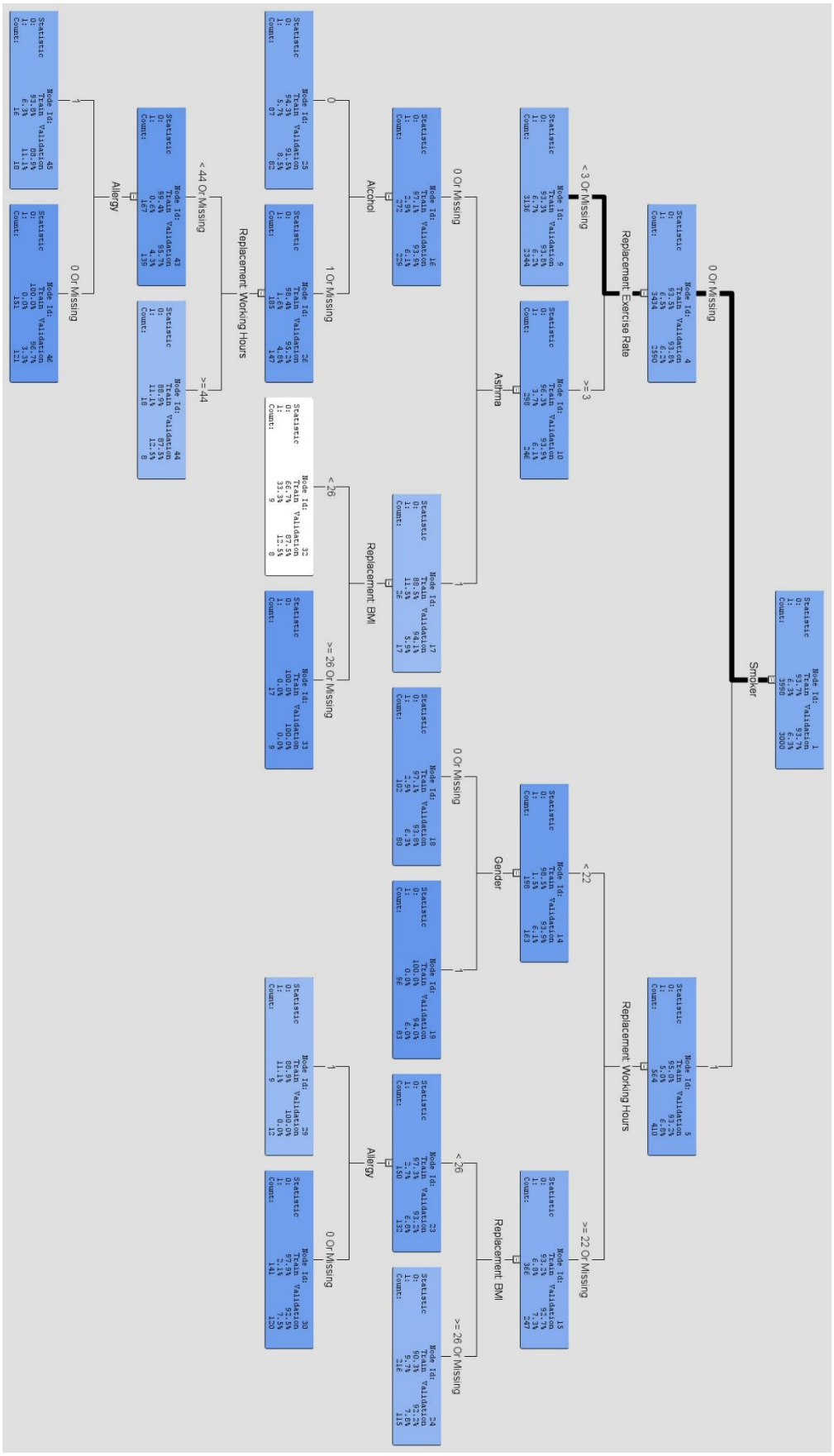
```
        fx = Gam * (x - C(k, 1))^p/ (C(k, 4)^q + Eps);
        f = f + int(fx, x, C(k, 1), C(k, 2));
    end

    g = 0.00;
    g = double(f)

     for k = 1: m1
        fx1 = Gam * (x - C(k, 1))^p/ (C(k, 4)^q1 + Eps);
        f1 = f1 + int(fx1, x, C(k, 1), C(k, 2));
    end

    g1 = 0.00;
    g1 = double(f1)

     for k = 1: m2
        fx2 = Gam * (x - C(k, 1))^p/ (C(k, 4)^q2 + Eps);
        f2 = f2 + int(fx2, x, C(k, 1), C(k, 2));
    end

    g2 = 0.00;
g2 = double(f2)
```

3. SAS Codes for Variables Statistics

```
proc capability data = sasuser.data10000;

   histogram age/normal
   midpoin = 10 20 30 40 50 60 70 80 90
   ctext = blue;
run;


proc freq data = sasuser.data10000;
 tables age;
 output out = sasuser.data10000_output;
run;


proc univariate data = sasuser.data10000;
   var age;
   histogram age/midpoints = 0 to 99 by 10;

run;


proc means data = sasuser.data10000 min max median q1 q3 range cv
skew kurt;
  var age;
run;


proc freq data = sasuser.data10000;
 tables gender alcohol allergy asthma Diabetes Family_His
Highbloodpressure illicit_drugs infected obesity
 pregnant Recent_Perscription_drugs smoker vegetarian;
 output out = sasuser.data10000_output;
run;


proc univariate data = sasuser.data10000;
   var bmi;
   histogram bmi/midpoints = 15 to 33 by 1;
run;


proc capability data = sasuser.data10000;
   var bmi;
   histogram bmi/normal
   midpoints = 15 to 33 by 1
   ctext = blue;
run;


proc means data = sasuser.data10000 min max median q1 q3 range cv
skew kurt;
  var bmi;
run;


proc means data = sasuser.data10000 min max median q1 q3 range cv
skew kurt;
```

```
  var exercise_rate;
  freq exercise_rate;
run;


proc capability data = sasuser.data10000;
   var exercise_rate;
   histogram exercise_rate/normal
   midpoints = 0 to 7
   ctext = blue;
run;

proc freq data = sasuser.data10000;
 tables exercise_rate;
 output out = sasuser.data10000_output;
run;


proc univariate data = sasuser.data10000;
   var exercise_rate;
   histogram exercise_rate/midpoints = 0 to 7 by 1;
run;


proc means data = sasuser.data10000 min max median q1 q3 range cv
skew kurt;
  var working_hours;
  freq working_hours;
run;


proc univariate data = sasuser.data10000;
   var working_hours;
   histogram working_hours/midpoints = 0 to 50 by 10;

run;


proc freq data = sasuser.data10000;
 tables working_hours;
 output out = sasuser.data10000_output;
   run;
```

4. The Structure of the Interactive Decision Tree



5. English Rules for the interactive decision tree

```
*-------------------------------------------------------------*
 Node = 9
*-------------------------------------------------------------*
if Smoker IS ONE OF: 0 or MISSING
AND Replacement: Exercise Rate < 3 or MISSING
then
 Tree Node Identifier   = 9
 Number of Observations = 3136
 Predicted: Infected=1  = 0.07
 Predicted: Infected=0  = 0.93

*-------------------------------------------------------------*
 Node = 18
*-------------------------------------------------------------*
if Smoker IS ONE OF: 1
AND Replacement: Working Hours < 22
AND Gender IS ONE OF: 0 or MISSING
then
 Tree Node Identifier   = 18
 Number of Observations = 102
 Predicted: Infected=1  = 0.03
 Predicted: Infected=0  = 0.97

*-------------------------------------------------------------*
 Node = 19
*-------------------------------------------------------------*
if Smoker IS ONE OF: 1
AND Replacement: Working Hours < 22
AND Gender IS ONE OF: 1
then
 Tree Node Identifier   = 19
 Number of Observations = 96
 Predicted: Infected=1  = 0.00
 Predicted: Infected=0  = 1.00

*-------------------------------------------------------------*
 Node = 24
*-------------------------------------------------------------*
if Smoker IS ONE OF: 1
AND Replacement: Working Hours >= 22 or MISSING
AND Replacement: BMI >= 26 or MISSING
```

```
then
 Tree Node Identifier   = 24
 Number of Observations = 216
 Predicted: Infected=1  = 0.10
 Predicted: Infected=0  = 0.90

*------------------------------------------------------------*
 Node = 25
*------------------------------------------------------------*
if Smoker IS ONE OF: 0 or MISSING
AND Replacement: Exercise Rate >= 3
AND Asthma IS ONE OF: 0 or MISSING
AND Alcohol IS ONE OF: 0
then
 Tree Node Identifier   = 25
 Number of Observations = 87
 Predicted: Infected=1  = 0.06
 Predicted: Infected=0  = 0.94

*------------------------------------------------------------*
 Node = 29
*------------------------------------------------------------*
if Smoker IS ONE OF: 1
AND Replacement: Working Hours >= 22 or MISSING
AND Replacement: BMI < 26
AND Allergy IS ONE OF: 1
then
 Tree Node Identifier   = 29
 Number of Observations = 9
 Predicted: Infected=1  = 0.11
 Predicted: Infected=0  = 0.89

*------------------------------------------------------------*
 Node = 30
*------------------------------------------------------------*
if Smoker IS ONE OF: 1
AND Replacement: Working Hours >= 22 or MISSING
AND Replacement: BMI < 26
AND Allergy IS ONE OF: 0 or MISSING
then
 Tree Node Identifier   = 30
 Number of Observations = 141
 Predicted: Infected=1  = 0.02
 Predicted: Infected=0  = 0.98
```

```
*-------------------------------------------------------------*
 Node = 32
*-------------------------------------------------------------*
if Smoker IS ONE OF: 0 or MISSING
AND Replacement: Exercise Rate >= 3
AND Replacement: BMI < 26
AND Asthma IS ONE OF: 1
then
 Tree Node Identifier   = 32
 Number of Observations = 9
 Predicted: Infected=1  = 0.33
 Predicted: Infected=0  = 0.67

*-------------------------------------------------------------*
 Node = 33
*-------------------------------------------------------------*
if Smoker IS ONE OF: 0 or MISSING
AND Replacement: Exercise Rate >= 3
AND Replacement: BMI >= 26 or MISSING
AND Asthma IS ONE OF: 1
then
 Tree Node Identifier   = 33
 Number of Observations = 17
 Predicted: Infected=1  = 0.00
 Predicted: Infected=0  = 1.00

*-------------------------------------------------------------*
 Node = 44
*-------------------------------------------------------------*
if Smoker IS ONE OF: 0 or MISSING
AND Replacement: Working Hours >= 44
AND Replacement: Exercise Rate >= 3
AND Asthma IS ONE OF: 0 or MISSING
AND Alcohol IS ONE OF: 1 or MISSING
then
 Tree Node Identifier   = 44
 Number of Observations = 18
 Predicted: Infected=1  = 0.11
 Predicted: Infected=0  = 0.89

*-------------------------------------------------------------*
 Node = 45
*-------------------------------------------------------------*
```

```
if Smoker IS ONE OF: 0 or MISSING
AND Replacement: Working Hours < 44 or MISSING
AND Replacement: Exercise Rate >= 3
AND Asthma IS ONE OF: 0 or MISSING
AND Allergy IS ONE OF: 1
AND Alcohol IS ONE OF: 1 or MISSING
then
 Tree Node Identifier   = 45
 Number of Observations = 16
 Predicted: Infected=1  = 0.06
 Predicted: Infected=0  = 0.94

*------------------------------------------------------------*
 Node = 46
*------------------------------------------------------------*
if Smoker IS ONE OF: 0 or MISSING
AND Replacement: Working Hours < 44 or MISSING
AND Replacement: Exercise Rate >= 3
AND Asthma IS ONE OF: 0 or MISSING
AND Allergy IS ONE OF: 0 or MISSING
AND Alcohol IS ONE OF: 1 or MISSING
then
 Tree Node Identifier   = 46
 Number of Observations = 151
 Predicted: Infected=1  = 0.00
 Predicted: Infected=0  = 1.00
```

6. Covariance and Correlation Matrices

Estimated Covariance Matrix																		
Parameter	Intercept	Gender	Age	Smoker	BMI	Obesity	Diabetes	Asthma	Alcohol	Recent Perscription Drugs	Illicit Drugs	Vegetarian	Exercise Rate	HighBloodPressure	Pregnant	Family His	Allergy	Working Hours
Intercept	0.094049	-0.00429	0.000196	0.000574	-0.00395	-0.00064	0.001721	-0.00272	0.001296	-0.00029	0.000666	-0.0033	-0.00004	0.001636	-0.00886	-0.00171	-0.00216	0.000039
Gender	-0.00429	0.00779	-8.47E-7	0.000053	0.000022	-0.00013	-0.00015	0.000031	-0.00014	-0.00013	-0.0006	0.000078	-0.00003	0.00017	0.004011	-0.00006	0.000127	-3.59E-6
Age	0.000196	-8.47E-7	6.446E-6	-0.00001	-0.00001	-0.00002	-0.00012	1.06E-6	-0.00007	-0.00002	-0.00002	-0.00002	-1.53E-6	-0.00012	0.000046	4.36E-6	3.634E-6	4.592E-7
Smoker	0.000574	0.000053	-0.00001	0.017097	-0.00004	-0.00033	0.000111	-0.00008	-0.00019	-0.00025	0.000922	0.000081	-0.00014	-0.00011	-0.00031	-0.00013	-0.00012	-0.00006
BMI	-0.00395	0.000022	-0.00001	-0.00004	0.000187	-0.00002	-0.00002	0.000012	-0.00007	-0.00004	-0.00002	0.000057	-1.26E-6	-0.00003	0.000159	-0.00003	5.709E-6	-5.15E-6
Obesity	-0.00064	-0.00013	-0.00002	-0.00033	-0.00002	0.009921	0.000308	0.000191	-0.00044	-0.00013	0.00065	-3.58E-6	-0.00009	-0.00013	0.000027	-0.00008	-0.00005	-0.00003
Diabetes	0.001721	-0.00015	-0.00012	0.000111	-0.00002	0.000308	0.041425	0.000227	0.000295	0.000012	0.000602	0.000195	-2.62E-6	-0.00011	0.000088	0.000247	0.000016	0.000039
Asthma	-0.00272	0.000031	1.06E-6	-0.00008	0.000012	0.000191	0.000227	0.024283	-0.00005	0.000091	-0.00079	0.000152	-0.00002	0.000145	-0.0001	0.00022	0.000091	0.000011
Alcohol	0.001296	-0.00014	-0.00007	-0.00019	-0.00007	-0.00044	0.000295	-0.00005	0.01041	-0.00027	-0.00116	0.000302	-0.0001	0.000378	-0.00038	-0.00014	-0.00047	-0.00009
Recent Perscription Drugs	-0.00029	-0.00013	-0.00002	-0.00025	-0.00004	-0.00013	0.000012	0.000091	-0.00027	0.010667	-0.00053	0.000123	-0.00002	-0.00003	-0.00064	0.000129	0.000095	-0.00002
Illicit Drugs	0.000666	-0.0006	-0.00002	0.000922	-0.00002	0.00065	0.000602	-0.00079	-0.00116	-0.00053	0.08737	0.001061	0.000039	-0.00049	-0.00273	0.000336	0.000108	-0.00003
Vegetarian	-0.0033	0.000078	-0.00002	0.000081	0.000057	-3.58E-6	0.000195	0.000152	0.000302	0.000123	0.001061	0.018948	0.000043	-0.00013	0.000682	-0.00016	-0.00035	0.000016
Exercise Rate	-0.00004	-0.00003	-1.53E-6	-0.00014	-1.26E-6	-0.00009	-2.62E-6	-0.00002	-0.0001	-0.00002	0.000039	0.000043	0.000904	-0.00002	-0.00018	-0.00009	4.359E-6	-0.00001
HighBloodPressure	0.001636	0.00017	-0.00012	-0.00011	-0.00003	-0.00013	-0.00011	0.000145	0.000378	-0.00003	-0.00049	-0.00013	-0.00002	0.013675	0.000235	0.000118	-0.00033	-2.37E-6
Pregnant	-0.00886	0.004011	0.000046	-0.00031	0.000159	0.000027	0.000088	-0.0001	-0.00038	-0.00064	-0.00273	0.000682	-0.00018	0.000235	0.065537	0.000276	0.000354	-0.00003
Family His	-0.00171	-0.00006	4.36E-6	-0.00013	-0.00003	-0.00008	0.000247	0.00022	-0.00014	0.000129	0.000336	-0.00016	-0.00009	0.000118	0.000276	0.016258	-0.00024	8.141E-6
Allergy	-0.00216	0.000127	3.634E-6	-0.00012	5.709E-6	-0.00005	0.000016	0.000091	-0.00047	0.000095	0.000108	-0.00035	4.359E-6	-0.00033	0.000354	-0.00024	0.022936	5.802E-6
Working Hours	0.000039	-3.59E-6	4.592E-7	-0.00006	-5.15E-6	-0.00003	0.000039	0.000011	-0.00009	-0.00002	-0.00003	0.000016	-0.00001	-2.37E-6	-0.00003	8.141E-6	5.802E-6	7.979E-6

Estimated Correlation Matrix																		
Parameter	Intercept	Gender	Age	Smoker	BMI	Obesity	Diabetes	Asthma	Alcohol	Recent Perscription Drugs	Illicit Drugs	Vegetarian	Exercise Rate	HighBloodPressure	Pregnant	Family His	Allergy	Working Hours
Intercept	1.0000	-0.1586	0.2523	0.0143	-0.9418	-0.0209	0.0276	-0.0569	0.0414	-0.0093	0.0073	-0.0782	-0.0039	0.0456	-0.1128	-0.0437	-0.0465	0.0451
Gender	-0.1586	1.0000	-0.0038	0.0046	0.0183	-0.0147	-0.0084	0.0023	-0.0153	-0.0138	-0.0231	0.0064	-0.0108	0.0164	0.1775	-0.0053	0.0095	-0.0144
Age	0.2523	-0.0038	1.0000	-0.0384	-0.4235	-0.0946	-0.2265	0.0027	-0.2529	-0.0764	-0.0329	-0.0572	-0.0200	-0.4121	0.0703	0.0135	0.0095	0.0640
Smoker	0.0143	0.0046	-0.0384	1.0000	-0.0248	-0.0255	0.0042	-0.0042	-0.0144	-0.0182	0.0239	0.0045	-0.0366	-0.0073	-0.0092	-0.0078	-0.0063	-0.1554
BMI	-0.9418	0.0183	-0.4235	-0.0248	1.0000	-0.0125	-0.0063	0.0057	-0.0480	-0.0273	-0.0057	0.0303	-0.0031	-0.0158	0.0453	-0.0154	0.0028	-0.1331
Obesity	-0.0209	-0.0147	-0.0946	-0.0255	-0.0125	1.0000	0.0152	0.0123	-0.0435	-0.0122	0.0221	-0.0003	-0.0301	-0.0112	0.0011	-0.0060	-0.0036	-0.0890
Diabetes	0.0276	-0.0084	-0.2265	0.0042	-0.0063	0.0152	1.0000	0.0072	0.0142	0.0005	0.0100	0.0070	-0.0004	-0.0048	0.0017	0.0095	0.0005	0.0674
Asthma	-0.0569	0.0023	0.0027	-0.0042	0.0057	0.0123	0.0072	1.0000	-0.0032	0.0057	-0.0171	0.0071	-0.0047	0.0079	-0.0025	0.0111	0.0039	0.0261
Alcohol	0.0414	-0.0153	-0.2529	-0.0144	-0.0480	-0.0435	0.0142	-0.0032	1.0000	-0.0257	-0.0385	0.0215	-0.0340	0.0316	-0.0144	-0.0108	-0.0304	-0.3187
Recent Perscription Drugs	-0.0093	-0.0138	-0.0764	-0.0182	-0.0273	-0.0122	0.0005	0.0057	-0.0257	1.0000	-0.0173	0.0086	-0.0079	-0.0028	-0.0243	0.0098	0.0061	-0.0652
Illicit Drugs	0.0073	-0.0231	-0.0329	0.0239	-0.0057	0.0221	0.0100	-0.0171	-0.0385	-0.0173	1.0000	0.0261	0.0044	-0.0142	-0.0361	0.0089	0.0024	-0.0306
Vegetarian	-0.0782	0.0064	-0.0572	0.0045	0.0303	-0.0003	0.0070	0.0071	0.0215	0.0086	0.0261	1.0000	0.0105	-0.0082	0.0193	-0.0090	-0.0166	0.0415
Exercise Rate	-0.0039	-0.0108	-0.0200	-0.0366	-0.0031	-0.0301	-0.0004	-0.0047	-0.0340	-0.0079	0.0044	0.0105	1.0000	-0.0058	-0.0229	-0.0229	0.0010	-0.1554
HighBloodPressure	0.0456	0.0164	-0.4121	-0.0073	-0.0158	-0.0112	-0.0048	0.0079	0.0316	-0.0028	-0.0142	-0.0082	-0.0058	1.0000	0.0078	0.0079	-0.0184	-0.0072
Pregnant	-0.1128	0.1775	0.0703	-0.0092	0.0453	0.0011	0.0017	-0.0025	-0.0144	-0.0243	-0.0361	0.0193	-0.0229	0.0078	1.0000	0.0084	0.0091	-0.0382
Family His	-0.0437	-0.0053	0.0135	-0.0078	-0.0154	-0.0060	0.0095	0.0111	-0.0108	0.0098	0.0089	-0.0090	-0.0229	0.0079	0.0084	1.0000	-0.0125	0.0226
Allergy	-0.0465	0.0095	0.0095	-0.0063	0.0028	-0.0036	0.0005	0.0039	-0.0304	0.0061	0.0024	-0.0166	0.0010	-0.0184	0.0091	-0.0125	1.0000	0.0136
Working Hours	0.0451	-0.0144	0.0640	-0.1554	-0.1331	-0.0890	0.0674	0.0261	-0.3187	-0.0652	-0.0306	0.0415	-0.1554	-0.0072	-0.0382	0.0226	0.0136	1.0000

7. SAS Score Node Code

```
*-------------------------------------------------------------
*;
*-------------------------------------------------------------
*;
* TOOL: Input Data Source;
* TYPE: SAMPLE;
* NODE: Ids;
*-------------------------------------------------------------
*;
*-------------------------------------------------------------
*;
* TOOL: Statistics Exploration;
* TYPE: EXPLORE;
* NODE: Stat;
*-------------------------------------------------------------
*;
*-------------------------------------------------------------
*;
* TOOL: Partition Class;
* TYPE: SAMPLE;
* NODE: Part;
*-------------------------------------------------------------
*;
*-------------------------------------------------------------
*;
* TOOL: Extension Class;
* TYPE: MODIFY;
* NODE: Repl;
*-------------------------------------------------------------
*;
* ;
*Variable: Age;
* ;
Label REP_Age= 'Replacement: Age';
REP_Age= Age;
if Age ne . and Age<-35.89243376 then REP_Age=-35.89243376;
if Age ne . and Age>112.63180345 then REP_Age=112.63180345;
*-------------------------------------------------------------
*;
* TOOL: Imputation;
* TYPE: MODIFY;
* NODE: Impt;
*-------------------------------------------------------------
*;
*-------------------------------------------------------------
*;
* TOOL: Variable selection Class;
* TYPE: EXPLORE;
* NODE: Varsel;
*-------------------------------------------------------------
*;
```

```
*-------------------------------------------------------------
*;
* TOOL: Neural;
* TYPE: MODEL;
* NODE: Neural;
*-------------------------------------------------------------
*;
**********************************;
*** Begin Scoring Code for Neural;
**********************************;
DROP _DM_BAD _EPS _NOCL_ _MAX_ _MAXP_ _SUM_ _NTRIALS;
 _DM_BAD = 0;
 _NOCL_ = .;
 _MAX_ = .;
 _MAXP_ = .;
 _SUM_ = .;
 _NTRIALS = .;
 _EPS =              1E-10;
LENGTH _WARN_ $4
;
      label Illicit_Drugs0 = 'Dummy: Illicit_Drugs=0' ;

      label I_Infected = 'Into: Infected' ;

      label U_Infected = 'Unnormalized Into: Infected' ;

      label P_Infected1 = 'Predicted: Infected=1' ;

      label P_Infected0 = 'Predicted: Infected=0' ;

      label  _WARN_ = "Warnings";

*** Generate dummy variables for Illicit_Drugs ;
drop Illicit_Drugs0 ;
if missing( Illicit_Drugs ) then do;
   Illicit_Drugs0 = .;
   substr(_warn_,1,1) = 'M';
   _DM_BAD = 1;
end;
else do;
   length _dm12 $ 12; drop _dm12 ;
    _dm12 = put( Illicit_Drugs , BEST12. );
   %DMNORMIP( _dm12 )
   if  _dm12 = '0'  then do;
      Illicit_Drugs0 = 1;
   end;
   else if  _dm12 = '1'  then do;
      Illicit_Drugs0 = -1;
   end;
   else do;
      Illicit_Drugs0 = .;
      substr(_warn_,2,1) = 'U';
      _DM_BAD = 1;
   end;
end;
```

```
*** ************************;
*** Checking missing input Interval
*** ************************;

IF NMISS(
   REP_Age   ) THEN DO;
   SUBSTR(_WARN_, 1, 1) = 'M';

   _DM_BAD = 1;
END;
*** ************************;
*** Writing the Node intvl ;
*** ************************;
*** ************************;
*** Writing the Node bin ;
*** ************************;
*** ************************;
*** Writing the Node Infected ;
*** ************************;
IF _DM_BAD EQ 0 THEN DO;
   P_Infected1  =    0.00056629486246 * REP_Age ;
   P_Infected1  = P_Infected1  +   -0.11940844734222 *
Illicit Drugs0 ;
   P_Infected1  =   -2.61417739119657 + P_Infected1 ;
   P_Infected0  = 0;
   _MAX_ = MAX (P_Infected1 , P_Infected0 );
   _SUM_ = 0.;
   P_Infected1  = EXP(P_Infected1  - _MAX_);
   _SUM_ = _SUM_ + P_Infected1 ;
   P_Infected0  = EXP(P_Infected0  - _MAX_);
   _SUM_ = _SUM_ + P_Infected0 ;
   P_Infected1  = P_Infected1  / _SUM_;
   P_Infected0  = P_Infected0  / _SUM_;
END;
ELSE DO;
   P_Infected1  = .;
   P_Infected0  = .;
END;
IF _DM_BAD EQ 1 THEN DO;
   P_Infected1  =    0.06253126563281;
   P_Infected0  =    0.93746873436718;
END;


*** Decision Processing;
label D_INFECTED = 'Decision: Infected' ;
label EP_INFECTED = 'Expected Profit: Infected' ;

length D_INFECTED $ 9;

D_INFECTED = ' ';
EP_INFECTED = .;

*** Compute Expected Consequences and Choose Decision;
_decnum = 1; drop _decnum;
```

```
D_INFECTED = '1' ;
EP_INFECTED = P_Infected1 * 1 + P_Infected0 * 0;
drop _sum;
_sum = P_Infected1 * 0 + P_Infected0 * 1;
if _sum > EP_INFECTED + 4.547474E-13 then do;
   EP_INFECTED = _sum; _decnum = 2;
   D_INFECTED = '0' ;
end;


*** End Decision Processing ;
*** ************************;
*** Writing the I_Infected  AND U_Infected ;
*** ************************;
_MAXP_ = P_Infected1 ;
I_Infected  = "1          " ;
U_Infected  =                1;
IF( _MAXP_ LT P_Infected0  ) THEN DO;
   _MAXP_ = P_Infected0 ;
   I_Infected  = "0          " ;
   U_Infected  =                0;
END;
*******************************;
*** End Scoring Code for Neural;
*******************************;
drop S_:;
*-------------------------------------------------------------
*;
* TOOL: Model Compare Class;
* TYPE: ASSESS;
* NODE: MdlComp;
*-------------------------------------------------------------
*;
if (P_Infected1 ge 0.06407466187056) then do;
b_Infected = 1;
end;
else
if (P_Infected1 ge 0.06360085914664) then do;
b_Infected = 2;
end;
else
if (P_Infected1 ge 0.06313032365367) then do;
b_Infected = 3;
end;
else
if (P_Infected1 ge 0.06299648237857) then do;
b_Infected = 4;
end;
else
if (P_Infected1 ge 0.06286290581796) then do;
b_Infected = 5;
end;
else
if (P_Infected1 ge 0.06272959352898) then do;
b_Infected = 6;
end;
```

```
else
if (P_Infected1 ge 0.06259654506911) then do;
b_Infected = 7;
end;
else
if (P_Infected1 ge 0.06246375999629) then do;
b_Infected = 8;
end;
else
if (P_Infected1 ge 0.06233123786882) then do;
b_Infected = 9;
end;
else
if (P_Infected1 ge 0.06219897824541) then do;
b_Infected = 10;
end;
else
if (P_Infected1 ge 0.0620999555309) then do;
b_Infected = 11;
end;
else
if (P_Infected1 ge 0.06196815422896) then do;
b_Infected = 12;
end;
else
if (P_Infected1 ge 0.06183661421962) then do;
b_Infected = 13;
end;
else
if (P_Infected1 ge 0.0617053350631) then do;
b_Infected = 14;
end;
else
if (P_Infected1 ge 0.06160704661605) then do;
b_Infected = 15;
end;
else
if (P_Infected1 ge 0.061476222895) then do;
b_Infected = 16;
end;
else
if (P_Infected1 ge 0.06137827552057) then do;
b_Infected = 17;
end;
else
if (P_Infected1 ge 0.06128047401185) then do;
b_Infected = 18;
end;
else
if (P_Infected1 ge 0.06115029858309) then do;
b_Infected = 19;
end;
else
do;
b_Infected = 20;
end;
```

```
*-------------------------------------------------------------
*;
* TOOL: Score Node;
* TYPE: ASSESS;
* NODE: Score;
*-------------------------------------------------------------
*;
*-------------------------------------------------------------
*;
* Score: Creating Fixed Names;
*-------------------------------------------------------------
*;
LABEL EM_SEGMENT = 'Segment';
EM_SEGMENT = b_Infected;
LABEL EM_EVENTPROBABILITY = 'Probability for level 1 of
Infected';
EM_EVENTPROBABILITY = P_Infected1;
LABEL EM_PROBABILITY = 'Probability of Classification';
EM_PROBABILITY =
max(
P_Infected1
,
P_Infected0
);
LENGTH EM_CLASSIFICATION $%dmnorlen;
LABEL EM_CLASSIFICATION = "Prediction for Infected";
EM_CLASSIFICATION = I_Infected;
```

CURRICULUM VITAE

Name: Shengpeng Jin

Address: Department of Industrial Engineering
J.B. Speed School of Engineering
University of Louisville
Louisville, KY 40292

Education: Bachelor of Science, Automation
Dalian University of Technology
2002-2006
Master of Science, Control Theory and Control Engineering
Dalian University of Technology
2006-2009
Master of Science, Industrial Engineering
University of Louisville
2009-2011

Awards & Honors:

Graduate Fellowship, University of Louisville
2009-2011
University Outstanding Student Scholarship, Dalian University of Technology
2002-2005

Work Experience:

Teaching Assistant
Industrial Engineering, University of Louisville, 2011-Represent

Professional Societies and Memberships:

Institute for Operations Research and the Management Sciences
Institute of Industrial Engineers
Vice President of INFORMS Student Chapter, University of Louisville

Publications & Presentations:

Shengpeng Jin, Suraj M. Alexander, Yang Liu, "Contact Network Based Risk Assessment to Prevent Potential Pandemics", International Journal of Collaborative Enterprise.

Shengpeng Jin, Suraj M. Alexander, Yang Liu, "Assessing Infection Risk with Contact Networks and Spatiotemporal

Activity Information", Proceedings of the 2013 Industrial and Systems Engineering Research Conference, May 18-22, 2013, San Juan, Puerto Rico.

Shengpeng Jin, Suraj M. Alexander, "Contact Network Based Risk Assessment to Prevent Potential Pandemics", INFORMS 2012 Annual Conference, October 14-17, 2012, Phoenix, Arizona.

Shengpeng Jin, Suraj M. Alexander, "Early Detection and Control of Potential Pandemics", INFORMS 2012 Annual Conference, November 13-16, 2011, Charlotte, North Carolina.

Shengpeng Jin, Suraj M. Alexander, "Early Detection and Control of Potential Pandemics", IIE 2011 Annual Conference, May 21-25, 2011, Reno, Nevada.

Shengpeng Jin, Suraj M. Alexander, "A Spatiotemporal GIS Approach for Agent-Based Potential Model", INFORMS 2010 Annual Conference, November 7-10, 2010, Austin, Texas.