5-2017

GIS INTERNSHIP AT UMASS MEDICAL SCHOOL INFORMATION TECHNOLOGY DEPARTMENT

Qiming Shi M.S.
Clark University, qshi@clarku.edu

GIS INTERNSHIP AT UMASS MEDICAL SCHOOL INFORMATION TECHNOLOGY DEPARTMENT

Qiming Shi

Degree will be conferred May 2017

A GISDE final project paper

submitted to the faculty of Clark University, Worcester, Massachusetts,

in partial fulfillment of the requirements for the degree of

Masters of Science in Geographic Information Sciences for Development and Environment

in the Department of International Development, Community, and Environment

Accepted on the recommendation of

Dr. Yelena Ogneva-Himmelberger, Project Advisor

Abstract

GIS INTERNSHIP AT UMASS MEDICAL SCHOOL IFORMATION TECHNOLOGY

Qiming Shi

This paper summarizes my internship at the University of Massachusetts Medical School Information Technology Department in Worcester, MA, which lasted for 1 year. I worked on several projects correlating health and disease outcomes to spatial and temporal factors where I not only brought my technical skills in GIS but also turned my solution into a useful web application, using R Shiny. It was a great learning and professional experience, and I learned a great deal about R programming and public health GIS. I would highly recommend Internship at Umass Medical School to anyone who is interested in the field of public health.

Academic History

Name: Qiming Shi Date: December, 2016

Place of Birth: Hohhot, China Date: August 31st, 1992

Baccalaureate School: China University of Geoscience Date: Jun 2015

Baccalaureate Subject: Earth Information Science and Technology

Occupation and Academic Connection since Baccalaureate Degree:

GIS Intern – Umass Medical School Information Technology, Worcester, MA

Dedication

I would like to dedicate this final paper to my mom, Yuanting Qu and my family, who have always been there for me and supported me all the time.

Acknowledgements

Several people have helped me during the Internship. First, I would like to give a big thanks to Dr. Vasu Chandrasekaran, my supervisor and the Team Lead of Data Science and Informatics at Umass IT Department who gave me this Internship opportunity. He is a great person to work for; even though he did not have too much time for communicating with me due to his busy schedule, he tried his best to mentor and guide me, and I learned a lot from him. I would also like to thank Zhongzheng Shu, another intern in Umass Medical School Information Technology. He always helped me in debugging R and Python program. I would also like to thank mentor Stanislav Prodanov for his support on career development.

Table of Contents

Figures

CHAPTER 1: INTRODUCTION

When in my undergraduate school, China University of Geoscience, I majored in Earth Information Science & Technology. My undergraduate major major mostly focused on geology and remote sensing. The reason that I choose GIS as my graduate degree was that I performed a landslide analysis using ArcGIS. I found that GIS was so attractive to me in terms of its mapping capacity and spatial functionality. After I came to Clark University, I found GIS not only conducted research on environment issues like pollution management, environment surveillance, and forest degradation but also humanity issues including epidemiology, food desert or planning. The humanity part just drew my attention and finally I found I wanted to devote myself to humanity field. Clark University gave students many choices based on students' interests. Clark University allowed us to have different career tracks and different GIS research directions. I was so honored that I joined the Clark University community and well-known GISDE group. My studies have matured and benefited from living in such a diverse and dynamic community.

My internship with the Umass Medical School IT Department allowed me to utilize my GIS skills to aid the discovery and intervention of diseases like C-diff and Influenza around central Massachusetts. I worked in the Albert Sherman Center in the main Umass campus in Worcester, where many of the main offices of different Umass department are also located. The main campus was a great place to work because of its diverse cultural and historical venues that provided opportunities for self-enrichment.

I was given the title of GIS Specialist and my main role was to visualize patient data to discover the pattern of disease outbreaks spatially and temporally. Throughout my time at Umass, I learned a lot about the real utility and capacity of GIS in medical field. Overall, I had a great experience and learned a great deal of skills that will undoubtedly be useful in my transition from school to career.

CHAPTER 2: DESCRIPTION OF ORGANIZATION

The University of Massachusetts Medical School was chartered in 1962 and opened in 1970. University of Massachusetts (UMASS) Medical School was one of the cohorts of medical schools founded in response to fears of a physician shortage. In Massachusetts, this translated into a call for more opportunities for the state's students to attend an affordable school where, it was hoped, they would deliver primary care to the people of their home state. (More, 2012)

2.1 Mission

The department I entered for was Data Sciences & Technology, a sub-department of the IT department, which is located at Albert Sherman Center 9th floor at Main Campus. Data Sciences & Technology maintains a variety of hardware and software applications for bioinformatics, biostatistics, and High Performance Computing (HPC). It provides clinical consulting services to solve researchers' important technology needs. Its mission is to create an effective and flexible computing facility to enable scientific computing for clinicians, researchers, and students.

2.2 Organizational Structure

The IT Departments at the University of Massachusetts Medical School are located at 333 South Street, Shrewsbury and Umass Medical School's main campus in Worcester, Massachusetts. It includes six main sub-departments; Academic computing services, data science & technology, customer services, engineering, infrastructure services and IT security. Each sub-department has its own projects and they cooperate with each other to accomplish the final goal of serving customers and partners. People in the IT department come from different parts of the world including Asia, America, Africa and Europe, which makes UMass Medical School a diverse place to work. The Data Science team works from the Albert Sherman Center, which was built in 2013 and work and personal life are well balanced at Umass,

colleagues take care of each other and they all collaborate with each other to achieve the different tasks which include data collection, data integration, data process and data analysis.

There are three sections in the department, including High Performance Computing, Clinical Data Solutions and Laboratory IT Solutions. "High Performance Computing (HPC) uses distributed computational cycles to decrease the amount of time a single job would take. HPC processing jobs typically consist of searching or time and process jobs. Processing string searching for genomic data comparisons assists with the speed-up of "needle" and "haystack" processing and analysis. Researchers utilize custom and open software to analyze, distribute, and calculate large data sets. Utilizing best practices users can decrease job run time several fold by using distribution and cluster related protocols such as Message Passing Interface (MPI). Examples of HPC distribution include: Monte Carlo computations, time and space computations, and string (over DNA, etc.) matching algorithms. Users can create programs and scripts on the cluster here at UMASS for pattern matching and general search based needs; for example, using HG18 we can create simple shell script(s) using the Perl scripting language for effective pattern matching. HPC team can assist with these needs and help create optimal routines as needed." (High Performance Computing,2014)

"The Clinical Data Team developed and maintained MiCARD, which is the UMass implementation of the i2b2 informatics platform for clinical research. MiCARD helps researchers overcome one of the greatest problems in population-based research; rapidly compile large groups of well characterized patients. UMass Medical School is a member of the REDCap Consortium. REDCap (Research Electronic Data Capture) is a mature, secure web application for building and managing online surveys and databases. REDCap was specifically designed to support data capture for research studies. Services include the harmonization of data from multiple systems and formats into a single data repository. The Clinical Data Team can take data from these data marts and mine it using complex

algorithms to develop robust reporting that can be used to improve patient outcomes." (Clinical Data Solutions & Consulting,2014)

"Laboratory IT – Solutions & Consulting is a division of Data Sciences & Technology. The team offers many specialized services to support the variety and uniqueness of the many university laboratories:

- Research and recommend new software and cloud based technology solutions.
- Collaborate on and help manage projects that help bring in new technologies or enhance old ones for the lab
- Assist with inventory management of instrument-connected computers (Windows & Mac) and broker preventive maintenance/calibration of specialized equipment not under service agreements with our UMMS IT Productivity Services team
- Vendor coordination/management to ensure that the latest software versions are compatible with systems
- Assist to coordinate legacy instrument support and personal server conversions (local to virtual)
- Arrange virus/malware protection for Laptop, PC and Networks
- Arrange/coordinate network segmentation for systems driven by sunset Operating Systems

The Lab IT assists with Workflow/Sample Management, Analysis and Infrastructure. The team will research and recommend new technologies/tools to secure and enhance lab operations and performance.

No matter what specific research lab is undertaking, it has the need to track data (samples, sample locations, trials, etc.) across and within research cores, labs and possibly outside of the UMMS system. The lab needs to protect its data and must be able to analyze and report on the data it is gathering. Workflow and sample management solutions can offer new and far greater functionality than current offerings." (Laboratory IT Solutions & Consulting,2014) Consultation and assistance is available for, or to find lab new tools for, the following:

- Data archival, integrity/manipulation and migration to and from systems
- Data management, analytics and reporting solutions and tools
- Workflow analysis, process streamlining and automation
- Collaboration, sharing and enhanced Visualization tools

2.3 GIS and Mapping within the Department

In the Data Sciences & Technology sub-department, the GIS's ability to visualize patient's data and find spatial correlation with certain diseases just recently draw their attention, and they just began the GIS research when I came to the department. However, there were other departments which already started GIS analysis. The primary goal of my work during the summer was to find the correlation of spatial distribution of C Difflcle patients and to create an interactive website for visualizing ICD9/10 codes. ICD 9 and ICD 10 were different versions of International Classification of Disease. For every kind of disease, ICD9/10 had its corresponding code. For example, 008.45 stands for C. difficile infection, a disease which is caused by bacterium. During the summer, several experienced colleagues helped me understand certain web-mapping packages to debug my code whenever I was stuck. Our group had a meeting to present last week's accomplishments to other group members so that the whole group could know each other's process well and then set the goal for next week. Strong communication and mutual assistance made the group organized, determined and efficient.

CHAPTER 3: INTERNSHIP RESPONSIBILITIES

In my internship with the Umass Medical School IT Department, I was responsible for three GIS projects. The first project was to create a website to visualize patient's data in Umass Medical database. The second project was to generate automatic workflow to geocode patients based on ICD9/10 codes,

and to run hot spot analysis using Arcpy. The third project was to analyze spatial and temporal pattern of C-diff patients. For every patient that came to Umass Memorial, his/her record went into the Umass medical database, and address of that patient was included, so could be geocoded.

To visualize patient's data, the first step was to get the patients' addresses from Umass Medical database. One spreadsheet in the database had patient ID, street address, city/town, state and zip code. Another spreadsheet in the database included patient ID, ICD9/10 codes (International Classification of Disease), patient age and service date. Every patient had his or her own patient ID, this ID was unique to every patient, so it wouldn't change no matter how many times this patient came to Umass Memorial.

When patient information was taken from various databases, SQL Server Management Studio was used to operate the query. The goal of this query was to get the patient's address, age, service date and ICD9/10 code together in one spreadsheet. One patient ID could have many service records, but only one distinct service record would be selected. The purpose of doing distinct selection was that the count of patients in particular geographic unit would be biased if one patient had many services for one type of disease. I tried to geocode every patient based on their street addresses, but it turned out that it might take couple of days. Then I found that patients would be visualized based on town/city level, so there was no need for geocoding every patient using street address. Using town/city field in the database to sum patient numbers would save a lot of geo-processing time. Then Python Package Pandas would be used to generate a data frame containing patient counts for each ICD9/10 code based on each town/city. The problem was that python and the database were in different servers and Umass had strict rules about installing software in the server for security purposes. The way to address this problem was to use Python in another server to query the SQL database server. Python Package Pyodbc was an open source Python module that made accessing Open Database Connectivity (ODBC) databases simple. It was used to make Python's connection with database and enabled me to write SQL syntax directly in Python. For accessing database server from Python server, I had to use a remote access twice to share

one disk which contains SQL database in one server with Python server so that Python in Python server could find the location of database in SQL database server. After using Pyodbc to set up the database name and access code, the loop code could be operated to generate the patient data frame.

R studio was utilized to create a website frame work. R was a strongly functional language and environment to statistically explore data sets but it was not very strong and comprehensive to generate a website. The reason why I used R was because my supervisor thought that would be easier for me to start. Also he found a R Package called Shiny which was designed for creating website. Shiny is an open source R package which can provide user easiest and most comprehensive way to establish the framework for web application. Another package called Leaflet is also used to enable the interactive functions of website. Leaflet is the leading open-source JavaScript library for mobile-friendly interactive maps and it can be called in R.

Basic functions of web map including pan and zoom are supported. The web map also has two select drop boxes, one slider bar, one distribution graph and one legend. Each town is represented by a circle in web map. The first select drop box controls colors of the circles. Factors controlling circle colors include household income, mean age, population, patient ratio and ICD9/10 code. The second drop box controls circle size based on ICD9/10 code, so the more patients we get in one town, the bigger the circle we get on the web map. Circles shown on the map are also controlled by the slider. That means the user can select his or her thresholds of patient count and only circles in that particular range are shown on the map. The distribution graph gives us the general idea of how the factors are correlated with each other from two select drop boxes. A regression line with correlation coefficient and coefficient of determination is obtained from distribution graph. The legend informs us the color range derived from first drop box. If you want to know the detailed information about certain town represented by circle, you can click the circle that you are interested in and a popup will be triggered to show zip code, town name, specific patient count and patient population ratio of that town.

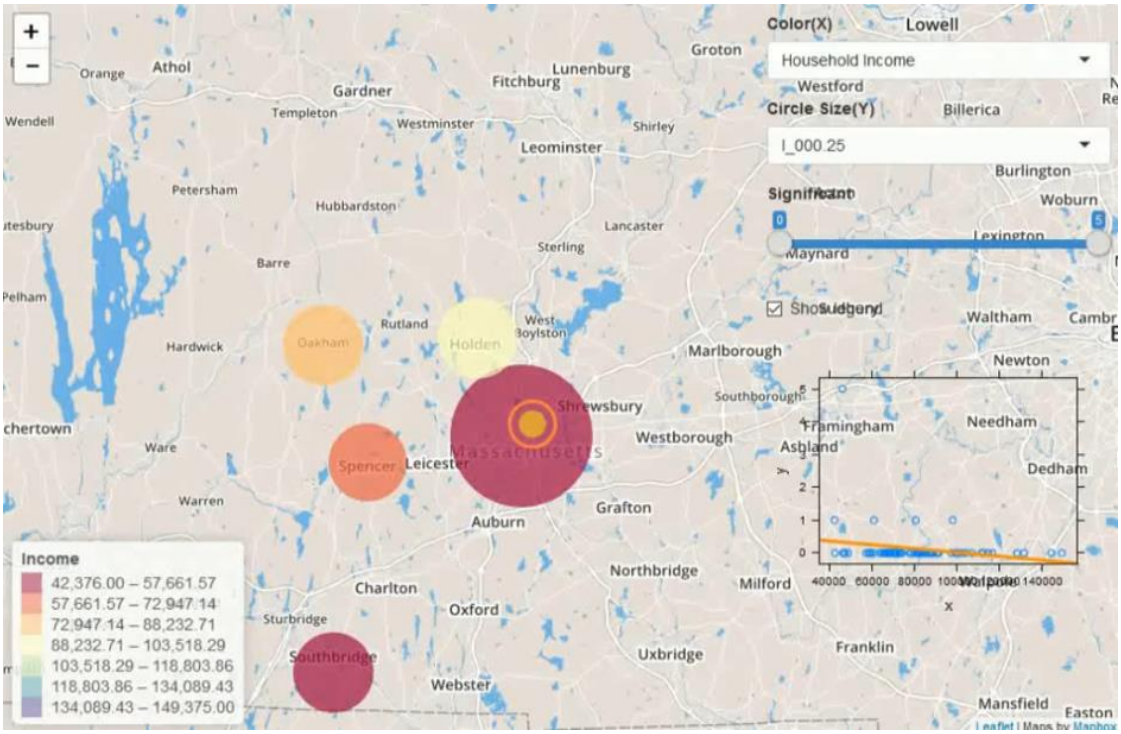
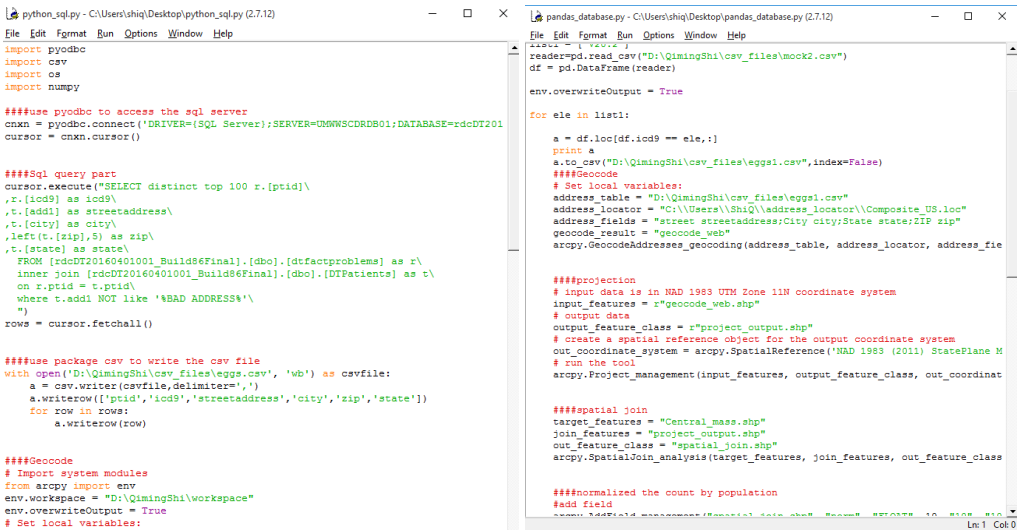


Figure 1: Web application for Umass IT

The second project is to conduct hot spot analysis using Arcpy. The purpose of this work it to generate the automatic workflow to query data in database, filter the bad address, geocode patients' address, spatial join, project, normalize patient by population and hot spot analysis. For every kind of disease, this Python code will create a layer.



```
python_sql.py - C:\Users\shiq\Desktop\python_sql.py (2.7.12)
File Edit Format Run Options Window Help
import pyodbc
import csv
import os
import numpy

####use pyodbc to access the sql server
cnxn = pyodbc.connect('DRIVER={SQL Server};SERVER=UMWWSCDRDB01;DATABASE=rdcDT201
cursor = cnxn.cursor()


####Sql query part
cursor.execute("SELECT distinct top 100 r.[ptid]\
,r.[icd9] as icd9\
,t.[add1] as streetaddress\
,t.[city] as city\
,left(t.[zip],5) as zip\
,t.[state] as state\
  FROM [rdcDT20160401001_Build86Final].[dbo].[dtfactproblems] as r\
  inner join [rdcDT20160401001_Build86Final].[dbo].[DTPatients] as t\
  on r.ptid = t.ptid\
  where t.add1 NOT like '%BAD ADDRESS%'\
  ")
rows = cursor.fetchall()


####use package csv to write the csv file
with open('D:\QimingShi\csv_files\eggs.csv', 'wb') as csvfile:
    a = csv.writer(csvfile,delimiter=',')
    a.writerow(['ptid','icd9','streetaddress','city','zip','state'])
    for row in rows:
        a.writerow(row)


####Geocode
# Import system modules
from arcpy import env
env.workspace = "D:\QimingShi\workspace"
env.overwriteOutput = True
# Set local variables:
```

```
pandas_database.py - C:\Users\shiq\Desktop\pandas_database.py (2.7.12)
File Edit Format Run Options Window Help
reader=pd.read_csv("D:\QimingShi\csv_files\mock2.csv")
df = pd.DataFrame(reader)

env.overwriteOutput = True

for ele in list1:

    a = df.loc[df.icd9 == ele,:]
    print a
    a.to_csv("D:\QimingShi\csv_files\eggs1.csv",index=False)
    ####Geocode
    # Set local variables:
    address_table = "D:\QimingShi\csv_files\eggs1.csv"
    address_locator = "C:\\Users\\ShiQ\\address_locator\\Composite_US.loc"
    address_fields = "street streetaddress;City city;State state;ZIP zip"
    geocode_result = "geocode_web"
    arcpy.GeocodeAddresses_geocoding(address_table, address_locator, address_fie


    ####projection
    # input data is in NAD 1983 UTM Zone 11N coordinate system
    input_features = r"geocode_web.shp"
    # output data
    output_feature_class = r"project_output.shp"
    # create a spatial reference object for the output coordinate system
    out_coordinate_system = arcpy.SpatialReference('NAD 1983 (2011) StatePlane M
    # run the tool
    arcpy.Project_management(input_features, output_feature_class, out_coordinat


    ####spatial join
    target_features = "Central_mass.shp"
    join_features = "project_output.shp"
    out_feature_class = "spatial_join.shp"
    arcpy.SpatialJoin_analysis(target_features, join_features, out_feature_class


    ####normalized the count by population
    #add field
Ln: 1  Col: 0
```

Figure 2: Python code for generating hot spot

The third project is to discover spatial and temporal pattern for C-diff patients. Around central Massachusetts, patient records from 1993 to 2015 are included in database. There are 12 thousand patients of Clostridium difficile infection in central Massachusetts from 1993 to 2015. The biggest discovery is that many C-diff patients are clustered in some assisted living, nursing homes and senior living facilities. Some assisted living and nursing homes have more than 100 C-difficile patients, which indicates the spread and diffusion of C-difficile patients is very severe at that location. Hot spot analysis was used to identify the high-density patient's area. Three geographical units were used to operate the analysis in census block, census block groups and census tract. These data are all from MassGIS and include demographic data. So patient data is normalized by population in case that patient counts are related to population count. Also, emerging hot spot analysis was used to identify the trend and space time cube was used to visualize the data in 3D. The result won't be shown here due to privacy concerns.

The IT EXPO has been held every year since 2013 and its purpose was to inform the public what IT department achieved last year and to find potential cooperation and investment. It contained 10 to 15 different booths including Clouding Computing, Visual Reality, Cyber Security, 3D Printing, GIS, Data Science, etc. this year. My GIS booth's name was Data Visualization and Pattern Identification. I was responsible for introducing my summer product to people from Umass or other companies who came with interests. During the IT EXPO time from 10am to 2pm on Sep 28th, I had 70 people stop by my booth as I introduced the potential GIS ability in the medical field. It was my pleasure to expose GIS power to the public. During the talk, many people were inspired by GIS power and methods in analyzing disease outbreak area, outbreak time, patients cluster and distribution trend. Many GIS methods used in preventing disease diffusion and recognizing disease spatial attributes were introduced to the public at IT EXPO.

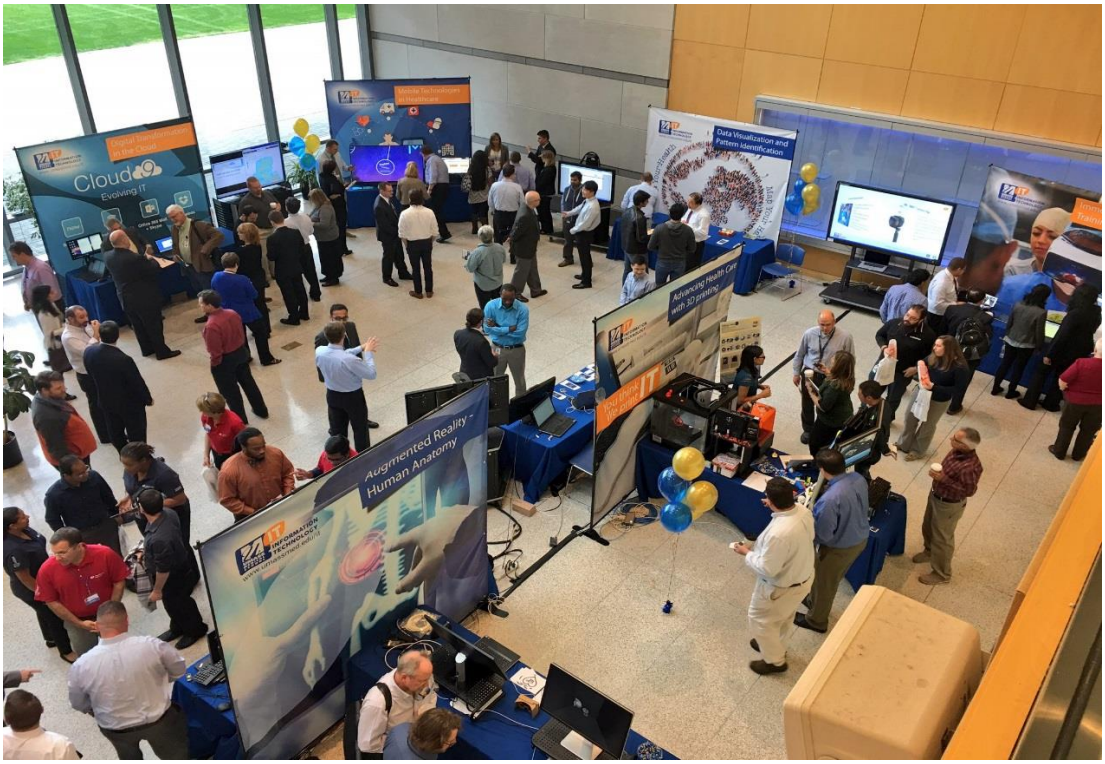


Figure 3: IT EXPO

CHAPTER 4: INTERNSHIP ASSESSMENT

My internship at Umass Medical School IT Department was an amazing experience overall. Under the great leadership of Chief Information Officer, Greg Wolf, the whole IT team was open, dynamic and energetic. This internship provided me opportunities to apply and reinforce my GIS skills in the real world, and more importantly, to learn new skills I did not learn in school. The internship also exposed me to a real working environment, which involves meetings, presentations and collaborations. This internship not only allowed me to explore how geo-spatial techniques can help in visualizing and analyzing medical data, but also trained me to become more professional and cooperative in real working environment.

In terms of new skills and techniques I learned from my internship, I have to say I didn't know approximately 70% when I just began my internship. During the first week of my internship, I had to finish the online training course "Protecting Human Research Participants" before I could do any work for my department.

I spent my first half month reading materials and papers given to me by my supervisor. Those materials were about spatial methods of visualizing medical data and spatial techniques of tracing disease diffusions. I had a lot of fun reading those papers. It helped me have a general knowledge about spatial technique history and current research stage in medical field. After I became familiar with the working environment at Umass, I devoted myself to researching the spatial pattern and data visualization of C-Diff patients. Anything related to GIS or ArcGIS was what I learned at Clark University. But I learned how to geocode addresses based on Excel table from some other resources including YouTube, Google and ArcGIS help. To get the Excel table from SQL Database, I had to learn SQL syntax from Youtube and couple SQL resources from website so that I could query the data and output the data into the desired format. This was new skill I got from my internship.

At the second month of my internship, I was asked to create a website to visualize patient's data in our database. R Shiny and Leaflet were used to establish the frame work and interactive part of the website. Packages I used for creating this website includes Shapefile, OGR, Pyodbc and Pandas. These were my favorite parts of my internship. I definitely learned a lot of new techniques and skills from making this product. I was not familiar with R programming and debugging at the beginning, but I was really confident and proficient in R programming after this product. Coursework at GISDE made this internship possible because I was well equipped with GIS knowledge from our diverse and professional GIS classes. I just had a little knowledge of ArcGIS software when I entered the Clark University. Throughout the internship, I used geocoding, spatial join, optimized hot spot and emerging hot spot analysis – all skills I learned from Clark University. I also used Python for generating the automatic work

flow for making the hot spot analysis map for all kinds of diseases. I learned how to code using Python with Professor Ylli Kellici at Clark. Couple Database packages in Python were used to finish my work, but the major part of my python code used Arcpy which was taught by professor Ylli Kellici. In terms of the website development using R, I didn't have that programming experience before the internship. I developed many skills from this project including R programming, R debugging, using R shiny, and Leaflet.

The internship aligned with my coursework in Clark University GISDE program well. I recommend that every student have some experience of web mapping before their summer interns because web-GIS becomes more popular in the job or intern market. I highly recommend our program give students opportunities to take Python class their first semester so that they can also take web mapping in the second semester before the summer internship. This way, I think GISDE students could have more opportunities on the job market.

With a great achievement of my summer internship, I definitely recommend this internship to other IDCE students. Our students would apply GIS techniques to medical patients, and gain lots of cutting edge IT techniques as well. The dynamic and diverse environment at Umass IT Department will benefit our students a lot in terms of innovative technology and strong network they would have with different departments in Umass medical school.

CHAPTER 5: CONCLUSION

I appreciated that I had the opportunity to have an internship working at Umass Medical School IT with many talented PHDs and interns. I was so pleased that my colleague at Umass taught me how to program and debug and helped me to understand the logic of computer hardware to software. This was

an awesome experience with so many things involved. During the process, I found my ability to solve the real world problems had been strengthened to another level and I thought I would benefit a lot from this internship in the future. I found that it was impossible to get the work done couple times during my internship. When I looked back to what I achieved during the internship, I found it was not that hard and unachievable. I just built up my confidence in handling different GIS projects. I am fully satisfied with this internship and would encourage anybody interested in applying GIS to epidemiology to pursue similar opportunities.

BIBLIOGRAPHY

More, Ellen S., "The University of Massachusetts Medical School, A History: Integrating Primary Care and Biomedical Research" (2012). *History of UMass Worcester.* Book 1.

Umass Medical School. (2014). High Performance Computing. Retrieved December 09, 2016, from http://www.umassmed.edu/it/services/research-computing/high-performance-computing/

Umass Medical School. (2014). Clinical Data Solutions & Consulting. Retrieved December 09, 2016, from http://www.umassmed.edu/it/services/research-computing/clinical-data-solutions--consulting/

Umass Medical School. (2014). Laboratory IT Solutions & Consulting. Retrieved December 09, 2016, from http://www.umassmed.edu/it/services/research-computing/laboratory-it-solutions--consulting/