CHAPTER I

INTRODUCTION

1.1 Background

The amount of demise caused by the 1918 influenza pandemic, otherwise known as Spanish Influenza was at least 40 million. This number is higher than the number of lives lost in World War I which expanded over 4 years prior to the pandemic itself [1]. Even after the invention of vaccines and various medications to counteract in the following years, the illness carried on and developed, and in recent decades it still brings a burden to both physical and economic situations in the world. In 2017, Centers for Disease Control and Prevention (CDC) approximated that flu-related hospitalizations that happened ranged from 140,000 to 710,000 and flu-related deaths ranged from 12,000 to 56,000 since 2010 [2].

Influenza-like illness itself is defined by the CDC as "fever (temperature of 100°F [37.8°C] or greater) and a cough and/or a sore throat without a known cause other than influenza [3]." CDC records the number of patients admitted with influenza-like illness in the United States by collecting information from volunteer public health departments at the state and local level and reports it on a weekly basis year-round. Because of the survey-nature of the data, by the time CDC releases the information those numbers will have been several weeks old and all interventions will not suffice. Other methods of collecting data are consultation rates of general practitioners [4] and school or workforce absenteeism figures [5].

Recently, many research has provided proof that user-generated data gives a better picture as to the real-time condition of current health issues; hence, different sources were tried such as web search logs [6, 7] and search logs in media platforms such as Twitter [8]. This particular field of research labeled as Computational Health has branched out a particular example of ILI rates modeling where it is shown that Yahoo search queries have a strong correlation with virological surveillance data [9] and another notable method has been turned into an applicative measure in the form of Google Flu Trends (GFT).

GFT was based on the works of Ginsberg et al. [6] and was launched in 2008. It is no longer publishing, but it was a notable success, showing a correlation between the frequency of web queries to ILI rates in the U.S. by comparing their predictions to the data CDC provided. Furthermore, in the following years, Lampos et al. [7] presented an automated tool with a web interface to track ILI in several regions of the United Kingdom using Twitter's microblogging service called Flu

Detector [8]. Flu Detector uses data from Twitter feed by selecting a group of features using encyclopedic and informal references that are related to influenza and flu-related word clusters created by Google Sets. They use a method called Bolasso, derived from LASSO method to ultimately compute flu score on a daily basis. Researchers hypothesized that the use of such information off media platform reduces the error in predictions.

Other approaches to forecast influenza models have also been done, such as using a state-space SIR model [10], time series model, or even the combination of different predictions into one forecast [11]. All of these attempts came to the same conclusion that predicting an exact model is literally impossible.

It is the reason why this thesis aims to model the probability distribution of ILI levels in the future, with methods proposed by Bollmann and Scherer [12] as the primary base of this work. Bollmann and Scherer applied time series modeling and pair-copula approach to create a simulation that shows the shape probability of influenza-like illness level in the future using historical data. The desired model produced from this work might be useful to health insurers and public health facilities who have to consider the costs from influenza epidemics in their financial planning, to estimate the necessary reserves to cover ILI-related claims or to perform stress tests to see the worst-case scenario of hospitalizations.

1.2 Problem Statement

The goal is to forecast the development of ILI activity in the future, using observed CDC-recorded ILI activity in the United States. The data collected by CDC is the numbers of patients admitted into over 3,000 hospitals and doctors' offices across the states every one to two weeks. Specific research questions include:

1. what characteristics do the data distribution have?
2. how can we create the desired model for each region?
3. how can we add in the interdependencies of regions into the model?
4. how will the future ILI spread look like?

The secondary research question is:

1. what possible applications can the simulated forecast have?

1.3 Objectives

The main objective this thesis strive to achieve is to verify our method through experiments accordingly to our problem statement. Below is listed an outline of our objective in this thesis:

1. develop a mathematical model using ARMA-GARCH models and the pair-copula construction to simulate paths of future ILI levels in the regions (jointly),
2. implement said model for a certain time horizon (1 year) in the form of a timeplot to depict what the ILI levels will look like in those time period, and
3. describe and interpret the results.

To achieve the objectives described above, this thesis proposed several steps, which are listed below:

1. review related research and literature,
2. transform CDC-given data to get a stationary time series,
3. model influenza-like illness activity using ARMA-GARCH models,
4. model the regional interdependencies using copula approach, and
5. use previous steps to create a future simulation of ILI activity.

1.4 Restrictions and Assumptions

The CDC's ILI data represent the collection of outpatient data from hospitals and doctors' offices across the U.S., and the data is supposedly related to the proportion of the U.S. population infected with influenza. The numbers recorded include patient visits and the number of those visits that were seen for ILI (defined as fever above 100°F and a cough or sore throat without a known cause other than influenza) which means that:

1. the data does not contain information about the proportion of the population that does not seek treatment, and
2. the data does not contain information about the proportion of the people seeking treatment who have flu symptoms but not the flu virus.

1.5 Benefits

Influenza-like illness level forecasting will hopefully be able to provide some benefits as listed below.

1.5.1 Theoretical Benefits

1. Shows the expected ILI levels dynamics in the future.
2. Calculates the accuracy of real ILI levels to follow the expected dynamics.

1.5.2 Practical Benefits

1. Helps in predicting the costs that might be caused by future ILI outbreaks in the future.
2. Allows health officials to predict national ILI levels in the future.

1.6 Thesis Structure

The writing structure of this thesis is as detailed below:

1. Chapter I describes the background, problem statement, objectives, restrictions, and methodology of this thesis.
2. Chapter II describes ARMA-GARCH model that will be used in this project. It opens with a brief history on ARMA model and GARCH model, complete with their uses mainly in the financial and economics area to forecast. Then, a brief description of ARMA model will be given since it is the more common time series sequence of the two, followed by a pure GARCH model description to show a comparison between each individual equations and how they are combined. These will be followed and closed by time series characteristics, theories, and regression concepts that were used in the process of simulating and inferring this thesis.
3. Chapter III describes the pair-copula method used in the thesis. Starting with the concept of a basic copula, followed by the concept of pair-copula mode and its methodological background. Then, this chapter will be closed with a literature review of sources concerning research using methods comparable to the ones used in this thesis.
4. Chapter IV contains the process of transforming raw given data to fit the desired model and how it will be able to forecast the future values. This chapter will also explain in detail of the data input itself and its minutiae.
5. Chapter V contains the experiment and results of implementing what have already been planned and mapped out in the previous chapter, diving into the technical part of the process and analyzing the results given to hopefully form a sustainable mathematical model and be able to forecast future ILI levels.
6. Chapter VI contains the summarization of this thesis' results, concludes and answers the problem statements and objectives stated in Chapter I, and gives suggestions for future research.