SPATIO-TEMPORAL MODELING OF THE

US COLLEGE CRIME DATA

by

Fatih Gezer

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Master of Science in Statistics

Summer 2017

SPATIO-TEMPORAL MODELING OF THE

US COLLEGE CRIME DATA

by

Fatih Gezer

Approved: _____

Xiaoke Zhang, Ph.D.
Professor in charge of thesis on behalf of the Advisory Committee

Approved: _____

Thomas Ilvento, Ph.D.
Chair of the Department of Applied Economics and Statistics

Approved: _____

Mark Rieger, Ph.D.
Dean of the College of Agriculture and Natural Resources

Approved: _____

Ann L. Ardis, Ph.D.
Senior Vice Provost for Graduate and Professional Education

ACKNOWLEDGEMENTS

In the name of Allah, the Most Beneficent, the Most Merciful. All praise is due to Allah, for his blessings and guidance in my life. Thank you for giving me the patience and strength to complete this thesis.

I would like to express the sincere gratitude and highest appreciation to my advisor Dr. Xiaoke Zhang, for all the invaluable knowledge, guidance, support, encouragement and his accessibility whenever I have a question. I am grateful for the chance to work with him and learn from him during my graduate study. I also would like to thank my thesis committee members, Drs. Thomas Ilvento and Paul Eggermont, for their support and feedback throughout this process. I am also grateful to Drs. John Pesek, Zeki Yildiz, and Kivanc Aksoy for writing a recommendation letter for my Ph.D. applications.

Finally, I would like to thank my loving mother Saliha Gezer, my father Bayram Gezer, and my sisters Betul and Tuba for their unconditional help and support in my entire life. I also would like to thank the person who is my only one, for her trust, motivation, and encouragement on me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

College crime is one of the most alarming social problems in the US today. To investigate important factors that are associated with college crime, we collected data from several publicly accessible sources and performed exploratory and statistical analyses. For the statistical analysis, Bayesian hierarchical modeling via Markov chain Monte Carlo and stepwise model selection procedures were applied to analyze such spatio-temporal data. We found the best models for California and Texas respectively in the sense that each model not only achieves a good balance between goodness-of-fit and interpretability but also satisfies spatial stationarity. A strong autoregressive effect was found for both states. The results additionally show that the proportion of undergraduate students and tuition are the most essential predictive factors that affect the college crime rate in California, while no strong factor is founded for Texas.

Chapter 1

INTRODUCTION

1.1 Background

After the rape and murder of Jeanne Clery, a 19-year-old Lehigh College student in her campus residence hall in 1986, the US Congress passed a law called "The Jeanne Clery Disclosure of Campus Security Policy and Campus Crime Statistics Act" in 1990 (Sloan, 1994). From then on, all universities and colleges are required to participate in a federal financial aid program, to report their crime statistics. The Clery Act requires the compliance of the disclosure of crime statistics based on the Uniform Crime Reporting (UCR) definitions by the Department of Education. The types of crimes are defined and determined under the law. The necessity of reporting a campus crime covers all cases no matter if it occurs in a campus property or it is committed by a student.

Universities and colleges are valuable centers for higher education, which is critical for the prosperity and stability of a country. Therefore, it is crucial for a government to monitor incidents in campuses, and take precautions. The Clery Act establishes a mandatory action for universities and colleges to collect and report their crime statistics, so that in-depth studies can be conducted to investigate important factors, and effective methods for prevention can be proposed.

1.2 Literature Review

There are approximately 4,200 universities and colleges, with roughly 16 million students attending those universities, in the entire US. After the higher education institutions received the federal funding to collect and report their crime statistics by 1992, very little research was conducted on campus crime. Sloan (1994), which is one of the first studies, introduced a preliminary model of the correlates of campus crime and a future framework. The study focused on the academic year of 1989-1990 for 494 colleges and universities. Regarding the statistical methodology, Sloan (1994) used factor analysis and multivariate analysis, and the results showed that 64 percent of the crimes are burglary, which is the highest type of crimes committed, and 2 percent of the crimes are violent crime, which is the lowest. He also concluded that drug-related offenses are positively and safety is inversely correlated with total crimes, and the number of minorities in an institution has a correlation with violent crime and vandalism.

Another research on dating violence of the students was conducted by Sellers and Bromley (1995). They found that aggressive acts in a relationship promote dating violence. However, serious actions such as the use of a gun or a knife rarely happen. They concluded their study through the regression analysis of a survey from a group of students. They also argued that even though violent crime has a low occurrence, it is still critical to develop strategies and prevention methods to avoid its serious consequences.

After 10 years since the Clery Act was signed, its impact on students' behaviors has been investigated. Janosig & Gehring (2003a) found that 27 percent of the students

are aware of the campus crime disclosure and even fewer read the annual reports. Meanwhile, most the students feel safe in their campuses. Therefore, the secure environment of a campus may reduce students' motivation to read the regulations and reports regarding the crime statistics. Janosig & Gregory (2003b) concluded in another study that the Clery Act has a minor effect on students' behaviors.

In addition to the investigation of the efficiency of the Clery Act, there have been explorations that aim to examine the impact of social learning and social control theories in the forecast of college crime. In their correlational and regression analysis, Payne & Salotti (2007) discussed that both theories have a strong association with college crime. Moreover, students who show a continuous class attendance and have a strong communication with their professors, as well as students who are aware of related policies, are less likely to commit a crime. Also, parental attachment shows a significant impact only on the prevention of drug use. Another study by Gardella et al. (2014) justified that students who have less connection to classes and campus are the risk for crimes. On the other hand, the questionnaires from 2,230 female students showed that the increase of their attachment to campus life and the frequency of party attendance increase the possibility of their sexual victimization (Franklin et al., 2012). All three studies mentioned in this paragraph have the same conclusion about the issues about white females being a sexual target.

Besides a preliminary exploratory analysis, Nobles et. al., (2012) mapped the on and off-campus crimes in a large southeastern university. To visualize crimes, they used the geographic information systems (GIS) technique. In addition to a spatial analysis,

their binary logistic regression showed predictive factors for the incidents. More importantly, they alleviated one of the biggest limitations of the Clery Act by showing city and campus crimes on the same map. The results showed that if a university is in a city, it is crucial to include its city-level crime statistics if one wants to investigate the factors affecting the campus crime. It indicated that the Clery Act needs more transparency for the data collection of higher education institutions. The thesis by LaRue (2013) applied spatial regression to study the property crime in Ottawa, Canada and concluded that universities are significant factors, which indicates the entanglement of city and university crimes. These studies motivated us to take into account city-level crime when analyzing college crime.

With respect to statistical modeling, Luan et al. (2016) are perhaps the closest work to this thesis, where they performed a Bayesian hierarchical modeling to investigate the spatio-temporal patterns of the police calls of incidents in Waterloo, Canada. The advantages of Bayesian modeling of spatio-temporal data were highlighted and it motivated the modeling procedure of this thesis. The difference between the objective of this thesis and that of their study is that this thesis studies campus crime rather than city crime, although, as previously mentioned, city-level crime may play an important role in analyzing campus crime. To study the relationship between diversity and campus crime, Wang et al., (2012) proposed a spatio-temporal generalized additive model to predict the possibility of a crime when a specific time and location is given, where predictors that represent ethnicity were included in the modeling. Another related work is Gonzales (2015), where ordinary least squares were used to predict intentional homicide crime rate,

and youth unemployment rate, an economic factor, was found to be significantly associated with the crime rate. These studies altogether strengthen the usefulness of spatio-temporal modeling on campus crime and the necessity of investigating the effect of economic factors, and university and college features, e.g., ethnical diversity.

1.3 Objectives

The objective of this thesis is to examine the spatio-temporal characteristics of the US campus crime and to identify key predictive factors from a statistical perspective. The data set used in this thesis is combined from a variety of publicly accessible databases. Instead of restricting ourselves with the data released by the Federal Bureau of Investigation (FBI) as required by the Clery Act regulation, we obtained a more comprehensive data set that consists of potential factors, including city-level crime statistics, economic factors, and other characteristics, e.g., ethnical diversity, of universities and colleges. These factors may be related to college crime, but cannot be reflected in the released data as required by the Clery Act.

Regarding the statistical methodology of this study, Bayesian hierarchical modeling was used, which is one of the newest approaches to the analysis of college crime. The spatial and temporal information revealed in the data set motivated us to apply such a method. Due to various limitations, we only focused on two states, California and Texas, and analyzed each state separately. The findings in the two states are different, which is somewhat unsurprising considering their cultural and sociological differences.

This study may not be nationally representative because we focus on particular states due to several limitations as discussed in Chapter 6.

1.4 Outline

The rest of the thesis proceeds as follows: Chapter 2 introduces the procedure for data collection and compilation. Summary statistics, and exploratory figures and charts are provided in Chapter 3 to show the spatial and temporal characteristics of the data set. Moreover, Chapter 4 presents the statistical framework of the study and the results are given in Chapter 5. Finally, conclusions and future work are discussed in Chapter 6.

Chapter 2

DATA DESCRIPTION

2.1 Data Sources

The data set was created from three different sources: the Federal Bureau of Investigation (FBI), the National Center for Education Statistics (NCES), and the Bureau of Labor Statistics (BLS). Since the raw data from the three sources have different formats, we had to make some adjustments when we merged them. Below we provide a detailed explanation of the data collection procedure from each of the three sources and a list of the chosen variables and their meanings.

2.1.1 Federal Bureau of Investigation (FBI)

After the US Congress passed the law of Clery Act, universities and colleges are required to report their crime statistics under the definitions of UCR, which is the nationwide annual data release by the FBI regarding crimes in the United States. The FBI does not collect the data; instead, they was provided by law enforcements agencies. There might be some law enforcement agencies which do not voluntarily share their data with the FBI although they keep the records. Therefore, the raw data set from the FBI does not include all higher education institutions' crime statistics.

We can obtain two categories of the crime statistics from the FBI data set, including college crime statistics, and city crime statistics. The latter refers to the crime statistics of the city where a university or a college is located. Figure 2-1 is a snapshot of the raw data of college crimes in 2004, which shows institution names, numbers of violent and property crimes with their subcategories as well as student enrollments. The list of variables that we used and their descriptions will be explained in subsequent sections.

Offenses Known to Law Enforcement by University and College, 2004											
University/College by state	Student enrollment	**Violent crime**	Murder and non-negligent man-slaughter	Forcible rape	Robbery	Aggravated assault	**Property crime**	Burglary	Larceny-theft	Motor vehicle theft	Arson[2]
CALIFORNIA											
Allan Hancock College	13,014	0	0	0	0	0	42	20	22	0	0
California State Polytechnic University:											
Pomona	19,821	6	0	3	2	1	343	23	286	34	0
San Luis Obispo	18,453	5	0	0	2	3	171	8	161	2	1
California State University:											
Bakersfield	7,765	1	0	0	0	1	81	15	62	4	0
Channel Islands[4]		4	2	0	0	2	22	8	14	0	0
Chico	16,246	11	0	4	2	5	283	33	229	21	1
Dominguez Hills	13,504	3	0	1	1	1	114	13	81	20	0

Figure 2-1. A snapshot of some college crime statistics from the FBI in 2004.

On the other hand, city-level crime statistics are considered as a crucial factor because of its possible impact on campus crimes, which was discussed in the literature review. Similarly, the raw data obtained from the FBI releases, given in Figure 2-2, consist of cities with their populations, and violent and property crimes together with their subcategories.

Offenses Known to Law Enforcement											
by City 10,000 and over in Population, 2004											
City by state	Population	**Violent crime**	Murder and non-negligent man-slaughter	Forcible rape	Robbery	Aggravated assault	**Property crime**	Burglary	Larceny-theft	Motor vehicle theft	Arson[1]
CALIFORNIA											
Adelanto	20,233	81	6	8	11	56	523	231	202	90	9
Agoura Hills	22,035	43	1	4	6	32	345	102	210	33	5
Alameda	72,633	342	1	12	89	240	2,231	376	1,522	333	15
Albany	16,589	42	1	3	29	9	718	130	446	142	8
Alhambra	88,766	251	2	17	115	117	2,403	474	1,427	502	27
Aliso Viejo	40,917	34	0	2	5	27	495	88	366	41	7
American Canyon	13,287	27	0	7	10	10	326	73	202	51	1
Anaheim	336,195	1,530	10	102	493	925	10,249	1,912	6,388	1,949	31
Anderson	10,008	69	1	4	5	59	514	166	304	44	3

Figure 2-2. A snapshot of some city-level crime statistics from the FBI in 2004.

2.1.2 National Center for Education Statistics (NCES)

The NCES is a federal entity of the Institute of Education Sciences, which is an independent statistical research and evaluation branch of the US Department of Education. The main purpose of the NCES is to collect, analyze and report the education-related statistics in the US. It also fulfills the international aim for the standardization of the terminology and definitions as well as the comparison of worldwide educational statistics.

The NCES data set makes the most important contributions to this thesis since it provides a large number of characteristics of universities and colleges. The raw data set from the NCES is extremely big, which includes 974 variables for roughly 4,000 institutions nationwide between 1987 and 2012. However, we only used a subset of these institutions and variables in our study. To give a brief idea about the appearance of the data set, a snapshot is shown in Figure 2-3 below as a sample.

academicyear	instname	city	state	zip	sector	iclevel	control	oberegion	fte_count
2012	Kaplan College-Chesapeake	Chesapeake	VA	23320	6	2	3	5	111
2012	The University of America	Murrieta	CA	92563	2	1	2	8	27
2012	Miller-Motte Technical College	Roanoke	VA	24018	6	2	3	5	172
2012	Rio Grande Bible Institute	Edinburg	TX	78539	2	1	2	6	108
2012	San Joaquin Valley College-Temecula	Temecula	CA	92590	6	2	3	8	168
2012	Western Shores Institute Inc	Dunkirk	MD	20754	9	3	3	2	1
2012	Fortis College-Montgomery	Montgomery	AL	36117	9	3	3	5	82
2012	South University–Savannah Online	Savannah	GA	31406	3	1	3	5	22544

Figure 2-3. A snapshot of some educational statistics from the NCES in 2012.

2.1.3 Bureau of Labor Statistics (BLS)

As a principal federal agency, the BLS is a unit of the US Department of Labor, which pursues a mission on collection, determination, and report of the vast field of labor economics and statistics. The crucial benefit of the data from the BLS is its contribution to providing regional economic information.

The literature implies that the economic conditions of a place may influence the crimes at the nearby locations. Therefore, the unemployment rate of a county in which an institution is located is selected as a predictor variable, since a lower level of economic statistics for a place is not publicly available. Figure 2-4 below represents a sample of the data set from the BLS for California labor force data in 2004.

Labor Force Data by County, 2004 Annual Averages								
LAUS Code	State FIPS Code	County FIPS Code	County Name State Abbreviation	Year	Labor Force	Employed	Unemployed	Unemploy-ment Rate (%)
CN0600100000000	06	001	Alameda County, CA	2004	741,617	698,223	43,394	5.9
CN0600300000000	06	003	Alpine County, CA	2004	534	491	43	8.1
CN0600500000000	06	005	Amador County, CA	2004	16,906	15,932	974	5.8
CN0600700000000	06	007	Butte County, CA	2004	97,434	90,271	7,163	7.4
CN0600900000000	06	009	Calaveras County, CA	2004	19,956	18,606	1,350	6.8
CN0601100000000	06	011	Colusa County, CA	2004	9,513	8,216	1,297	13.6
CN0601300000000	06	013	Contra Costa County, CA	2004	507,762	480,274	27,488	5.4
CN0601500000000	06	015	Del Norte County, CA	2004	10,626	9,772	854	8.0
CN0601700000000	06	017	El Dorado County, CA	2004	88,838	84,174	4,664	5.3
CN0601900000000	06	019	Fresno County, CA	2004	404,288	362,204	42,084	10.4
CN0602100000000	06	021	Glenn County, CA	2004	11,451	10,385	1,066	9.3

Figure 2-4. A snapshot of some data from the BLS in 2004.

2.2 Variables

The total number of all variables is more than 1,000, and most of them are from the NCES. Since listing all variables here is unrealistic, we only present a small number of variables, which were considered to be useful to model the US college crime as in previous studies. Table 2-1 lists these variables, and their sources, types, and formula if any transformations and definitions.

Table 2-1. A subset of the variables from the FBI, BLS, and NCES databases.

Source	Label	Type	Formula	Definition
FBI - I	year	Num		Year
	name	Char		Name of an institution
	enroll	Num	Total enrollment	Total enrollment of an institution
	crime_total	Num	violent + property	Total crimes in an institution
	violent	Num	v1+v2+v3+v4	Total of violent crimes in an institution
	v1	Num		Murder and nonnegligent manslaughter
	v2	Num		Forcible rape
	v3	Num		Robbery
	v4	Num		Aggravated assault
	property	Num	p1+p2+p3+p4	Total of property crimes in an institution
	p1	Num		Burglary
	p2	Num		Larceny-theft
	p3	Num		Motor vehicle theft
	p4	Num		Arson
	rcrime	Num	Crime/enroll*1000	Total crimes per 1000 persons
	rviolent	Num	Formula	Violent crimes per 1000 persons
	rproperty	Num	Formula	Property crimes per 1000 persons
BLS	county	Char		County the institution belongs to
	labor	Num		Labor force
	employed	Num		Number of people employed
	unemployed	Num		Number of people unemployed
	unemp_rate	Num	unemployed/labor	Unemployment rate

NCES	instname	Char		Name of an institution
	city	Char		City an institution belongs to
	control	Cat	Public: 1, Private: 2	Publicly or Privately controlled
	hsi	Cat	No: 0, Yes: 1	Hispanic-serving institution
	cpi_scalar	Num	CPI_index/2012 Fiscal year CPI Index number	Consumer price index
	hepi_scalar	Num	HEPI_index/2012 HEPI Index	Higher education price index
	heca_scalar	Num	HECA_index/2012 HECA Index	Higher education coast adjustment
	tuition_t	Num		Total revenue from tuition and fees
	tuition	Num	tuition_t/enroll/ 1,000	Tuition per capita (rescaled by 1,000)
	gom_t	Num		Difference between total revenues and total expenditures
	gom	Num	gom_t/enroll/ 1,000,000	Difference between total revenues and total expenditures per capita (rescaled by 1,000,000)
	undergrad_t	Num		Number of undergraduate students
	undergrad	Num	undergrad_t/enroll	Proportion of undergraduate students
	grad_t	Num		Number of graduate students
	grad	Num	grad_t/enroll	Proportion of graduate students
	amin_t	Num		Number of American-Indian students
	amin	Num	amin_t/enroll	Proportion of American-Indian students
	asian_t	Num		Number of Asian students
	asian	Num	asian_t/enroll	Proportion of Asian students
	black_t	Num		Number of Black students
	black	Num	black_t/enroll	Proportion of Black students
	hisp_t	Num		Number of Hispanic students
	hisp	Num	hisp_t/enroll	Proportion of Hispanic students
	white_t	Num		Number of White students
	white	Num	white_t/enroll	Proportion of White students
	multi_t	Num		Number of Multi-race students
	multi	Num	multi_t/enroll	Proportion of Multi-race students
	unkn_t	Num		Number of unknown race students
	unkn	Num	unkn_t/enroll	Proportion of unknown race students
	nonres_t	Num		Number of Non-resident students
	nonres	Num	nonres_t/enroll	Proportion of Non-resident students

FBI - II	c_city	Char		City an institution belongs to
	c_population	Num		Population of the city
	c_crime_total	Num	c_violent+ c_property	Total crimes in the city
	c_violent	Num	c_v1+c_v2+c_v3 +c_v4	Total of violent crimes in the city
	c_v1	Num		Murder and nonnegligent manslaughter
	c_v2	Num		Forcible rape
	c_v3	Num		Robbery
	c_v4	Num		Aggravated assault
	c_property	Num	c_p1+c_p2+c_p3 +c_p4	Total of property crimes in the city
	c_p1	Num		Burglary
	c_p2	Num		Larceny-theft
	c_p3	Num		Motor vehicle theft
	c_p4	Num		Arson
	c_rcrime	Num	c_crime_total/enroll *1000	Total crimes per 1000 persons
	c_rviolent	Num	c_violent/enroll *1000	Violent crimes per 1000 persons
	c_rproperty	Num	c_property/enroll *1000	Property crimes per 1000 persons

For each crime type, its corresponding crime rate is calculated as:

$$\boldsymbol{crime\ rate} = \frac{\boldsymbol{number\ of\ crimes}}{\boldsymbol{enrollment}} \times \mathbf{1000}.$$

This crime rate represents the number of crimes per 1000 persons, which is a standard form adopted in the literature.

In addition to the variables from the three sources, longitude and latitude are manually created variables that we needed to specify the spatial information for each institution and to perform spatio-temporal modeling. Longitude and latitude are the coordinates for each institution, and they were collected from the Google Maps. As

spatial information was obtained through longitude and latitude, the variable "year" enables us to construct a time series. Therefore, the data contain both spatial and temporal information.

2.3 Data Compilation

The data set for the subsequent spatio-temporal modeling was constructed from the three sources, the FBI, NCES, and BLS, as mentioned previously. However, there are some challenges to merge the raw data due to their different formats. One of the primary principles is to create the data set of a particular structure that is adaptive to the computational package. Therefore, each variable obtained from a source was transformed and combined in an appropriate form. Another crucial issue to consider is the necessity of the attentiveness during the combination of the variables from different sources because each observation needs to be matched with the corresponding value from another source.

Although the three sources cover all fifty states and Washington, D.C., we chose only to study California and Texas for the following two reasons: First, they are culturally and socially representative but mutually different. Second, each of them has a large number of universities and colleges compared to the other states. For instance, California has around 300 and Texas has around 400 universities and colleges.

Since the data set was created from three different sources, we were only able to include the common universities in all sources. Universities, which appear in only one or two sources, were not eligible to be included. This led to a dramatic decrease in the number of institutions in our data set. In addition, some universities such as Stanford

University or California Institute of Technology decided not to voluntarily share their data with the public, so it was impossible to include such universities in our data set. It is obvious that private institutions usually prefer not to share their crime statistics. Also, most of the universities, which voluntarily share their data, are the 4-year-insitutions.

Similarly, the time range was decided due to its availability in all sources. We wanted our time series not to be very short. The data set from the NCES is until 2012, which determined the upper limit of the years. The lower limit was determined as 1997 to maximize the numbers of universities, which leads to the 16-year time series from 1997 to 2012 in our final data set.

We cannot use all the variables for model fitting because of a few conceptual principles. For instance, if a variable is obtained through a linear combination of several others, it is not valid to use such transformed variable together with its components in the model fitting procedure because they are linearly dependent. Moreover, to avoid multicollinearity, we may not include highly correlated predictors together in the model. The scatter plot matrix in Figure 2-5 shows that year, cpi_scalar, hepi_scalar and heca_scalar variables have strong linear relationships. Therefore, only one of them, cpi_scalar is chosen for the model fitting purpose. Also, all predictors which contain any missing values are not eligible to be in the model because the computational package fails to incorporate this scenario.

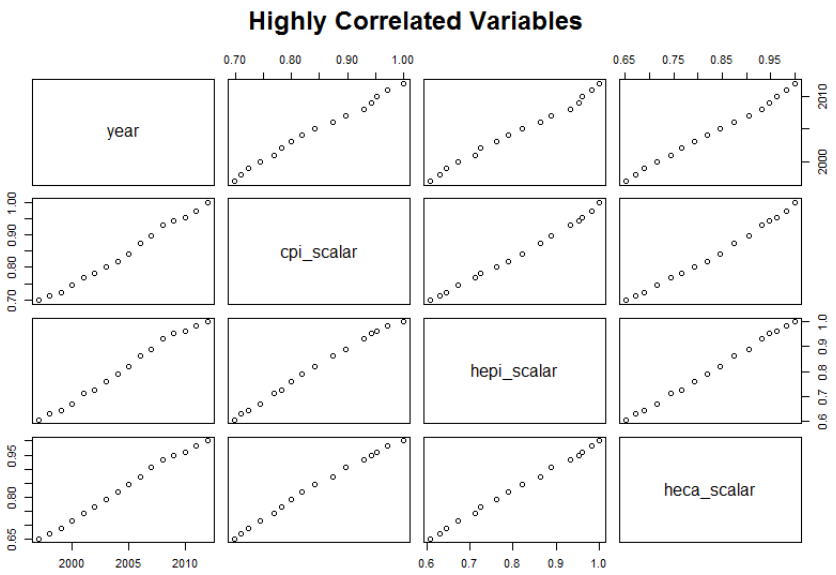

Figure 2-5. Several highly correlated variables in the data set, including year, cpi_scalar, hepi_scalar, and heca_scalar.

The full list of variables for later analysis is given in Table 2-2. All variables but one are the same to study both California and Texas. The variable “hsi” was included for California but not for Texas while “control” was used for Texas but not for California. The reason for using “hsi” but not “control” in California is that there are a considerable number of institutions that are Hispanic-serving, but there are few private institutions in the complied data for California. On the contrary, there are very few Hispanic-serving institutions in the final dataset in Texas. Moreover, we can use the variable “control” for Texas because there is a balanced distribution of publicly and privately controlled schools.

Table 2-2 The list of the variables used in the subsequent analysis.

Variable Number	Variable Name
1	unemp_rate
2	hsi (CA) / control (TX)
3	cpi_scalar
4	tuition
5	gom
6	undergrad
7	amin
8	asian
9	black
10	hisp
11	white
12	nonres
13	c_rcrime

Chapter 3

EXPLORATORY ANALYSIS

This chapter provides selective exploratory analysis results, including some summary statistics and visualizations of the data. Especially, the characteristics of the response variable, “crime rate”, will be visualized by bubble graphs and line charts for both California and Texas, which can demonstrate its evolvement over time and space.

3.1 Summary Statistics

Table 3-1 provides brief information on the data sets for California and Texas. As seen in Table 3-1, we have a 16-year time series of 32 institutions in California and 39 in Texas.

Table 3-1: Brief information on the data sets for California and Texas.

	California	**Texas**
Year Range	1997 - 2012	1997 - 2012
Numbers of Universities	32	39
Numbers of Predictors	12 + hsi	12 + control

Table 3-2 gives a few summary statistics of the college crime rates for the two states. As defined in Chapter 2, the crime rate here represents the number of crimes per 1000 persons. For each state, three types of crime rates are provided, including total crime, violent crime and property crime. Obviously, the property crime accounts for a

majority of the total crime, and California has a higher occurrence of all types of crimes per 1000 persons than Texas.

Table 3-2. Summary statistics of three types of college crime rates for California and Texas.

	California			**Texas**		
	Total Crime	Violent Crime	Property Crime	Total Crime	Violent Crime	Property Crime
Min	2.574	0	2.248	0.4898	0	0.490
Median	14.702	0.328	14.228	8.6602	0.140	8.464
Mean	21.232	0.722	20.510	12.7666	0.285	12.482
Max	185.185	20.701	182.280	77.7646	3.158	76.575
SD	23.865	1.533	23.183	11.856	0.443	11.666

3.2 Visualization

Spatial and temporal patterns are essential features of our data, which can be visualized by bubble plots and line charts. They will enable us to understand the change of crime rates over locations and time. Below we provide the bubble plots and line charts for total crime rate, violent crime rate, and property crime rate respectively for both states.

3.2.1 Bubble Plots for California

Bubble graphs are used to visualize both spatial and temporal patterns. Figures 3-1 (a) (b) refer to the yearly bubble plots of the total crime rates for the 32 institutions in California from 1997 to 2012. The position of a bubble signifies the location of its corresponding university, and a bigger and darker bubble represents a higher crime rate. Figures 3-1 (a) (b) show that overall the total crime rate declines from the late 1990s to

2012, but its variability across institutions is high. Among the 32 institutions, Santa Rosa Junior College is one of the safest universities, while the University of California, San Francisco, and California State University, Monterey Bay have the highest total crime rates, which makes them the most unsafe universities in California. The bubble plots were created by using the R program packages ggmap and ggplot2 (Kahle and Wickham, 2013).

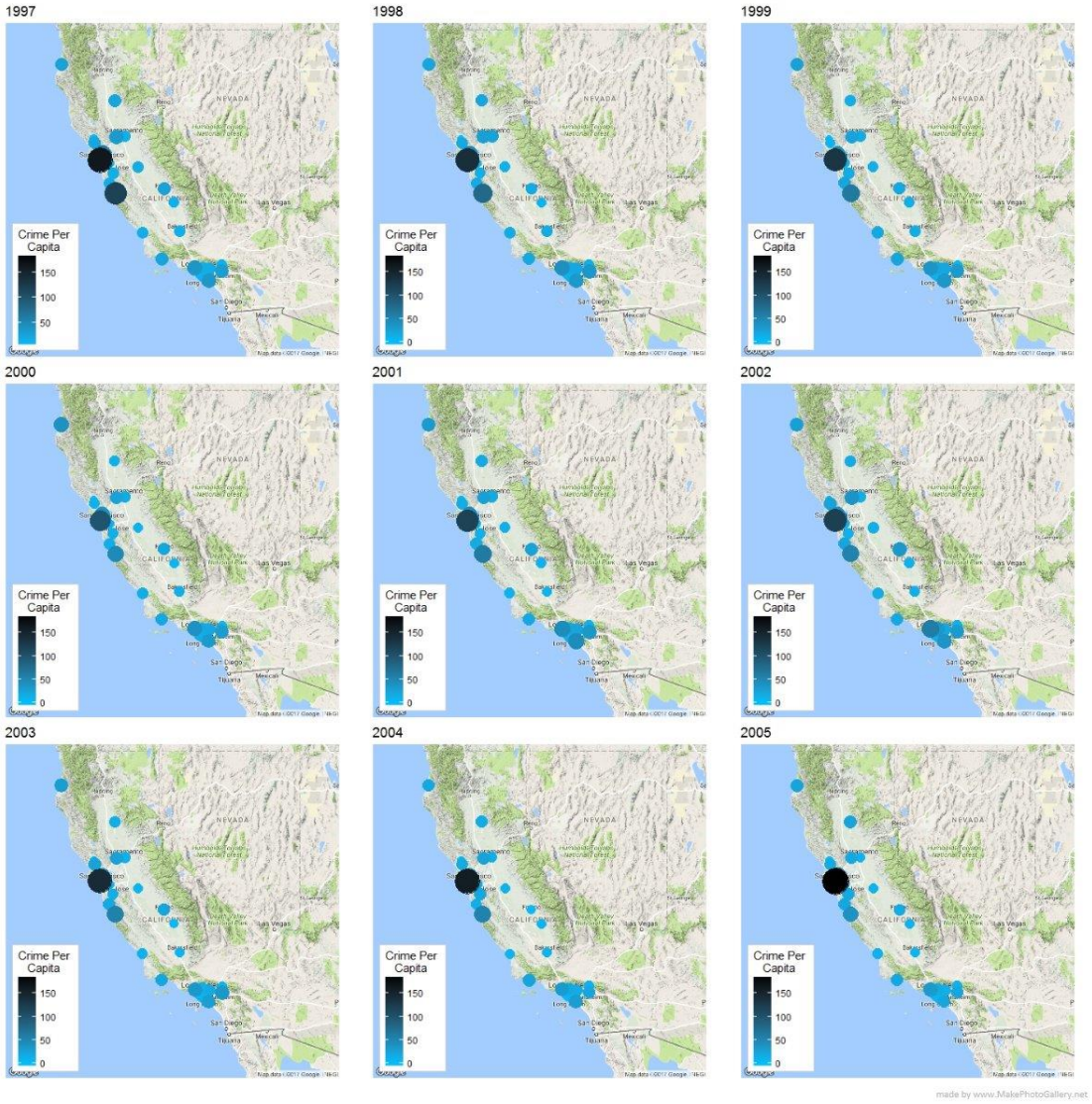

Figure 3-1 (a). Yearly bubble plots of the total crime rates for the 32 institutions in California from 1997 to 2005.

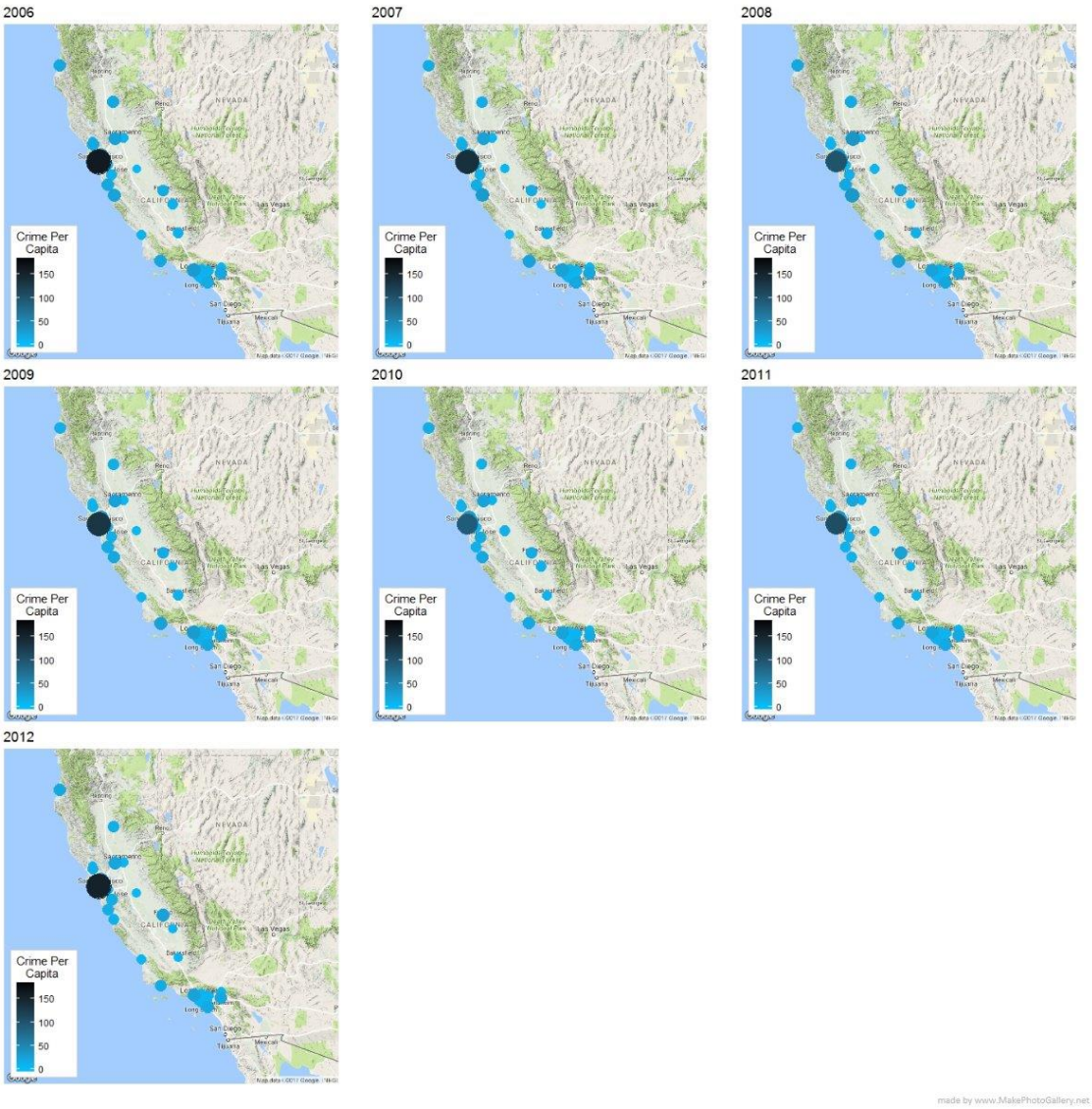

Figure 3-1 (b). Yearly bubble plots of the total crime rates for the 32 institutions in California from 2006 to 2012.

Similarly, the bubble plots of the two sub-categories of the crime rate, violent crime rate and property crime rate, are given in Figures 3-2 (a) (b) and Figures 3-3 (a) (b) respectively. Recall that the violent crime takes a small proportion of the total crime. Many institutions even have zero violent crimes for some years. Therefore, most

universities are safe in terms of violent crime. Among these institutions, the University of California, Hastings College of Law, and California State University, Monterey Bay have higher violent crime rates than the others. In addition, the violent crime rate decreases over years especially in the places where violent crimes were observed the most.

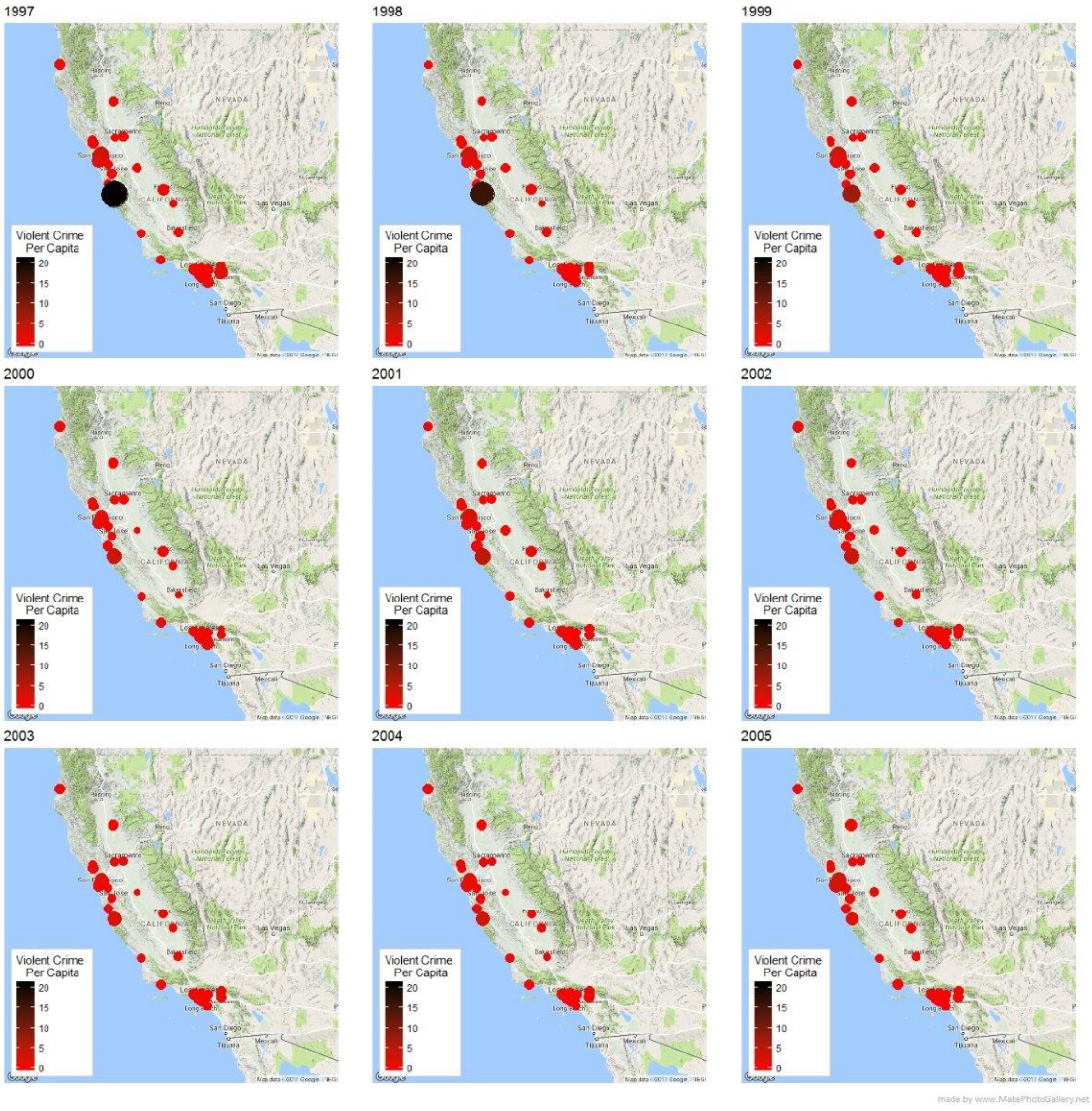

Figure 3-2 (a). Yearly bubble plots of the violent crime rates for the 32 institutions in California from 1997 to 2005.

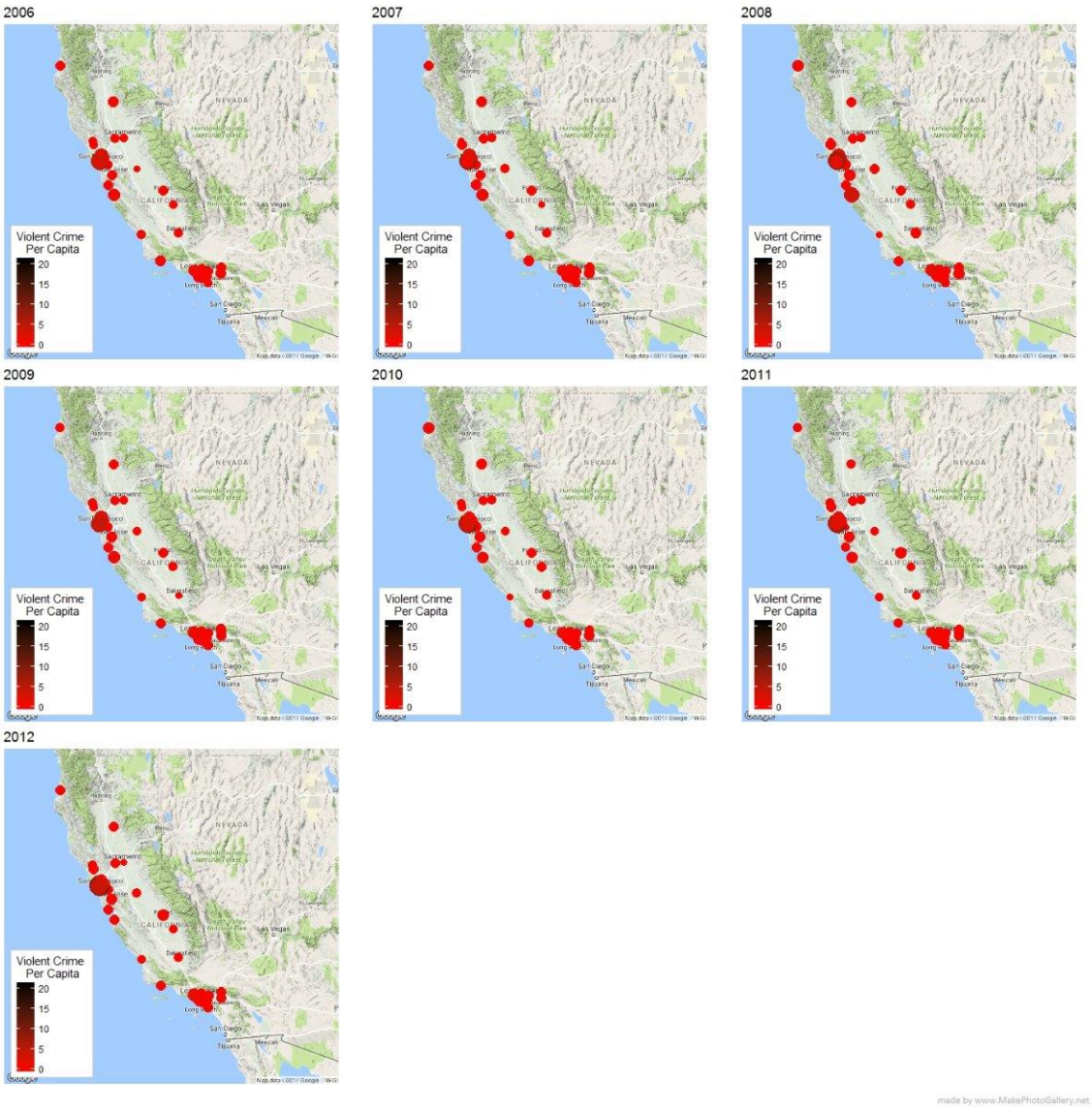

Figure 3-2 (b). Yearly bubble plots of the violent crime rates for the 32 institutions in California from 2006 to 2012.

We also showed here the yearly bubble plots of the property crime rates in Figure 3-3 (a) (b). Since the property crime rate is very close to the total crime rate, the discussion on the total crime also applies to the property crime. Figures 3-3 (a) (b)

indicate that the most property crimes per 1,000 persons were committed in the University of California, San Francisco, and California State University, Monterey Bay, while Santa Rosa Junior College has the smallest property crime rate. Also, we again observe the shrinking sizes of the bubbles over years, which is the sign of a decline in property crime rate over time.

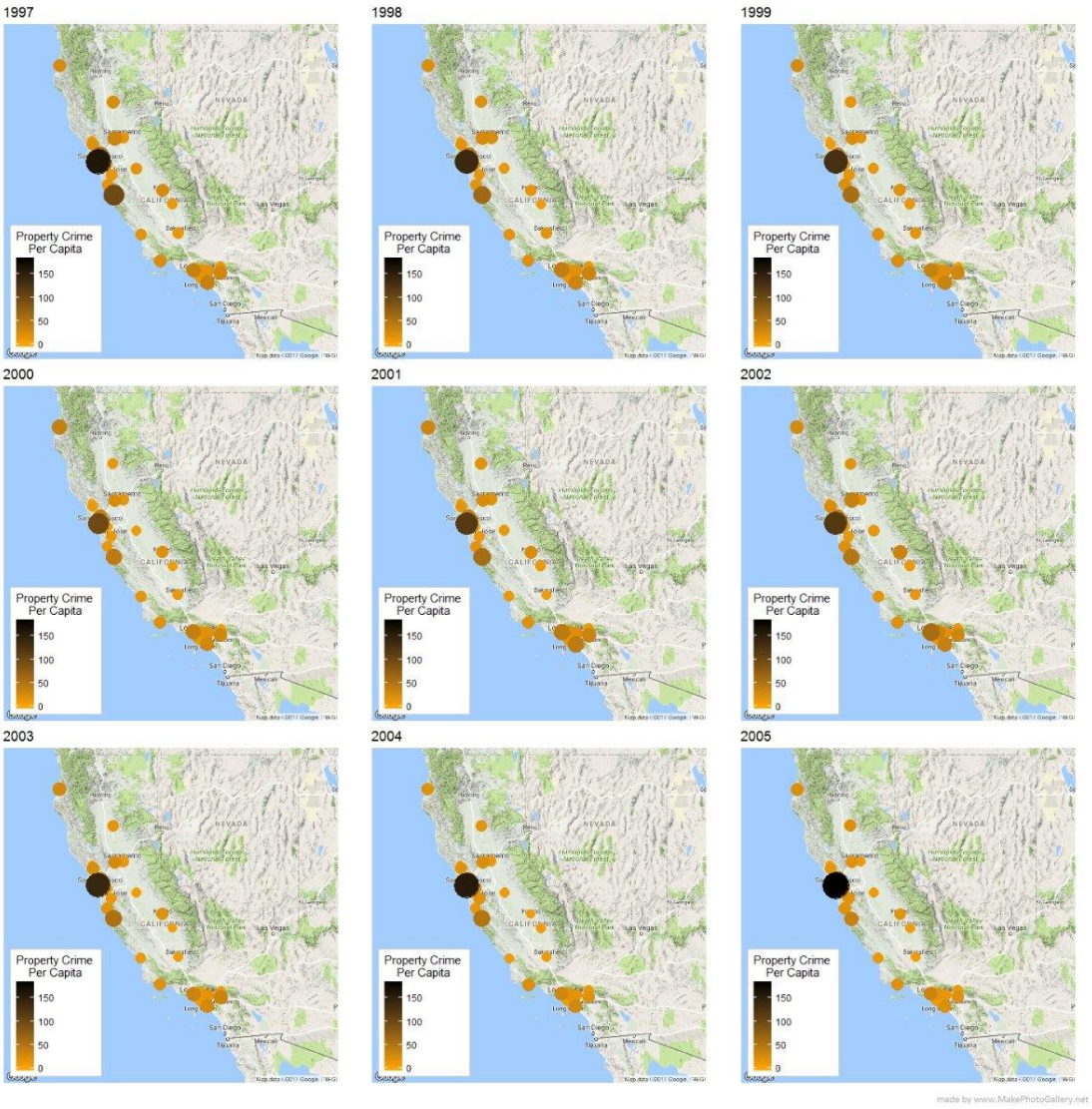

Figure 3-3 (a). Yearly bubble plots of the property crime rates for the 32 institutions in California from 1997 to 2005.

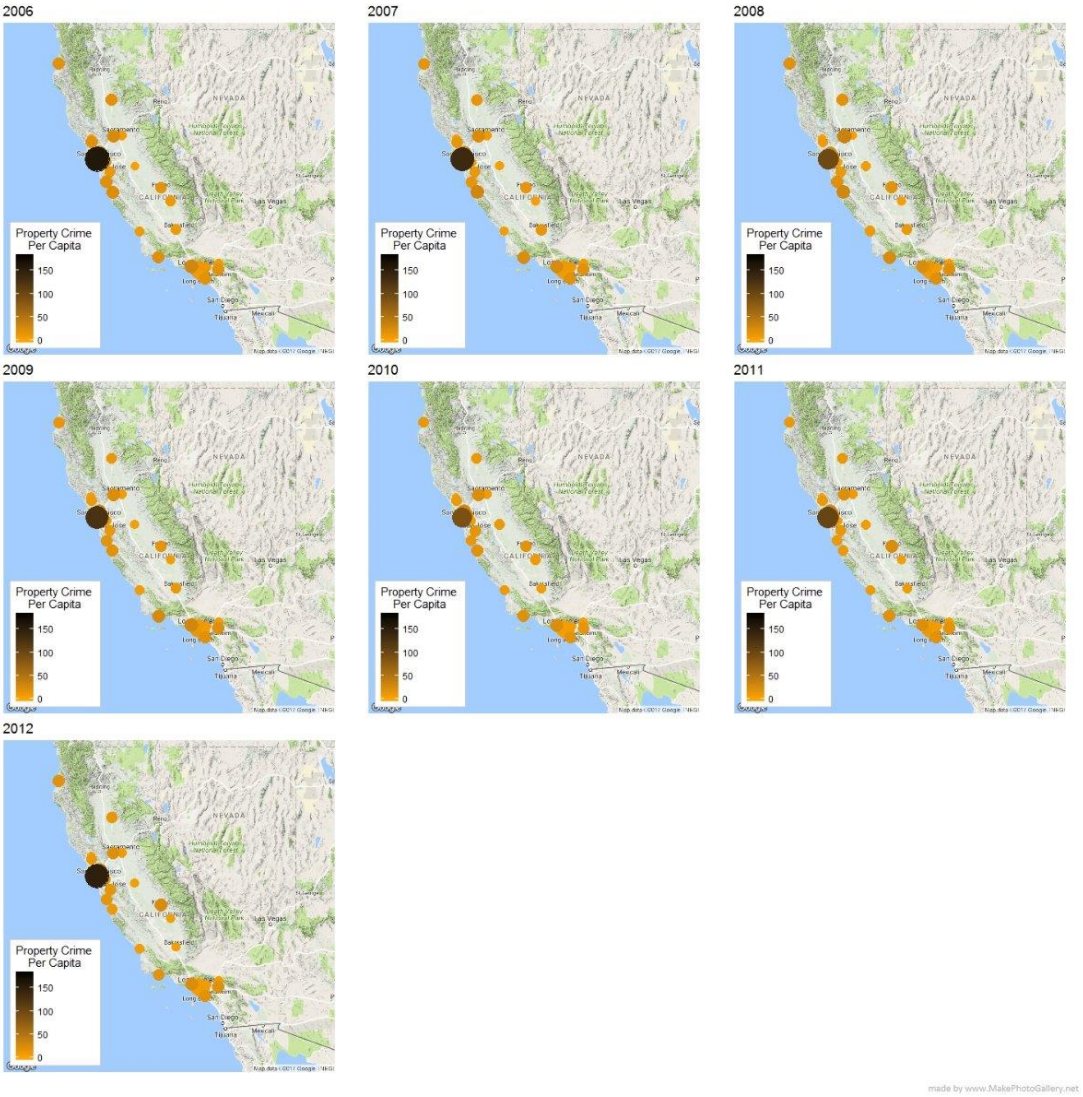

Figure 3-3 (b) Yearly bubble plots of the property crime rates for the 32 institutions in California from 2006 to 2012.

Lastly, we aggregated the crimes of each institution over years and created three bubble plots in Figure 3-4 to visualize the space-only patterns of the three types of crime rates. The total crime rate and property crime rate were achieved at their highest level in the University of California, San Francisco. Most of the universities are safe in terms of

violent crime, and California State University, Monterey Bay, and the University of California, Hastings College of Law have relatively high violent crime rates. Santa Rosa Junior College is the notable college as the safest place in terms of both total and property crime. Consequently, we ended up with the same conclusions from the aggregated bubble plots as we made from the yearly bubble graphs.

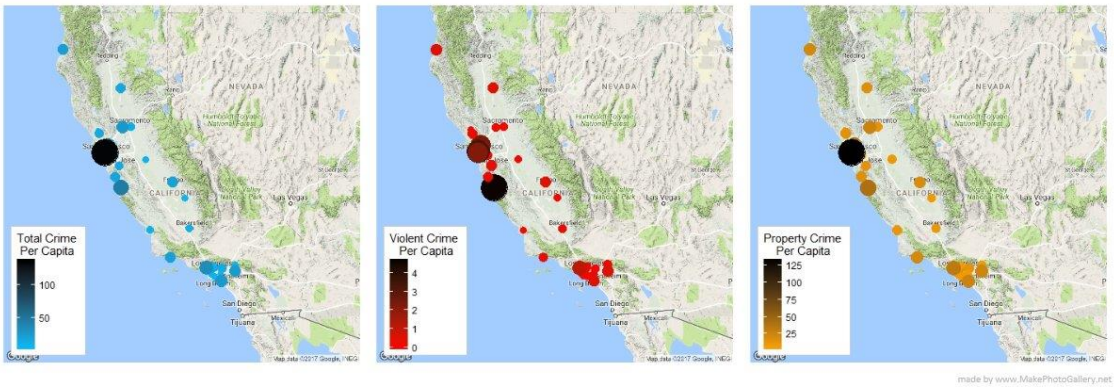

Figure 3-4. Yearly aggregated plots of the three types of crime rates for the 32 institutions in California.

3.2.2 Line Charts for California

Line charts were used here to present time-only patterns more effectively. For each year, we averaged the crime rates over the 32 institutions, and created a line chart for each crime type over these years. Figures 3-5, 3-6 and 3-7 respectively show the change of the rate of the total crime, violent crime and property crime over time. As observed in the bubble plots, the total crime rate and property crime rate both decrease from 1997 to 2012. The temporal pattern of the violent crime rate is different, which is

somewhat unsurprising. Even though the violent crime rate in 2012 shows a decline from 1997, it is overall very stable over the years.

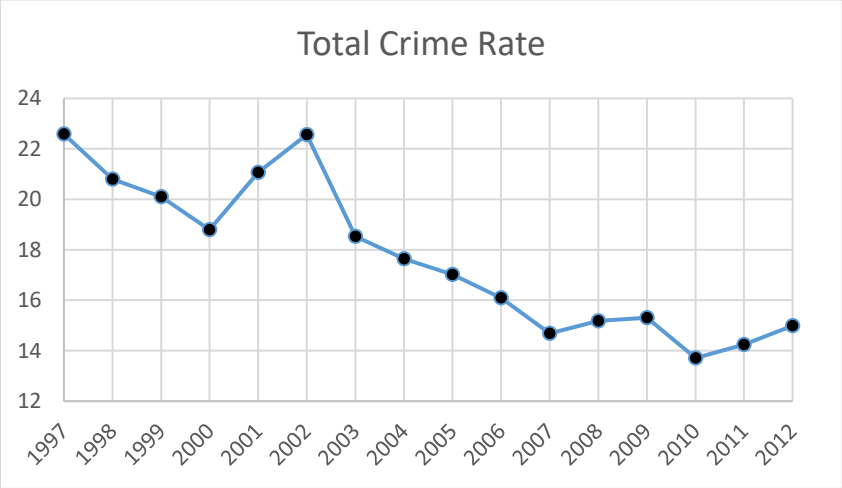

Figure 3-5. The total crime rate over time in California.

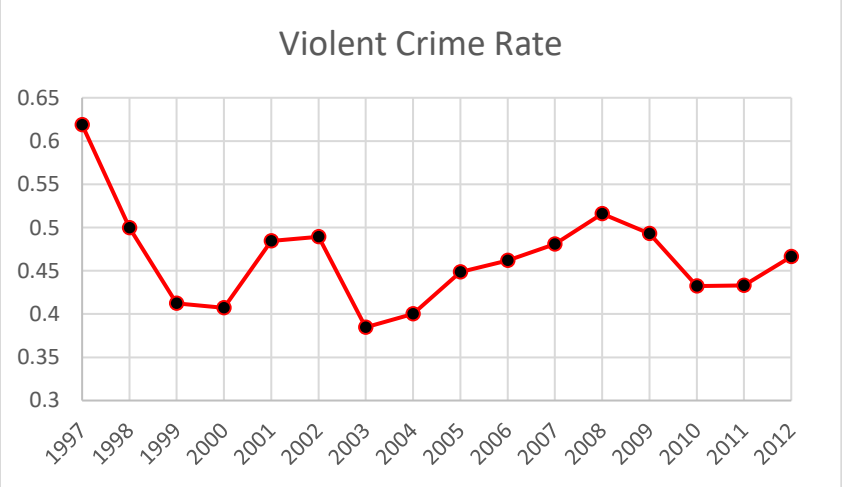

Figure 3-6. The violent crime rate over time in California.

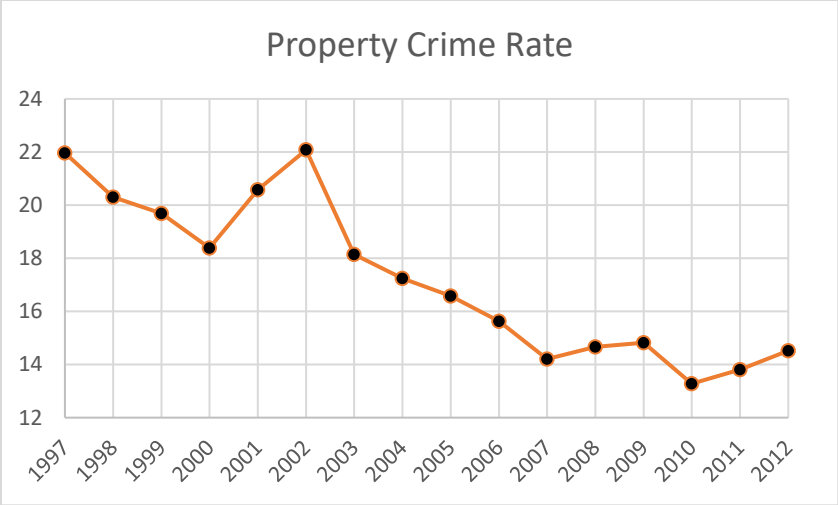

Figure 3-7. The property crime rate over time in California.

3.2.3 Bubble Plots for Texas

Bubble plots and line charts for Texas were created following the same procedure as for California. The initial information that we get from the yearly bubble graphs is the decrease in the total crime rate over time. Even though the universities in Texas do not show as high crime rates as in California, several universities such as Rice University, Trinity University and University of North Texas Health Science Center are the most dangerous universities in Texas while Central Texas College and South Plains Colleges are the safest places, in terms of both property and total crime rates. Similar to California, there are many universities in Texas where no violent crime was ever committed, which makes most of the institutions very safe in terms of violent crime, but Texas Southern University and Rice University are on the top of the list of the institutions with relatively high violent crime rates. More importantly, Rice University gets the attention on having the highest total crime rate.

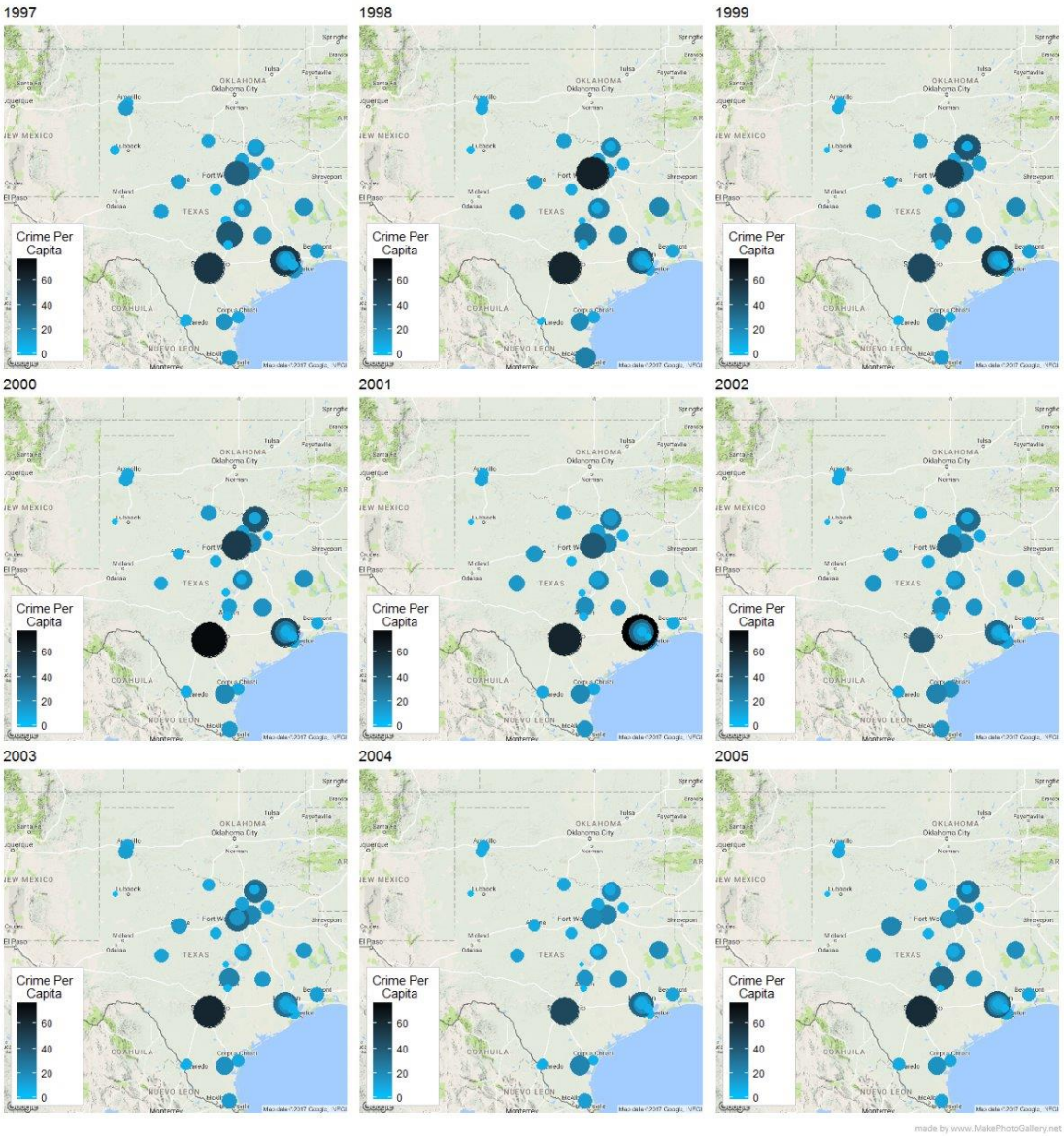

Figure 3-8 (a). Yearly bubble plots of the total crime rates for the 39 institutions in Texas from 1997 to 2005.

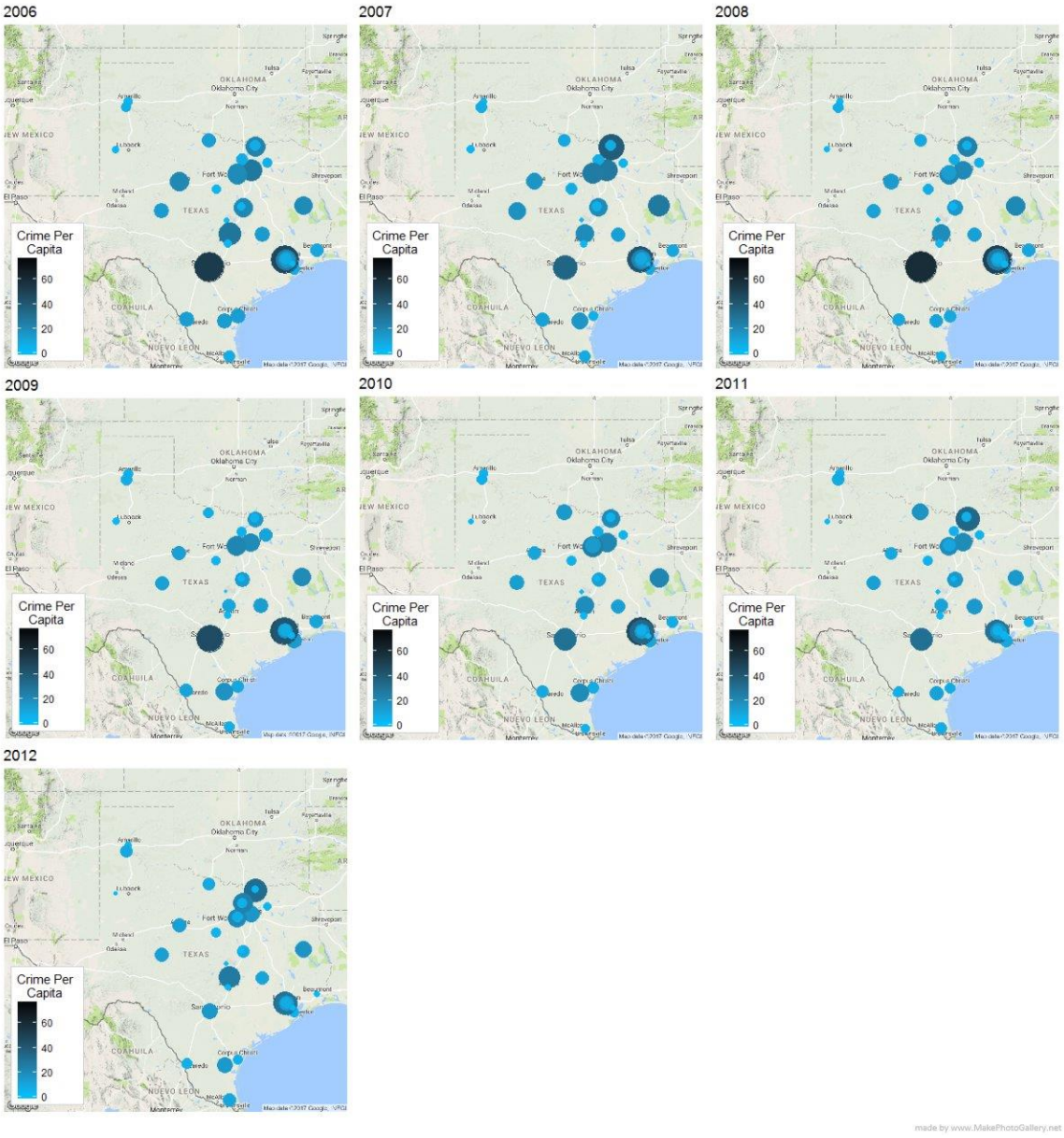

Figure 3-8 (b). Yearly bubble plots of the total crime rates for the 39 institutions in Texas from 2006 to 2012.

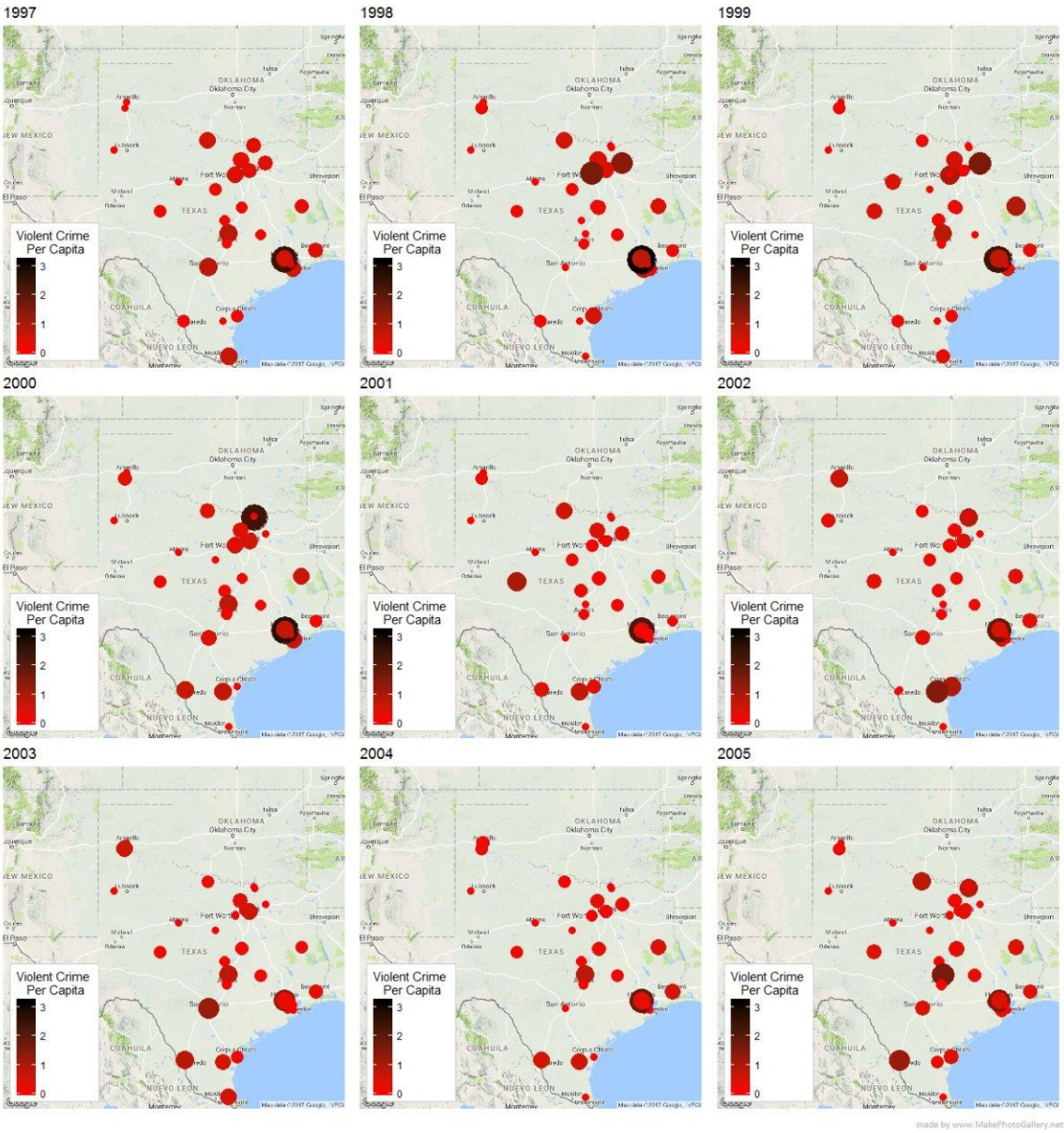

Figure 3-9 (a). Yearly bubble plots of the violent crime rates for the 39 institutions in Texas from 1997 to 2005.

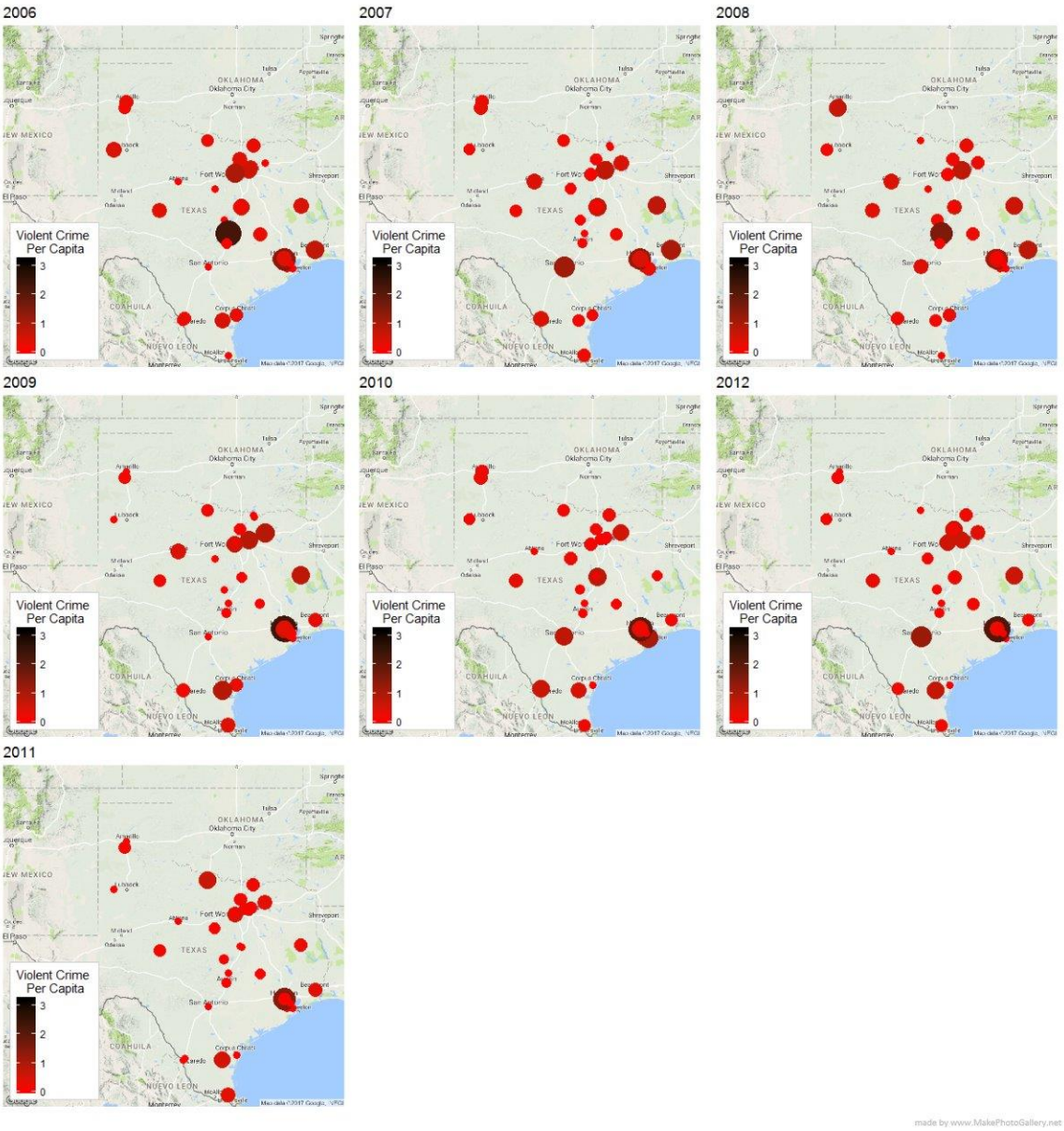

Figure 3-9 (b). Yearly bubble plots of the violent crime rates for the 39 institutions in Texas from 2006 to 2012.

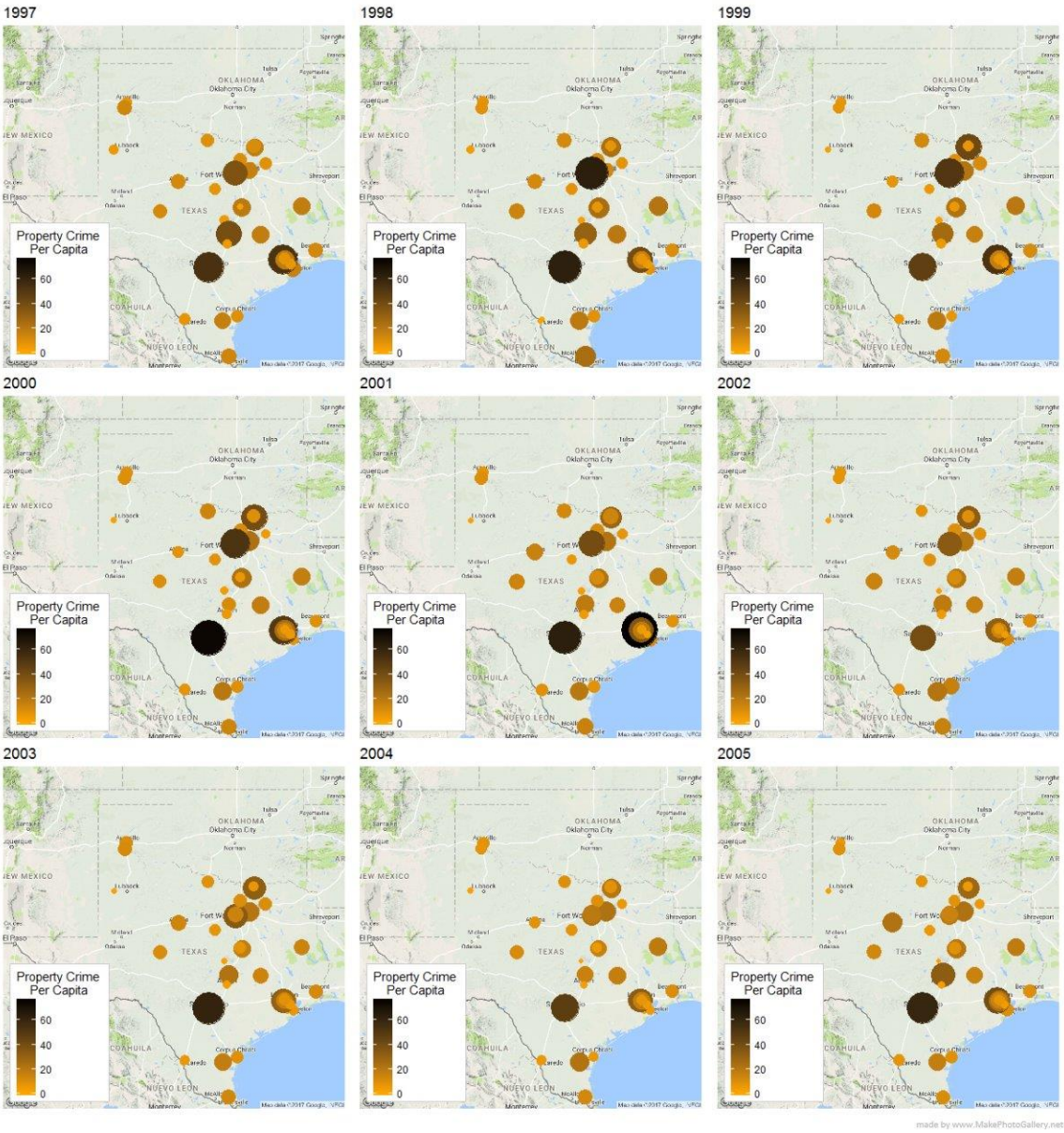

Figure 3-10 (a). Yearly bubble plots of the property crime rates for the 39 institutions in Texas from 1997 to 2005.

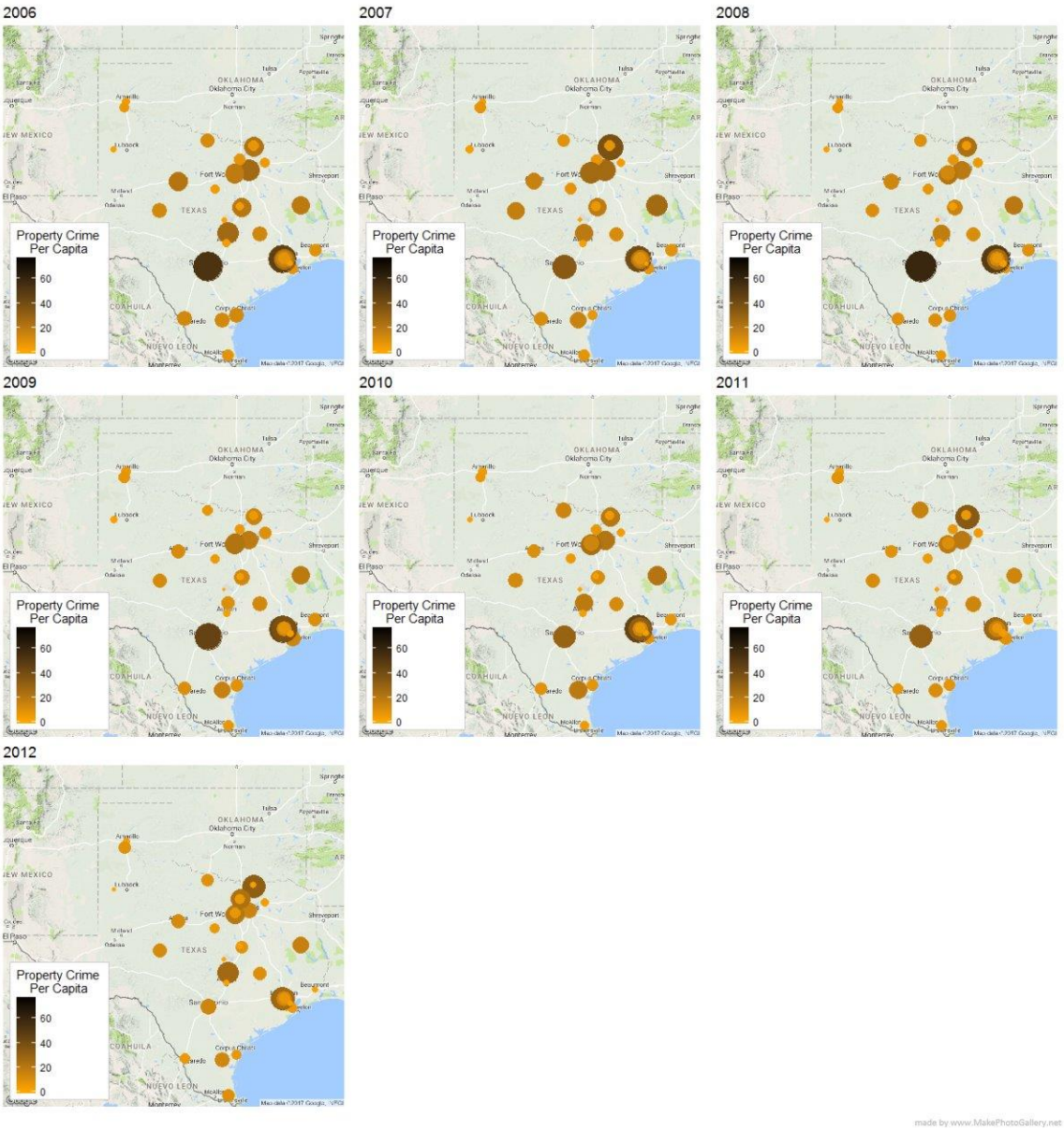

Figure 3-10 (b). Yearly bubble plots of the property crime rates for the 39 institutions in Texas from 2006 to 2012.

Yearly aggregated plots, which only reveal spatial information, allow us to justify our comments with yearly bubble plots for Texas in Figure 3-11. Same universities showed similar characteristics even though the time dimension was eliminated.

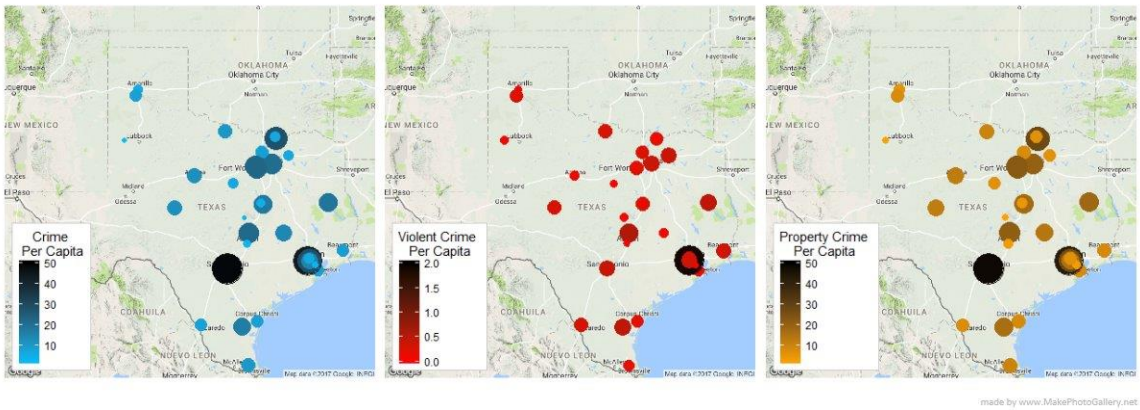

Figure 3-11. Yearly aggregated plots of the three types of crime rates for the 39 institutions in Texas.

3.2.4 Line Charts for Texas

Line charts indicate that there is a dramatic decrease for all crime rates over time in Texas. A majority of the total crime is the property crime in Texas, so the patterns of the two crime types are very similar. The violent crime rate in each year is very small, and it declines over the years, although the trend is not monotone. However, the reason for its rapid increase from 2011 to 2012 is unknown and thus requires additional investigations in the future.

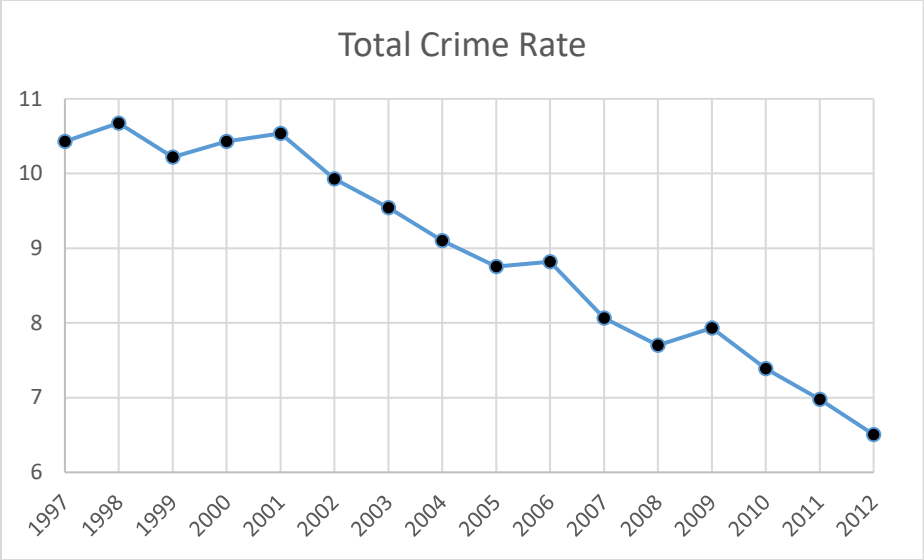

Figure 3-12. The total crime rate over time in Texas.

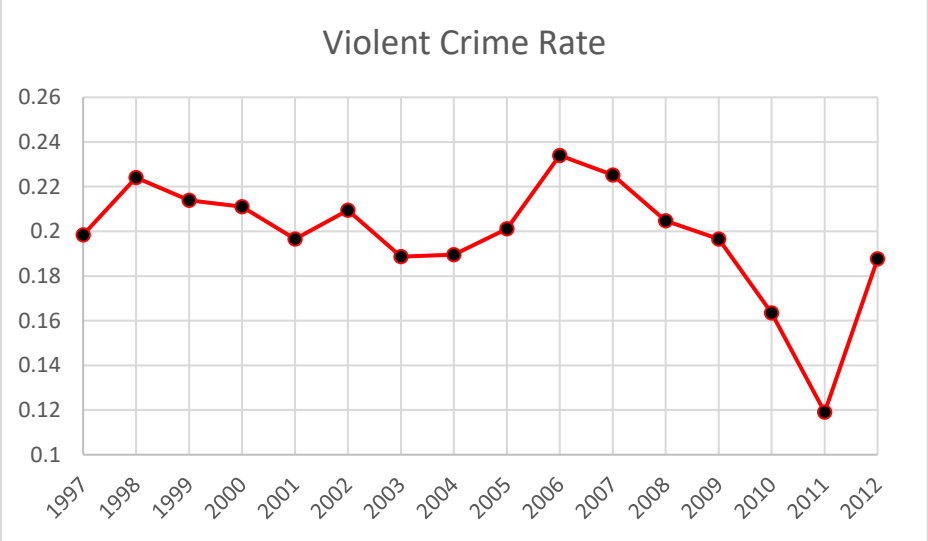

Figure 3-13. The violent crime rate over time in Texas.

Figure 3-14. The property crime rate over time in Texas.

From now on, we focus on the total crime rate as the response variable in the subsequent statistical analysis. The total crime rate reflects the overall safety level, but as seen in both California and Texas, it primarily reflects the likelihood of the property crime occurrence. Since the violent crime has too many zero values, it may be studied differently, e.g., by a generalized spatial-temporal model, but the computational software that can fit such model is currently unavailable. This is beyond the scope of this thesis, and is a promising future research topic.

3.3 The Relationship Between the Response and Predictors

To explore the association between the predictors and response, i.e., the total crime rate, we present scatterplots and correlation matrices for continuous predictors, and boxplots for categorical predictors.

3.3.1 California

As shown in Figures 3-15 and Table 3-3, the total crime rate has a moderate and positive correlation with the gross operating margin and tuition. It is negatively correlated with the proportion of the undergraduate students, which is the strongest relationship. Also, the proportion of Asian students is positively and that of Hispanic students is negatively related to the crime rate but their associations are not very strong. The predictor "hsi" is not contained in Table 3-3 since it is a categorical variable and the correlation coefficient is not applicable.

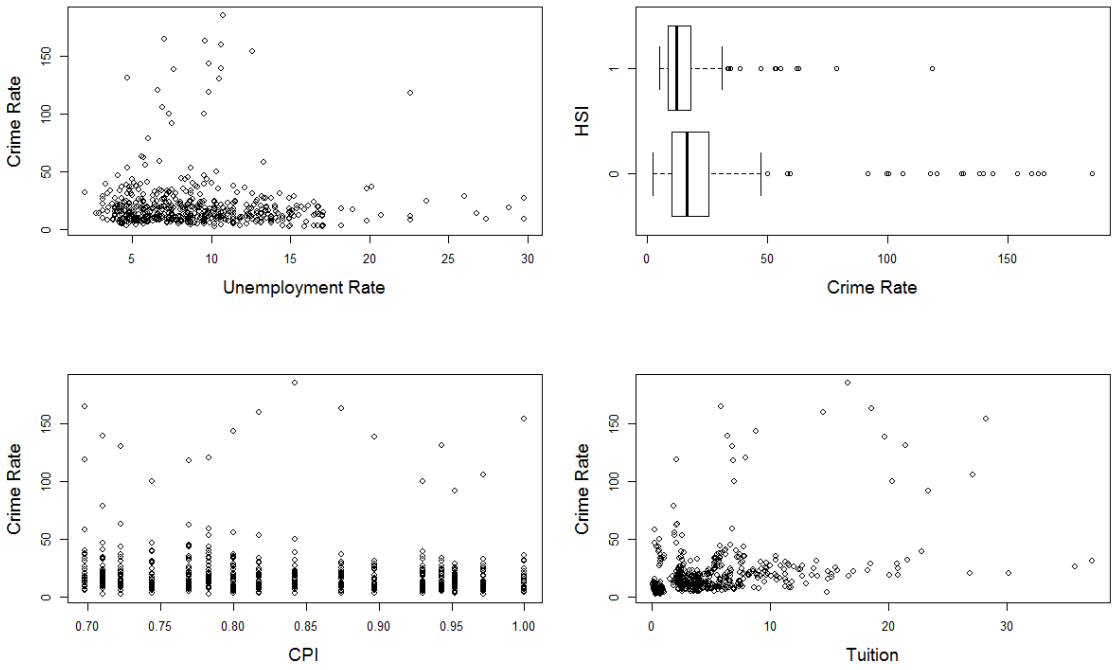

Figure 3-15 (a). California: Three scatterplots for three continuous predictors respectively: unemp_rate, cpi, and tuition, and two boxplots for the two levels of a categorical predictor: hsi.

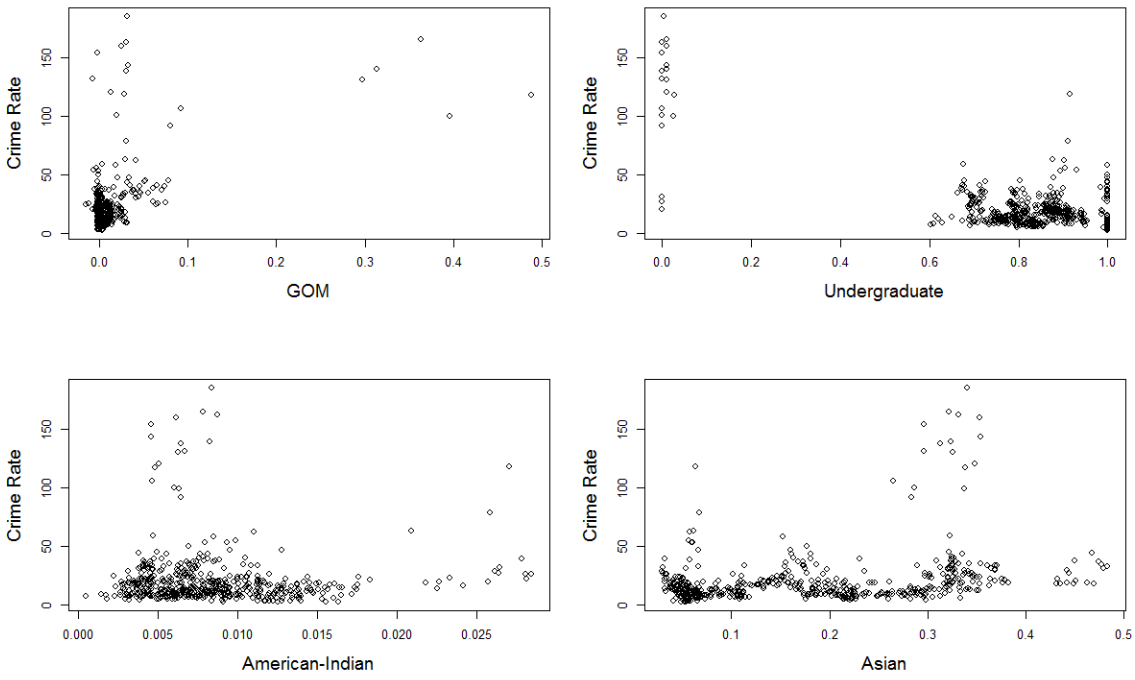

Figure 3-15 (b). California: Four scatterplots for four continuous predictors respectively: gom, undergrad, amin, and asian.

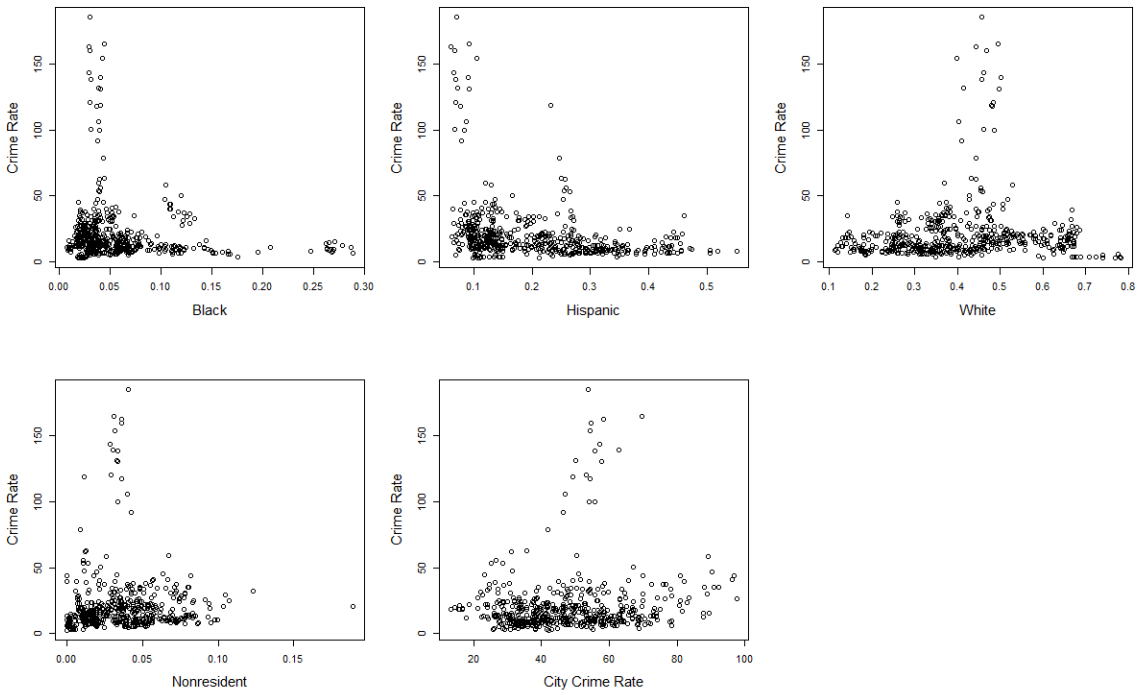

Figure 3-15 (c). California: Five scatterplots for five continuous predictors respectively: black, hisp, white, nonres, c_rcrime.

Table 3-3. Correlation coefficients between the response and all continuous predictors in California.

	rcrime
unemp_rate	-0.0248
cpi_scalar	-0.122
tuition	0.370
gom	0.535
undergrad	-0.691
amin	-0.004
asian	0.279
black	-0.132
hisp	-0.313
white	0.092
nonres	0.047
c_rcrime	0.133

3.3.2 Texas

Similar to California, we created Figure 3-16 (a)-(c) and Table 3-4 for Texas. In Texas, the management of an institution, whether it is publicly or privately controlled, seems to have a strong association with the crime rate. Tuition is positively and the proportion of the undergraduate students are negatively correlated with the crime rate, as discovered in California. However, checking the impacts of these variables allows us to summarize that, compared with California, the tuition has a higher correlation with the response but the proportion of undergraduate students has a lower correlation in Texas. Meanwhile, in both California and Texas, the gross operating margin, and the proportions of the Asian and Hispanic students all have a similarly low association with the crime rate. The variable “control” was not included in Table 3-4 because it is a categorical variable.

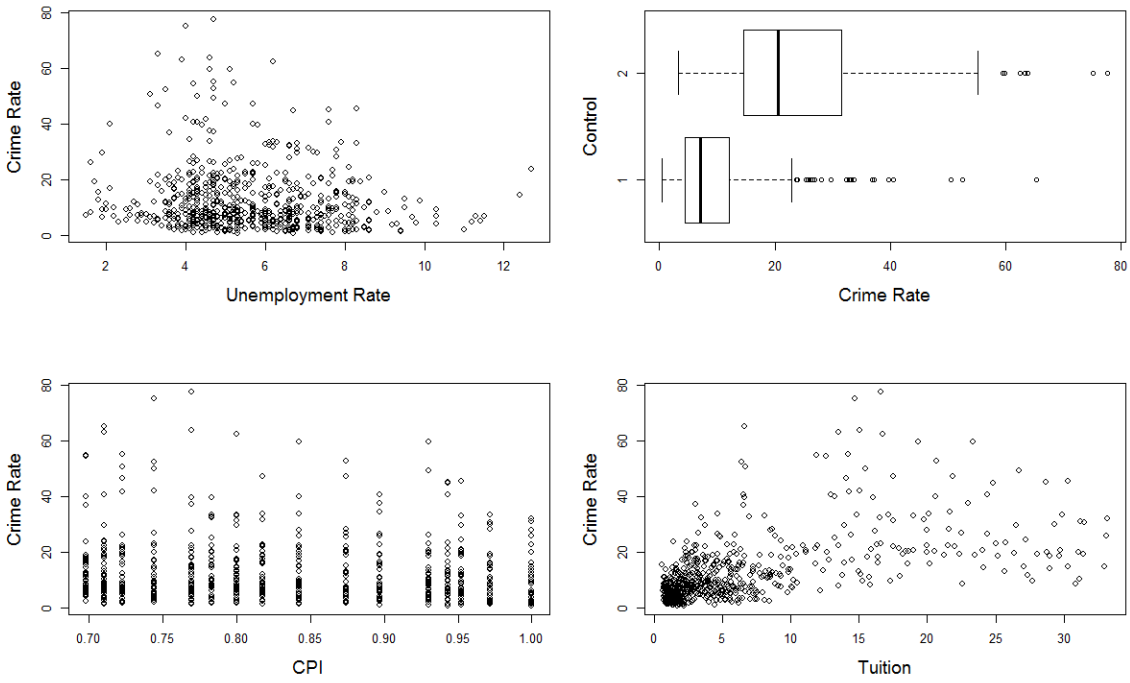

Figure 3-16 (a). Texas: Three scatterplots for three continuous predictors respectively: unemp_rate, cpi, and tuition, and two boxplots for the two levels of a categorical predictor: control.

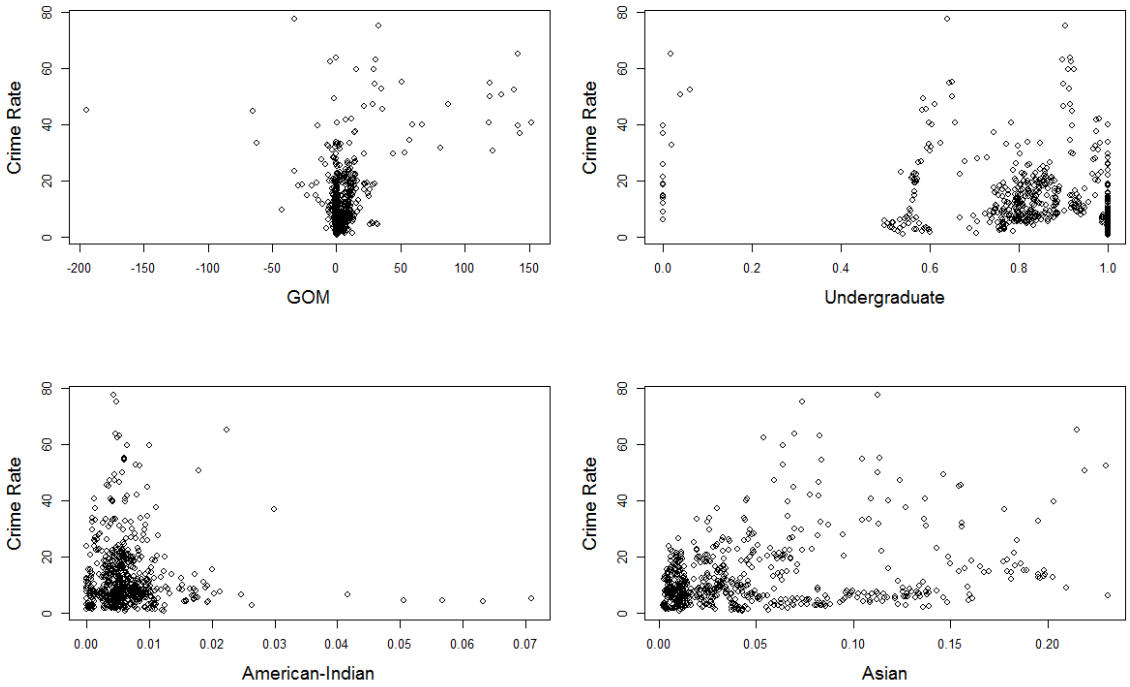

Figure 3-16 (b). Texas: Four scatterplots for four continuous predictors respectively: gom, undergrad, amin, and asian.

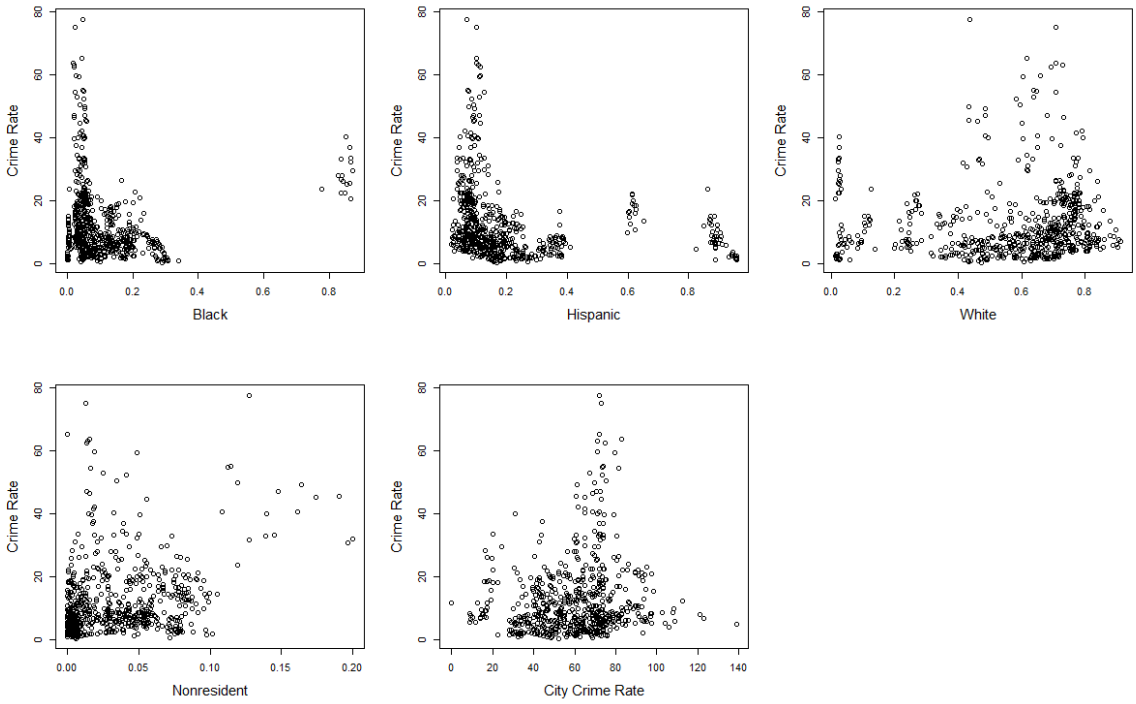

Figure 3-16 (c). Texas: Five scatterplots for five continuous predictors respectively: black, hisp, white, nonres, c_rcrime.

Table 3-4. Correlation coefficients between the response and all continuous predictors in Texas.

	rcrime
unemp_rate	-0.124
cpi_scalar	-0.134
tuition	0.566
gom	0.356
undergrad	-0.278
amin	-0.052
asian	0.284
black	0.034
hisp	-0.241
white	0.068
nonres	0.282
c_rcrime	0.126

Chapter 4

STATISTICAL ANALYSIS

In this chapter, we introduce the framework of the statistical analysis. We first specify the autoregressive model that is adaptive to the spatio-temporal structure of our data. Then we introduce the estimation procedure in a Bayesian paradigm, and Gibbs sampling, a Markov chain Monte Carlo (MCMC) method for computation. Moreover, due to a large number of predictors, model selection is needed for preferable interpretations, so the criterion for model assessment and the procedure of model selection will be presented comprehensively. In addition, to check whether spatial homogeneity, i.e., spatial stationarity, is valid, which is an essential assumption of the model, we will apply the global models on sub-regions to validate its performance. Global models' performance on smaller regions helps us to choose the best model for California and Texas respectively. Finally, we will briefly introduce the spTimer R package to succeed all essential computations.

4.1 Autoregressive Model

We fitted an autoregressive (AR) model, which was developed by Sahu et al. (2007), for our spatio-temporal data. The AR model indicates that the current value of the response depends on both its previous value in time and the current values of the

predictors. Let $Z_l(s_i,t)$ denote the observed response at location s_i, $i=1,...,n$, and time $t=1,...,T_l$, $l=1,...,r$, where l is a long time unit and t is a short time unit. Denote the true value of $Z_l(s_i,t)$ by $O_l(s_i,t)$, with $Z_{lt}=(Z_l(s_1,t),...,Z_l(s_n,t))'$ and $O_{lt}=(O_l(s_1,t),...,O_l(s_n,t))'$. Suppose that X_{lt} is a $n\times p$ matrix with p predictors including the intercept. The regression coefficients are in a $p\times 1$ vector form as $\beta=(\beta_0,...,\beta_p)'$ where p is the number of the predictors in the model. The autoregressive parameter is denoted by ρ.

The AR model takes the following form:

$$Z_{lt}=O_{lt}+\varepsilon_{lt},$$
$$O_{lt}=\rho O_{lt-1}+X_{lt}\beta+\eta_{lt}.$$

Here the error term or the nugget effect is denoted as $\varepsilon_{lt}=(\varepsilon_l(s_1,t),...,\varepsilon_l(s_n,t))'$ which assumed to be independently and normally distributed as $N(0,\sigma_\varepsilon^2 I_n)$, where σ_ε^2 is the unknown error variance and I_n is the identity matrix of the n locations. Similarly with the point-referenced observed data and the true values, the spatio-temporal random effects are $\eta_{lt}=(\eta_l(s_1,t),...,\eta_l(s_n,t))'$ and they are assumed to follow a normal distribution as $N(0,\Sigma_\eta=\sigma_\eta^2 S_\eta)$, where σ_η^2 is the invariant spatial variance for locations and S_η is the spatial correlation matrix.

In our study $r=1$, and $t=1997,...,2012$. If a longer time unit than year, e.g, century, is available, the notation l will be used in the equation. For generality, we

always keep the subscript l in all formulas below. In this study, the values of p will be up to 13 because we have 13 predictors in total. The range of the autoregressive parameter ρ is from -1 to 1, and if $\rho = 0$, then no autoregressive effect exists, which implies that the current value of the response is independent of its history.

Each initial term O_{l0} has mean μ_l, and covariance matrix $\sigma_l^2 S_0$, where σ_l^2 is the variance of each time slot. The correlation matrix S_0 can be created through a spatial correlation function and the explanation will be deferred in Chapter 4.2.

4.2 Spatial Correlation

The spatial variance for the locations $\Sigma_\eta = \sigma_\eta^2 S_\eta$ as we mentioned in Chapter 4.1 contains the spatial correlation matrix S_η. We have the control for choosing the type of spatial correlation in the spTimer package. The performance of the Matérn and exponential correlations for the best models was compared.

The Matérn correlation (Matérn 1986; Handcock and Stein 1993; Handcock and Wallis 1994) takes the form:

$$S_\eta(s_i, s_j; \phi, \nu) = \frac{1}{2^{\nu-1}\Gamma(\nu)}(2\sqrt{\nu}\left\|s_i - s_j\right\|\phi)^\nu K_\nu(2\sqrt{\nu}\left\|s_i - s_j\right\|\phi), \qquad \phi > 0, \nu > 0 .$$

where $\Gamma(\nu)$ is the standard gamma function, K_ν is the second kind Bessel function with the order ν, and $\left\|s_i - s_j\right\|$ is the distance between two locations. The rate of decay for the

correlation is controlled by the parameter ϕ while the distance increases between two locations (Banerjee et al. 2004; Cressie 1993).

On the other hand, the exponential correlation (Sahu et al., 2007) takes the form:

$$S_\eta(s_i - s_j;\phi) = S_\eta(d_{ij},\phi) = \exp(-\phi d_{ij}).$$

where d_{ij} is the distance between two sites similarly with the Matérn correlation. Since the exponential correlation always led to a better performance in terms of the predictive model selection criteria values, which will be introduced in Chapter 4.5, than the Matérn correlation for the best models we selected, below we only report the results where we fit the spatial structure using the exponential correlation.

4.3 Bayesian Framework

Bayesian framework is predominantly used to fit spatio-temporal data. By Gelfand (2012), a hierarchical structure can be established to analyze Bayesian spatio-temporal models. Three hierarchies are specified below to represent the distributions of data, process, and parameters as:

First Stage: [Data | Process, Parameter];

Second Stage: [Process | Parameter];

Third Stage: [Parameter].

The Bayes' rule allows us to obtain the posterior distribution of the process and parameters given the data as follows (Cressie and Wikle, 2011; Gelfand et al., 2010):

[Process, Parameter | Data] α [Data | Process, Parameter]

× [Process | Parameter]

× [Parameter]

In our case, to represent all parameters, a generic notation θ will be used as $\theta = (\beta, \rho, \sigma_\varepsilon^2, \sigma_\eta^2, \phi, \nu, \mu_l, \sigma_l^2)$ where O contains all O_{lt}. All observed data are denoted as z and missing data are denoted as z*, although our data set does not contain any missing data. Following the Bayes' rule, the logarithm of the joint posterior distribution is as follows:

$$\begin{aligned}\log \pi(\theta, O, z^* \mid z) \quad \alpha \quad & -\frac{N}{2}\log\sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2}\sum_{r=1}\sum_{t=1}^{T_l}(Z_{lt}-O_{lt})'(Z_{lt}-O_{lt}) - \frac{\sum_{r=1}T_l}{2}\log\left|\sigma_\eta^2 S_\eta\right| \\ & -\frac{1}{2\sigma_\eta^2}\sum_{r=1}\sum_{t=1}^{T_l}(O_{lt}-\rho O_{lt-1}-X_{lt}\beta)'S_\eta^{-1}(O_{lt}-\rho O_{lt-1}-X_{lt}\beta) \\ & -\frac{1}{2}\sum_{r=1}\log\left|\sigma_l^2 S_0\right| - \frac{1}{2}\sum_{r=1}\frac{1}{\sigma_l^2}(O_{l0}-\mu_l)'S_0^{-1}(O_{l0}-\mu_l) + \log\pi(\theta).\end{aligned}$$

Appropriate prior distributions are assumed for the AR model above. Three specifications to describe are the mean, variance, and correlation. Besides the random effects, all the parameters describing the mean are assumed to be in an independent normal prior form. Their means and variances can be specified as (μ_β, μ_ρ) and $(\delta_\beta^2, \delta_\rho^2)$. Under the assumption of having a flat prior distribution, all means were set to be 0 and all

variances set to be 10^4. In addition, the AR model has an n-dimensional vector μ_l for each of its components which are also assumed to follow an independent normal prior distribution with mean 0 and variance 10^4 and each variance component follows a gamma distribution with mean a/b and variance a/b^2 where a and b are specified as $a=2$ and $b=1$ in our study to have a proper prior distribution (Gelman et al., 2004). The smoothing parameter ν and the decay parameter ϕ follow discrete uniform prior distributions, each in a proper range. In our study, all prior distributions will be specified as the given assumptions of the algorithm and they will be explained in detail in Chapter 4.8.

4.4 Gibbs Sampling

We used the Gibbs sampling (Gelfand and Smith, 1990), which is an MCMC algorithm, to fit the AR model computationally. The fundamental idea of the Gibbs sampling is about creating a Markov chain of samples, where each of the samples has a similarity with nearby samples. It is convenient to sequentially sample from a conditional distribution when a multivariate distribution is given but direct sampling is difficult.

To implement the algorithm, we first decide k numbers of the samples of $W=(w_1,...,w_n)$ from a joint distribution $p(w_1,...,w_n)$. We express the i-th sample as $W^{(i)}=(w_1^{(i)},...,w_n^{(i)})$. Then, we can start with an initial value $W^{(i)}$ and define the next sample as $W^{(i+1)}$ where $W^{(i+1)}=(w_1^{(i+1)},...,w_n^{(i+1)})$. Moreover, each component of this vector $W_j^{(i+1)}$ is sampled while its distribution is conditioned on all other recently sampled

components. The emphasis is on conditioning on the components of the second sample $W^{(i+1)}$ until $w_{j-1}^{(i+1)}$. Later, conditioning on the components of the first sample $W^{(i)}$ starting from $w_{j+1}^{(i)}$ up to $w_n^{(i)}$. From the first component, all components are sampled sequentially. Therefore, we end up with the distribution as $p(w_j^{(i+1)} \mid w_1^{(i+1)},...,w_{j-1}^{(i+1)}, w_{j+1}^{(i)},...,w_n^{(i)})$ and repeat it k times.

In the context of the AR model, all parameters other than ν and ϕ are assumed by prior distributions are sampled from the full conditional distributions. Since ν and ϕ have a non-standard conditional distribution, they can be sampled if they are assumed to follow a discrete uniform prior distribution. Another option, which is only applicable for ϕ, is assumed to follow a continuous uniform prior distribution with an interval or with a gamma prior distribution. Then, samples can be produced using the Metropolis-Hastings algorithm. We will permit the spTimer package to create the samples and give us the approximation of the parameters.

4.5 Model Assessment

The spTimer package allows us not only to fit a model but also to evaluate the quality of the fitted model by reporting the predictive model selection criteria values (PMCC) (Gelfand & Ghosh, 1998):

$$PMCC = \sum_{i=1}^{n}\sum_{r=1}\sum_{t=1}^{T_l}\left\{E(Z_l(s_i,t)_{rep} - z_l(s_i,t))^2 + Var(Z_l(s_i,t)_{rep})\right\}.$$

The first term in the parenthesis reflects the goodness-of-fit, and the second is the penalty term for model complexity where $Z_l(s_i,t)_{rep}$ assesses the future replica of $z_l(s_i,t)$. The PMCC values can be automatically calculated in the package.

Since PMCC reflects the trade-off between goodness-of-fit and model complexity, we can use it to compare different models in the intermediate steps of model selection. Since different models selection procedures may lead to multiple models, we can also use their PMCC values to decide the best one, if all of them satisfy the assumptions of the AR model, e.g., spatial stationarity.

4.6 Model Selection

Due to a large number of predictors, we used forward selection and backward elimination methods to select predictors. These two methods are based on the idea of repeatedly adding or dropping a predictor. Forward selection is a sequential process, which starts from a model containing no predictor, then repeatedly adds a predictor such that the resulted model is the best among all models with one additional predictor until all predictors are included in the model. On the other hand, backward elimination allows the opposite operation that starts with the model with all predictors, and then repeatedly removes the predictor such that the resulted model is the best among all models with one predictor removed until no predictor is contained in the model. We performed each model selection method depending on the smallest goodness-of-fit or the smallest PMCC value

in the intermediate steps. The organization of the four model selection methods can be described as:

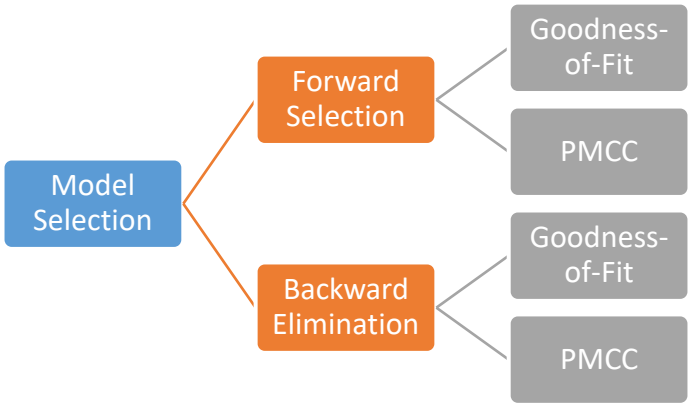

Figure 4-1. The diagram of forward and backward model selections with two selection criteria in the intermediate steps.

To clarify the procedure, we explain in detail how to perform forward selection with goodness-of-fit as the selection criterion in the intermediate steps as an example. We start with the null model, denoted by M_0, and add one of the thirteen predictors such that the goodness-of-fit value of the resulted model, denoted by M_1, is smallest among all models with only one predictor. Following the same principle, we obtain M_2, …, M_{13} by repeatedly adding more predictors until all predictors are included. Among M_0, …, M_{13}, we choose the candidate model which has the smallest corresponding PMCC value.

The diagram in Figure 4-1 indicates that we will end up with four candidate models. We then compare these models and investigate which one is the best for the data set. If one candidate model meets the assumptions of the AR model, e.g., spatial

stationarity, while another does not, then the former is chosen as the better one, regardless of their PMCC values. If multiple models satisfy the assumptions of the AR model, we will choose the best model that corresponds the smallest PMCC values.

4.7 Spatial Stationarity

The AR model assumes spatial stationarity to ensure that the model is applicable to the entire space. If spatial stationarity is violated, the results of the corresponding model are no longer trustworthy. Therefore, it is essential to check stationarity to guarantee the reliability of our results.

A popular method called Geographically Weighted Regression (GWR), which was first introduced by Fotheringham et al. (2002), motivates us how to check spatial stationarity in our study. The intuitive idea of GWR is simple: If spatial stationarity holds on the whole space, then the global model is supposed to describe local regions well too, so the parameter estimates obtained from the global model should be similar to those obtained from local models which are fitted on sub-regions (Bivand, 2015). This idea can successfully be applied to our study to detect whether spatial stationarity is violated. Once we obtained a global candidate model for each state as described in Chapter 4.6, we divided the state map into two sub-regions and applied the global model to them to see how the posterior distributions would change. Similar posterior densities from the entire map and two sub-regions imply spatial stationarity; otherwise, spatial stationarity is considered to be violated and separate model selections for each sub-region may be

needed. The criteria for choose sub-regions are different for California and Texas, and details will be given in Chapter 5.

4.8 Software and Example Code

The R 3.4.0 software together with the spTimer package (Bakar & Sahu, 2015) was used for computation. The spTimer package allows us to apply the MCMC-based Bayesian fitting of the AR model for spatio-temporal data. Despite other abundant features such as prediction and handling missing response values, we only used basic functions of the package for model fitting and assessment.

One of the most attractive features of the package is its flexible model specification. The main function we used is spT.Gibbs(). This function permits us to create MCMC samples using the Gibbs sampling approach, which was explained in Chapter 4.4. Before using this function, we prepared the data set such that its format conforms to the requirements by the function. The spT.Gibbs() function is capable of fitting three kinds of models, which are the Gaussian Process (GP) model, Autoregressive (AR) model, and Gaussian Predictive Process model (Bakar & Sahu, 2015) respectively. Here we only take fitting the AR model for example, which is what we did in our study as discussed in Chapter 4.1, and we only explain crucial inputs of the function.

```
post.ar_full = spT.Gibbs(formula= rcrime ~ unemp_rate + hsi +
      cpi_scalar_2012 + tuition + gom + undergrad + amin + asian
      + black + hisp + white  + nonres + c_rcrime,
      data=cali, model="AR", time.data=time.data, coords=coords,
      priors=priors, initials=initials, nItr=nItr, nBurn=0,
      report=nItr, tol.dist=0.005, distance.method="geodetic:km",
      cov.fnc="exponential", scale.transform="SQRT",
      spatial.decay=spatial.decay)
```

The first input of the function is “formula”, where we define the relationship between the response and predictors. We let the model be “AR”, since the default is the GP model. The temporal components are designated in “time.data” in terms of the time units in the model such as days, months, years, etc, but in our data set year is the only time unit. The spatial locations of the institutions can be defined in “coords”, either as a $n \times 2$ data frame or as a formula, which contains their longitude and latitude coordinates. The input “cov.fnc” is required to specify the structure of the spatial covariance function, which is “exponential” as we used in this study. The default value for the numbers of the iterations “nItr” is 5,000, but we increased it to 10,000, which improved computational stability. All other inputs are set as default. This example is only for California. For Texas, we created a different data set but the code is similar.

Once a model is fitted by the function spT.Gibbs(), the spTimer package can provide its PMCC value, together with the values of its goodness-of-fit and penalty term.

An example of the output for the PMCC values is shown as:

```
         Goodness.of.fit     Penalty        PMCC
values:            6.11      122.69       128.8
```

One drawback of the spTimer package is that it is unable to perform model selection automatically. Therefore, we had to perform the forward or backward selection manually in terms of PMCC or goodness-of-fit values.

Chapter 5

RESULTS

This chapter provides the results of the spatio-temporal modeling of college crime data for California and Texas. Following the procedure as in Chapter 4, for California and Texas separately, we first obtain a few candidate models after performing forward selection and backward elimination based on the smallest goodness-of-fit and PMCC values respectively. Then we check if spatial stationarity is valid for any of these candidate models by applying each candidate (and global) model to sub-regions. Since we are able to find one global model for each state such that the assumption on spatial stationarity is approximately satisfied, we do not need to perform model selection with respect to each sub-region. The interpretations of the two final models are given at the end.

5.1 California

5.1.1 Model Selection

Following the model selection procedure introduced in Chapter 4, we obtained the following five candidate models for California shown in the Table 5-1.

Table 5-1. Five candidate models for California, with their corresponding predictors and PMCC values.

Model	Predictors	PMCC Value
1: Forward - GoF-1	tuition, undergrad	88.15
2: Forward - GoF-2	hsi, tuition, gom, undergrad	88.08
3: Forward - PMCC	hsi, cpi_scalar, tuition, nonres	86.85
4: Backward - GoF	unemp_rate, tuition, asian, nonres	86.92
5: Backward - PMCC	hsi, tuition, amin, black, nonres	87.07

There are two candidate models from the forward selection based on the smallest goodness-of-fit values in the intermediate steps, of which corresponding PMCC values are very close to each other. Therefore, both of them were chosen as candidates.

5.1.2 Checking Stationarity

Applying each candidate model above on smaller regions on the map will enable us to check the validity of spatial stationarity. As mentioned in Chapter 4.7 we may spatially divide the entire map and compare the posterior distributions for the sub-regions and the entire state when the same candidate model is fitted. Therefore, we divided the California map into two subsets after removing a few institutions. Figure 5-1 illustrates how the two subsets were obtained.

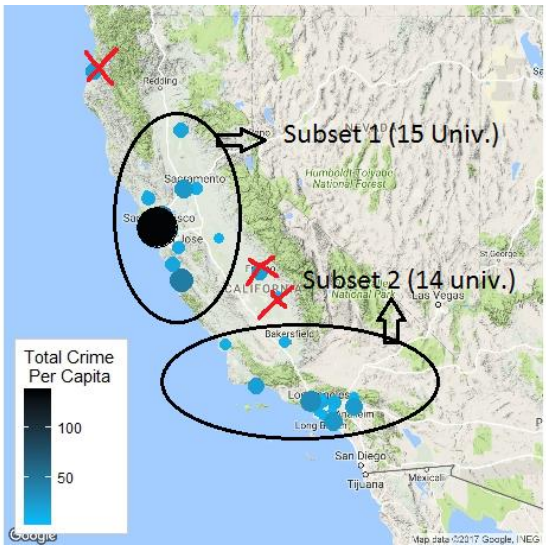

Figure 5-1. The selection of the two subsets in California to check spatial stationarity.

The subsets are ideal if they have enough institutions to fit each candidate model, be representative of the entire state (e.g., not all Hispanic-serving institutions), and optionally graphically interpretable (e.g., the San Francisco Bay Area). With this guideline, we obtained two subsets, with 15 and 14 institutions respectively, and they accumulated around San Francisco and Los Angeles. Three institutions were excluded in both subsets since one is up north of San Francisco, too far away from all others, and two are located in the middle of two subsets, which makes the assignment difficult. Then the five global models were carried out for the two subsets to see if they show a similar posterior distribution pattern with the global model.

Figure 5-2 shows the posterior distributions of the two coefficient estimates in Model 1 for the whole domain, along with the two subsets. Since the three posterior distributions are similar, spatial stationarity can be considered to be satisfied for Model 1,

which was obtained from forward selection based on goodness-of-fit. Therefore, the results from Model 1 are reliable.

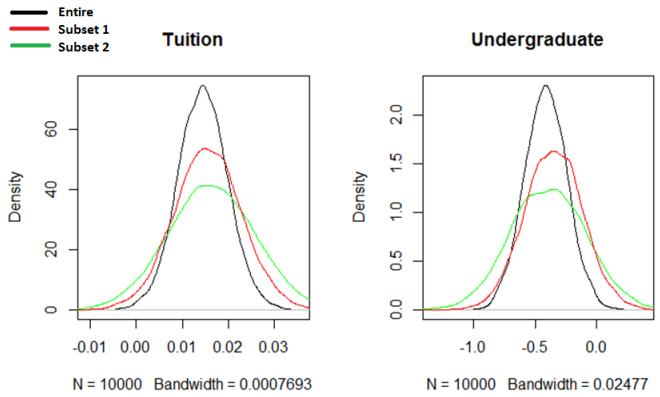

Figure 5-2. Posterior distributions of the coefficient estimates in Model 1, when it is applied to the whole domain and two subsets.

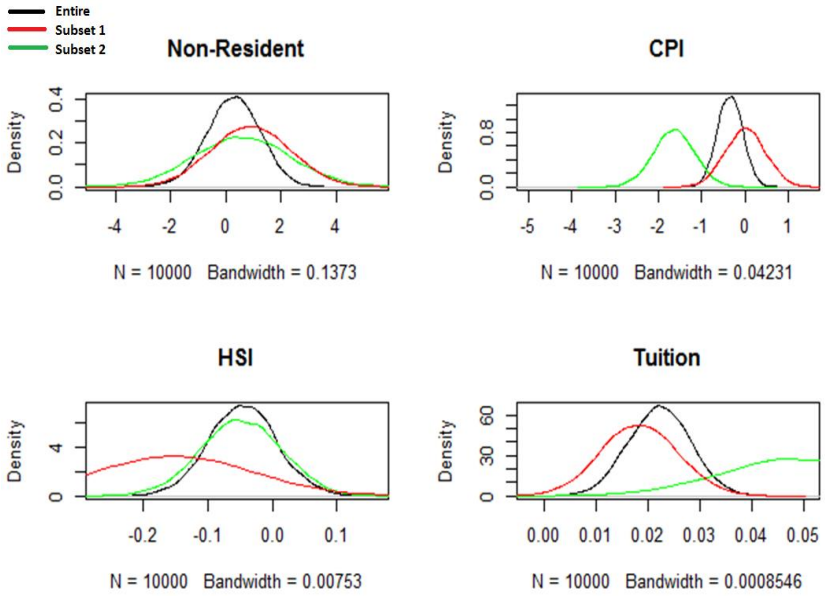

Figure 5-3. Posterior distributions of the coefficient estimates in Model 3, when it is applied to the whole domain and two subsets.

In contrast, all other candidate models violate spatial stationarity. In Figure 5-3, we provide the posterior distributions of the coefficient estimates in Model 3 as an example. Obviously, the posterior distributions for the global model and subsets are different, so spatial stationarity is invalid when fitting Model 3 and the results of Model 3 are not trustworthy. Therefore, Model 1 is the best model we chose for California.

The output for Model 1 was obtained below:

```
-----------------------------------------------------
Model: AR
Call: rcrime ~ tuition + undergrad
Iterations: 10000
nBurn: 0
Acceptance rate for phi (%): 85.2
-----------------------------------------------------
        Goodness.of.fit Penalty  PMCC
values:             0.5   88.15 88.65
-----------------------------------------------------
Computation time: 20.78  - Sec.
-----------------------------------------------------
Parameters:
               Mean  Median     SD Low2.5p Up97.5p
(Intercept)  0.6230  0.6249 0.2018  0.2244  1.0160
tuition      0.0145  0.0145 0.0054  0.0037  0.0253
undergrad   -0.4133 -0.4129 0.1736 -0.7574 -0.0667
rho          0.8998  0.8998 0.0164  0.8673  0.9315
```

The output shows that tuition and the proportion of undergraduate students are highly associated with the college crime rate in California, since zero does not fall into either of their 95% credible intervals. According to the mean value of the posterior coefficient estimates of the two predictors, higher tuition fees are related to higher college crime rates; on the contrary, a larger proportion of undergraduate students corresponds to a lower crime rate. Meanwhile, zero is outside the 95% credible interval for the autoregressive term ρ, and the mean of its posterior estimate is 0.8998 which is very

close to its upper bound 1. Therefore, the crime rate has an essential autoregressive effect, i.e., the present crime rate is strongly and positively influenced by its value in the previous year.

5.2 Texas

5.2.1 Model Selection

The same procedure was carried out for Texas and the five candidate models we obtained are given in Table 5-2.

Table 5-2. Five candidate models for Texas, with their corresponding predictors and PMCC values.

Model	Predictors	PMCC Value
1: Forward - GoF	unemp_rate, asian	121.76
2: Forward - PMCC	undergrad	124.74
3: Backward - GoF-1	control, gom, amin, nonres, c_rcrime	128.80
4: Backward - GoF-2	gom, amin, nonres	126.65
5: Backward - PMCC	control, cpi_scalar	124.93

Similarly with the situation we had for California, we obtained two candidate models, Models 3 and 4, both from backward elimination based on the smallest goodness-of-fit in the intermediate steps.

5.2.2 Checking Stationarity

As illustrated in Figure 5-4, we also obtained two subsets of institutions for Texas following the same guideline as in Chapter 5.1.2. After removing three universities which are very far from the others, the first subset contains 19 institutions, while the second

subset contains 17. There is no natural graphical accumulation of the universities in Texas, so we used a latitude line to divide the map.

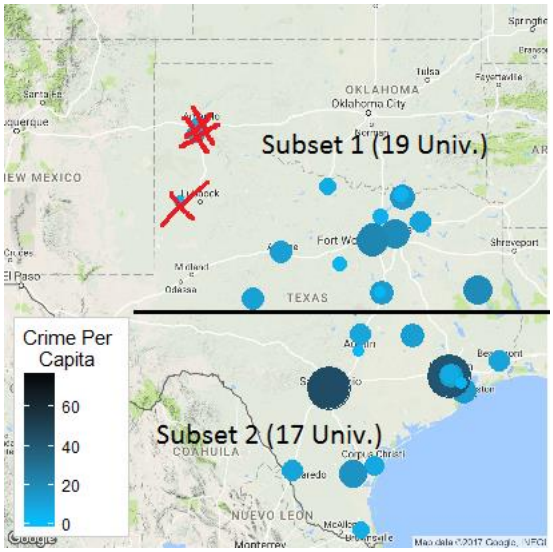

Figure 5-4. The selection of the two subsets in Texas to check spatial stationarity.

After fitting the five candidate models on the two subsets shown above, we found that Model 1, resulted from forward selection based on goodness-of-fit, implies spatial stationarity. Figure 5-5 shows the posterior distributions of the global Model 1 and its counterparts on the sub-regions.

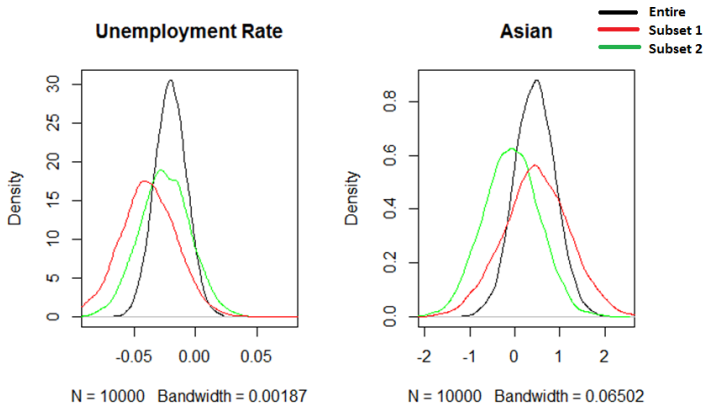

Figure 5-5. Posterior distributions of the coefficient estimates in Model 1, when it is applied to the whole domain and two subsets.

Model 1 is the only model which satisfies the stationarity assumption. All the other four models show different posterior distributions when they were applied to the sub-regions. Therefore, Model 1 is selected as the best model for Texas.

```
-------------------------------------------------
Model: AR
Call: rcrime ~ unemp_rate + asian
Iterations: 10000
nBurn: 0
Acceptance rate for phi (%): 75.97
-------------------------------------------------
        Goodness.of.fit Penalty   PMCC
values:            5.84  115.92 121.76
-------------------------------------------------
Computation time: 34.36  - Sec.
-------------------------------------------------
Parameters:
               Mean  Median     SD Low2.5p Up97.5p
(Intercept)  0.2908  0.2911 0.0987  0.0976  0.4861
unemp_rate  -0.0202 -0.0202 0.0131 -0.0460  0.0052
asian        0.4434  0.4448 0.4559 -0.4585  1.3410
rho          0.9246  0.9246 0.0164  0.8928  0.9566
```

Model 1 includes two predictors, unemployment rate and the proportion of Asian students. The output for Model 1 shows that they are not highly related with the crime rate since their 95% credible intervals contain zero. On the other hand, the autoregressive effect is also strong in Texas, which indicates the predictability of the historical crime rate.

5.3 Comparisons

Here we compare the two best models for California and Texas respectively. The similarity of the two models is that the college crime rate in both states has a strong autoregressive effect. Other than that, the two models are very different. First, the two models do not share a common predictor. Moreover, the predictors in Model 1 for California are both highly related to the response, but neither are the predictors in Model 1 for Texas. These findings, to some extent, indicate different patterns of college crime in the two states.

Chapter 6

CONCLUSIONS

The main objective of this thesis is to study the spatial and temporal patterns of the US college crime. Two states, California and Texas, that contain the highest numbers of institutins, were explored. The data set was created from three different sources: the FBI, NCES, and BLS. We fitted an AR model to study the relationship between the total crime rate and a few predictors. We obtained candidate models from forward selection and backward elimination, and then selected the best model that implies the validity of spatial stationarity.

For California, the best model includes two predictors, tuition and the proportion of undergraduate students; for Texas, the best model contains unemployment rate and the proportion of Asian students. The autoregressive effects are strong for both states, which highlights the necessity of fitting AR models. These models are adaptive to their corresponding states, but may not be generalized nationwide.

This thesis has a few limitations. First, there are too many missing values in the raw data sources. We hence removed numerous universities, so the sample size is dramatically reduced and the results from the thesis may not be generalized. We also had

to exclude some important variables, e.g., SAT scores, graduation rate, and the numbers of students per faculty member, for the same reason. The NCES's data was available until 2012, so we were unable to fit our model for recent years. Moreover, the information on a lot of private universities, e.g., Stanford University and California Institute of Technology, is unavailable from the three public sources. This makes the data set unrepresentative of the entire higher education system in the US, and the scope of this thesis restrictive. Finally, the spTimer package caused some difficulties during the analysis. Its inability to handle missing values in predictors ought to be improved. Another improvement may be needed regarding the MCMC because when the sample size is slightly larger than the number of predictors, the current version of the algorithm fails since it may generate NaN values during iterations.

In this thesis, we only focused on the total crime rate, which is a continuous response. To study violent crime, however, a different spatio-temporal modeling method is needed, since the number of violent crimes is often small and mostly zero. One direction of future work is to extend the AR model to fitting integer-valued responses. Additional and automated model selection procedures are also desirable.

REFERENCES

1. Sloan, J. J. (1994). The correlates of campus crime: An analysis of reported crimes on college and university campuses. *Journal of Criminal Justice* 22(1), 51-61.

2. Sellers, S. S., & Bromley, M. L. (1996). Violent Behavior in College Student Dating Relationships: Implications for Campus Service Providers. *Journal of Contemporary Criminal Justice* Vol. 12 No. I.

3. Janosik, S. M., & Gehring, D. D. (2003a). The Impact of the Clery Campus Crime Disclosure Act on Student Behavior. *Journal of College Student Development* Volume 44, Number 1, pp. 81-91.

4. Janosik, S. M., & Gregory, D. E. (2003b). The Clery Act and Its Influence on Campus Law Enforcement Practices. *NASPA Journal* 41:1, 182-199.

5. Payne, A. A., & Salotti, S. (2007). A Comparative Analysis of Social Learning and Social Control Theories in the Prediction of College Crime. *Deviant Behavior* 28:6, 553-573.

6. Gardella, J. H., Nichols-Hadeed, C. A., Mastrocinque, J. M., Stone, J. T., Coates, C. A., Sly, C. J., & Cerulli, C. (2014). Beyond Clery Act statistics: a closer look at college victimization based on self-report data. *Journal of Interpersonal Violence* 1–19, DOI: 10.1177/0886260514535257.

7. Franklin, C. A., Franklin, T. W., Nobles, M. R., & Kercher, G. A. (2012). Assessing the Effect of Routine Activity Theory and Self-Control on Property, Personal, and Sexual Assault Victimization. *Criminal Justice and Behavior* Vol. 39, No. 10, 1296-1315. DOI: 10.1177/0093854812453673.

8. Nobles, M. R, Fox, K. A., Khey, D. N., & Lizotte, A. J. (2012). Community and Campus Crime: A Geospatial Examination of the Clery Act. *Crime & Delinquency* 59(8) 1131-1156.

9. Luan, H., Quick, M., & Law, J. (2016). Analyzing local spatio-temporal patterns of police calls-for-service using Bayesian integrated nested Laplace approximation. *ISPRS International Journal of Geo-Information* 5(9), 162.

10. Gonzales, A. (2015). Education: The Secret to Crime Reduction?, Thesis, New York University.

11. Wang, X., & Brown, D.E. (2012). The spatio-temporal modeling for criminal incidents. *Security Informatics* 20121:2 DOI: 10.1186/2190-8532-1-2.

12. LaRue, E. (2013). Patterns of Crime and Universities: A Spatial Analysis of Burglary, Robbery and Motor Vehicle Theft Patterns Surrounding Universities in Ottawa. Thesis, Simon Fraser University.

13. Bakar, K. S., & Sahu, S. K. (2015). sptimer: Spatio-temporal Bayesian modelling using R. *Journal of Statistical Software*, 63(15), 1-32.

14. Crime in Schools and Colleges. Retrieved from https://ucr.fbi.gov/nibrs/crime-in-schools-and-colleges.

15. National Center for Education Statistics. Retrieved from https://nces.ed.gov/ipeds/datacenter.

16. Local Area Unemployment Statistics. Retrieved from https://www.bls.gov/lau/.

17. Uniform Crime Reporting. Retrieved from https://ucr.fbi.gov/.

18. Kahle, D., & Wickham, H. (2013). ggmap: Spatial Visualization with ggplot2. *The R Journal* Vol. 5/1.

19. Sahu, S. K., Gelfand, A. E., & Holland, D. M. (2007). High-resolution space–time ozone modeling for assessing trends. *Journal of the American Statistical Association* 102(480), 1221-1234.

20. Gelfand, A. E. (2012). Hierarchical Modeling for Spatial Data Problems. *Spat Stat.* 1: 30–39. doi:10.1016/j.spasta.2012.02.005.

21. Cressie N. A. C., & Wikle C. K. (2011). Statistics for Spatio-Temporal Data. *John Wiley & Sons*, New York.

22. Gelfand A. E., Diggle P. J., Fuentes M., & Guttorp P. (2010). The Handbook of Spatial Statistics. *Chapman & Hall/CRC*, New York.

23. Gelman A., Carlin J. B., Stern H. S., Rubin D. B. (2004). Bayesian Data Analysis. 2nd edition. *Chapman & Hall/CRC*, Boca Raton.

24. Gelfand A. E., & Smith A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85(410), 398-409.

25. Gibbs sampling. Retrieved from https://en.wikipedia.org/wiki/Gibbs_sampling

26. Bivand, R. S., Pebesma, E. & Gómez-Rubio V. (2008) Applied Spatial Data Analysis with R, *Springer-Verlag* 305–308 , New York.

27. Fotheringham, A. S., Brunsdon, C., & Charlton, M. E. (2002). Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. *Wiley*, Chichester.

28. Matérn, B. (1986). Spatial Variation. 2nd edition. *Springer-Verlag*, Berlin.

29. Handcock M. S., & Stein M. L. (1993). A Bayesian Analysis of Kriging. *Technometrics*, 35, 403-410.

30. Handcock M. S., & Wallis, J. (1994). An Approach to Statistical Spatial-Temporal Modelling of Meteorological Fields. *Journal of the American Statistical Association*, 89, 368-390.

31. Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2004). Hierarchical Modeling and Analysis for Spatial Data. *Chapman & Hall/CRC*, Boca Raton.

32. Cressie, N. A. C. (1993). Statistics for Spatial Data. *John Wiley & Sons*, New York.