The Pennsylvania State University
The Graduate School
College of Information Sciences and Technology

URBAN COMPUTING WITH MOBILITY DATA: A UNIFIED APPROACH

A Comprehensive Document in
College of Information Sciences and Technology
by
Hongjian Wang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

April 2016

The comprehensive document of Hongjian Wang was reviewed and approved* by the following:

Jessie Li
Assistant Professor of College of Information Sciences and Technology
Thesis Advisor, Chair of Committee

C. Lee Giles
Professor of College of Information Sciences and Technology

Anna Squicciarini
Associate Professor of College of Information Sciences and Technology

Daniel Kifer
Associate Professor of Department of Computer Science & Engineering

*Signatures are on file in the Graduate School.

Abstract

With the advent of information age, various types of data are collected in the context of urban spaces , including taxi pickups/drop-offs, tweets from users, air quality measure, noise complaints, POIs, and many more. It is crucial to use these data to understand region interactions in the city. Research questions that I am interested in are: 1) understand nodes using links; 2) understand links using nodes; 3) identify causal structure.

For my thesis work, I want to capture the complicated interactions of regions in the urban space. Traditionally, due to lack of flow data, spatial similarity is widely used to capture interactions. Recently, the availability of movement data enables us to study the interactions incurred by social flow. While various approaches are proposed, they still have the following drawbacks. 1) The definition of interactions from different data sources is ad-hoc. 2) Most models assume uniform correlation among different spatial regions. As a matter of fact, in my preliminary study I have observed that when training separate models on Chicago south and north, two models are different and the estimation results will be better.

The goal of this thesis is to develop a unified probabilistic graphical model to capture the interactions of heterogeneous flow in the urban context. Starting from a preliminary study on estimating the Chicago community level crime with POI and taxi flow. The intuition is that the POI complements the demographics features, and the taxi flow acts as a hyperlink to connect non-adjacent community areas. The results suggest that both newer type of features correlates with the crime and improves the estimation significantly. Next, the urban data are categorized as nodal feature and dyadic feature, which belong to a spatial unit and a pair of units respectively. The graphical model is a natural way to capture the complicated interaction. Lastly, model spatial variations, i.e., the same features in different regions have different parameters. In the graphical model, we use separate nodes to represent different regions, so that learned relations are locally specific.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express the deepest appreciation to my committee Dr. Jessie Li, Dr. C. Lee Giles, Dr. Anna Squicciarini, and Dr. Daniel Kifer.

I would like to express my special gratitude to my adviser Dr. Jessie Li, as well as my collaborator Dr. Corina Graif and Dr. Daniel Kifer, who gave me the golden opportunity to work on this interesting problem on the topic crime inference in Chicago, during which also helped me in learning new theories and developing new models. I am really thankful to them.

Lastly i would also like to thank my parents and friends who helped me a lot in finalizing this project within the limited time frame.

Chapter 1
Introduction – Urban Computing

1.1 Model Region Interactions Incurred By Social Flow

In the urban space, the movement of human population connects two disjoint regions, and brings influence from one to the other. *My research focus is to capture the complicated interactions in the urban space with human mobility data.*

In recent years, there are some large datasets of urban taxi going public [] under the FOIL[1] The taxi data usually contains the pickup/drop-off time and location of one trip. By aggregating the taxi data, we are able to get the social flow among different regions. Social flow data (e.g., commuting flow, taxi trajectories) are sensitive resources that urban planners can use to address city issues.

Consider the following two examples. These two examples show that considering interactions of regions is an important problem to answer.

Example 1. *Policy makers are deciding where to construct a shelter for families that are victims of violence. They understand the value of locating the shelter geographically far from violent neighborhoods. One possible choice is to locate the shelter in a neighborhood that is 10 miles from the violent neighborhoods where vulnerable families previously lived. However, a deeper analysis may reveal that the new neighborhood, though geographically removed from the old neighborhood, may still have strong social flows (connections caused by commutes, family visits) with the old neighborhood. Emerging research suggests that a great deal of crime happens*

[1]Freedom of information request law.

in areas that are socially connected to offenders' neighborhoods. This suggests that shelters may benefit from being located in a neighborhood that is also socially isolated from violence (e.g., with weak communication and commuting interactions with the violent neighborhoods that shelter residents fled from) while socially connected to jobs, services, and resources.

Example 2. *Another example. The smart route selection scenario does not stand, since google map provides real time feedback on traffic jam. No need to infer*

1.2 Research Problems

In the urban space, a model of region **interactions** bring solutions to the several fundamental data mining questions. We propose to view the city as a spatial network of communities linked by "hyperlink" flow. The questions we can answer are

- Understand nodes using links. Estimate an unobserved property of focal community, given the observations on other communities and other types of data. For example, the crime rate in a residential neighborhood could be impacted by non-adjacent but flow-connected neighborhoods, because the residents in the neighborhoods are also exposed to and influenced by the environment in the workplace.

- Understand links using nodes. Are certain properties of two connected nodes associated with the type and volume of the flow connections? For example, given the crime profiles of two connected communities, which type of interactions (taxi, LEHD, or space continuity) is more important in forming the crime properties?

- Identify the fundamental dependency structure among properties of nodes. For example, the crime rate of two connected communities may show a strong correlation. However, this does not necessarily mean that high crime rate in one community lead to the high crime in another. The crime is very likely to be caused by other properties.

1.3 Existing Work

Traditionally, most urban research do not use flow data, since the data is not available. For example, researchers have used demographic information (e.g., population poverty level, socioeconomic disadvantage, racial composition of population) to estimate the crime rate in a community [6]. However, such demographic information only contains partial information about the neighborhoods and does not dynamically reflect the changes in the community. Using only demographic information will result in a relative error of at least 30% for crime rate estimation in Chicago (refer to experiment section in the paper).

Use spatial influence as interactions. Since there is no flow data available, spatial similarity is widely accepted assumption to model the region interactions. **Spatial similarity** assumes that spatially adjacent regions tend to have similar properties. In urban space, most data reflect the properties of human beings, such as the traffic volume, crime count, and geo-tagged tweets. Therefore, these data are attached to human crowd instead of a specific location. In the urban space, the intuition behind the **spatial similarity** is that human movement is regular, and most of our daily activities are conducted in a limited area.

There is one existing study using the geographical influence [7] to estimate the crime rate, i.e., the crime in the nearby communities can be propagated to the focal community. But this geographical influence is of little help in improving the crime inference on top of demographic feature, with at most 0.4% relative improvement in our experiments. This is probably because the nearby communities also share similar demographics, which limits the additional benefit of geographical influence. In the study, spatial autoregressive model is used

$$y = \rho_1 W^{spatial} y + X\beta + \epsilon, \tag{1.1}$$

where y is the crime count, X is the demographics property, and $W^{spatial}$ is an $n \times n$ spatial distance matrix.

Simple extension on spatial influence model It is easy to extend the spatial autoregressive model in Equation 1.1 with a social flow matrix.

$$y = \rho_1 W^{spatial} y + \rho_2 W^{network} y + X\beta + \epsilon, \tag{1.2}$$

where $W^{network}$ is an $n \times n$ social distance matrix. It is clearly that the interaction term $\rho_2 W^{network} y$ is ad-hoc, since there could be other combinations like $\rho_3 W^{network} x_1$.

1.4 Challenges

The interactions usually take the the following form $m(f_{ij}, x_j)$.

Given one type of social flow, define interaction is non-trivial. Given only one type of social flow, there are too many possibilities in constructing interactions. 1) The flow matrix can take various form. For example, we can chose normalize it or not. When normalizing the flow matrix, we can chose whether normalize by in-flow or out-flow. 2) There are also many nodal properties that could interact with their neighbors, such as various demographics features. 3) Different kinds of function can be used to define interactions. Take the product of flow and nodal properties is the most straightforward choice. However, sometimes it also makes sense to further apply a distance exponential decay on previous product. Therefore, the construction of interaction is totally ad-hoc.

Given multiple types of social flow, the interaction is more difficult to define. It is possible we have multiple types of social flow (e.g. taxi flow, commuter transit). For one pair of regions, should we build separate interaction over different social flows, or sum over all flows to get one interaction? When we take the sum, should we weight different social flow differently, and how?

The interaction could be spatial non-stationary. Most models in existing work are global model, which assumes the statistic interaction does not vary over space. However, some urban data have spatial non-stationary property. Use Chicago crime as example, shown in Figure 1.1. It is clear that using global estimates of relationships can present misleading interpretations of local relationships.

1.5 A Unified Graphical Model

We propose a unified graphical model to capture the interactions among regions. Since various flow connect regions into various network. A graphical model is a natural way to represent the whole system. Additionally, this model can address the two drawbacks mentioned above.

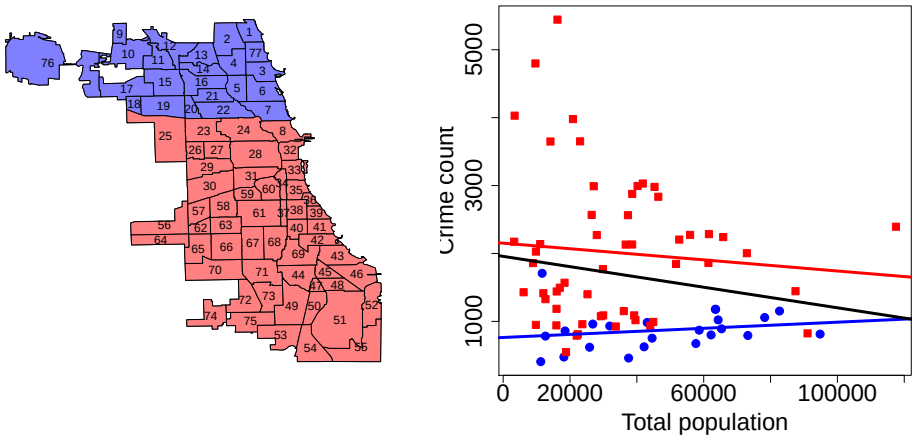


Figure 1.1. The crime count vs. total population relationship shows spatial non-stationary property.

To address the non-trivial definition of interaction. In the graphical model we model the interaction of two communities is a hidden variable. This hidden variable is connected with the observations on both flow and nodal properties. We can learn the interaction from the network.

To address the spatial non-stationarity. Each region is built as separate node in the graphical model. This way, the conditional probability

$$P(y_i|X_i)$$

is different at different regions, which is equivalent to have mutliple local models to capture the different relations between y and X.

In next chapter we study the first problem - "using links to understand nodes". We use crime inference as an example, in which we use an enhanced spatial autoregressive model to predict crime count of a community using its neighbors. In Chapter 3. We first discuss the missing property of a spatial autoregressive model, which is the spatial non-stationarity. To address this, we propose a conditional random field based graphical model. This model is superior than other method in the literature, such as geographically weighted regression. Finally, there is a research plan in Chapter 4.

Chapter 2
Understand Nodes Using Links

In this chapter we address the first proposed research question. Namely, we use the observations in other regions and other types of data to estimate one unobserved property in focal region. We call this problem **inference problem**. At the very beginning, we give a generalized inference problem definition. Following the general definition, we look at one example of crime inference.

2.1 General Problem Definition

In this section, we give a generalized definition of the inference problem. Suppose we have a set of regions $r_1, r_2, \cdots, r_n$, and we are interested in one property y_i for region r_i. In addition to $\vec{y}$, we also have other properties $\vec{x}_i$ observed on region r_i, and the set of all auxiliary properties are denoted as X. It is noteworthy that both $\vec{y}$ and X are nodal properties. The spatial adjacency among all regions are known as W^0, where W^0 is spatial adjacency matrix. The hyperlink mobility flow is also observed as W^k for type $k = \{1, 2, \cdots\}$. The entry w_{ij}^k in W^k refers to the quantity flow from r_i to r_j of type k.

The inference problem is that for a given region r_t, whose y_t is unobserved, we try to use $\{y_i\}\backslash y_t$ together with X and $W^k, k \in \{0, 1, 2, \}$ to estimate y_t. Mathematically, we have

$$\hat{y_t} = f(\{y_i\}\backslash y_t, X, W^k), \tag{2.1}$$

where f is the estimation model of any choice.

In the next section, we will use the crime inference as an example to show

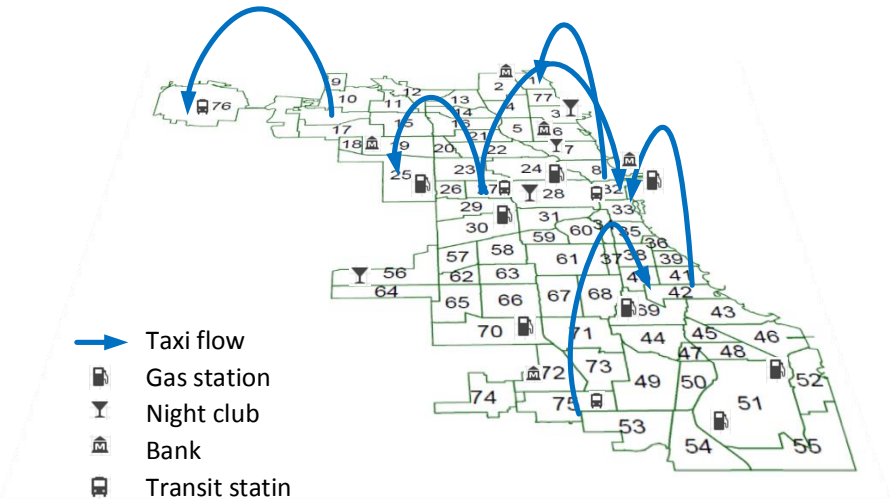


Figure 2.1. An illustration of various types of features we used in Chicago. The POI distribution across community areas reflects profiles of the region functionality. The taxi flow connects nonadjacent regions and act as a "hyperlink".

how this inference problem is solved in the literature, and how do we enhance the existing model.

2.2 Crime Inference as One Example

In Figure 2.1, we show that taxi flow as a newer type of big data could provide us new insights to understand some traditional socioeconomic urban problems. A huge amount of taxi flow data reflect how people commute in the city. In previous studies, when using geographical influence [7], people assume that a community is affected by the spatially nearby communities. However, communities are not only affected by spatially-close communities. Even if two communities are distant in geographical space, they could have a strong correlation if there are many people frequently travel between these two communities [8]. We hypothesize that taxi flows may be considered as "hyperlinks" in the city that connect the locations and we use such data to estimate crime rates. Our experiments show very promising results – adding taxi flow data on top of all other features can further decrease the

error by 5%.

2.2.1 Related Work

In the criminology literature researchers have studied the relationship between crime and various features. Examples are historical crime records [9, 10], education [11], ethnicity [12], income level [13], unemployment [14], and spatial proximity [7]. In data mining field, newer type of data are used in the study. For example, there are works using twitter to predict crime [15, 16], and works using cellphone data [17, 18] to evaluate crime and social theories at scale.

Overall, the existing work on crime prediction can be categorized into three paradigms.

Time-centric paradigm. This line of work focuses on the temporal dimension of crime incidents. For example, in a study [9], the authors propose to use a self-exciting point process to model the crime and gain insights into the temporal trends in the rate of burglary. In another study [19], the authors investigate the temporal constraints on crime, and propose an offender travel and opportunity model. This paper validates the claim that a proportion of offending is driven by the availability of opportunities presented in the offender's routine lives.

Place-centric paradigm. Most existing work adopt a place-centric paradigm, where the research question is to predict the location of crime incidents. The predicated crime location is usually refereed by the term *hotspot*, which has various geographical size. There are plenty of works on exploration of the crime hotspots. For example, in a study [20] the authors use criminal offense records to identify spatio-temporal patterns at multiple scales. They employ various quantitative tools from mathematics and physics and identify significant correlation in both space and time in the crime behavioral data. Short *et al.* [21] use a simple model to study the dynamics of crime hotspots and identify stable hotspots, where criminals are modeled as random walkers. Bogomolov *et al.* [18] use human behavioral data derived from mobile network and demographic sources, together with open crime data to predict crime hotspots. They compare various classifiers and find random forest has the best prediction performance. The paper [15] bases on automatic semantic analysis to understand natural language Twitter posts, from which the crime incidents are reported. Some other work [22, 23] employ the kernel density

estimation (KDE) to identify and analyze crime hot spots. Those works form another form of crime prediction, which relies on the retrospective crime data to identify areas of high concentrations of crime. In [24], the authors extend the crime cluster analysis with a temporal dimension. They employ the space-time variants of KDE to simultaneously visualize geographical extent and duration of crime clusters.

Population-centric paradigm. In the last paradigm, research focuses on the criminal profiling at individual level and community level. At the individual level, [10] aim to automatically identify crimes committed by same individual from the historical crime database. The proposed system called *Series Finder*, is designed to find and classify modus operandi (M.O.) of criminals. At the community level, Buczak *et al.* [25] use fuzzy association rule mining to find crime pattern. The rules they found are consistently held across all regions. The paper constructs association rules from population demographics in community. In another paper [17], the authors use computation method to validate various social theories at a large scale. The data they used is mobile phone data in London, from which they mine the people dynamics as features to correlate with crime.

Our problem is different from the first two categories of work, mainly because our innovation mostly lies in using newer type of data to enhance the commonly used traditional counterpart. More specifically, we use POI to enhance the demographics information, and use taxi flow as hyper link to enhance the geographical proximity correlation. Although our problem does not consider the temporal dimension of crime in depth, it could be a promising supplement to better profile crime. Our problem dose not predict the location of any particular crime incident. Therefore the methods proposed in place-centric method are not applicable in our problem. However, the features we proposed may be incorporated in those crime prediction model. Our problem falls into the third paradigm, because we are trying to profile the crime rate for Chicago community areas. In our problem, the community areas are well-defined and stable geographical regions. The newly proposed POI feature and taxi hyper link provide a unique perspective in profiling the crime rate across community areas.

2.2.2 Data Description and Feature Extraction

The crime dataset in Chicago has detailed information about the time and location (i.e., latitude and longitude) of crime and the types of crime. In our problem, when we use term crime count, we often refer to crime count in a region (i.e., community area) in a year. The *community area* is used as our geographical unit of study, since it is well-defined, historically recognized and stable over time [26]. In total, there are 77 community areas in Chicago. Crime rate is the crime count normalized by the population in a region. We use vector $\vec{y} = [y_1, y_2, \ldots, y_n]$ to denote the crime rate in region i.

The crime data of Chicago are obtained from City of Chicago data portal [27]. Chicago is the city with most complete crime data that are made public online. The crime dataset contains the incident date, location (strict name and GPS coordinates), and primary type from year 2001 to 2015. In total there are $5,856,414$ recorded crime incidents over 15 years, which is an average $390,417$ crimes incidents per year. We visualize the crime normalized by population in Figure 2.2, from which we can see that the downtown area has the highest crime rate.

In this example we study the crime rate inference problem. More specifically, we estimate the crime rate of some regions given the information of all the other regions. Without loss of generality, we assume there is one community area t with crime rate y_t missing, and we use the crime rate of all the other regions $\{y_i\}\backslash y_t$ to infer this missing value. Our problem is mathematically formalized as follows

$$\hat{y_t} = f(\{y_i\}\backslash y_t, X), \tag{2.2}$$

where X refers to observed extra information of all those community areas.

We consider two types of features X for inference:

- Nodal feature. Nodal features describe the characteristics of the focal region. Such features include demographic information and Point-of-Interest (POI) distribution. Demographics are frequently used in literature, but POI is a newer type of big data, which we find significantly improve the crime inference accuracy.

- Edge feature: (1) Geographical influence. Geographical influence considers the crime rate of the nearby locations. This feature has been extensively used

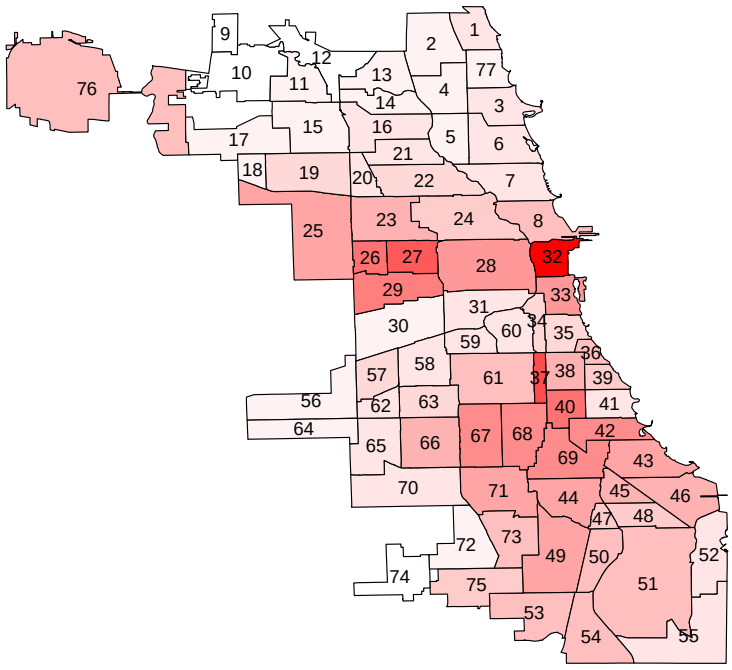


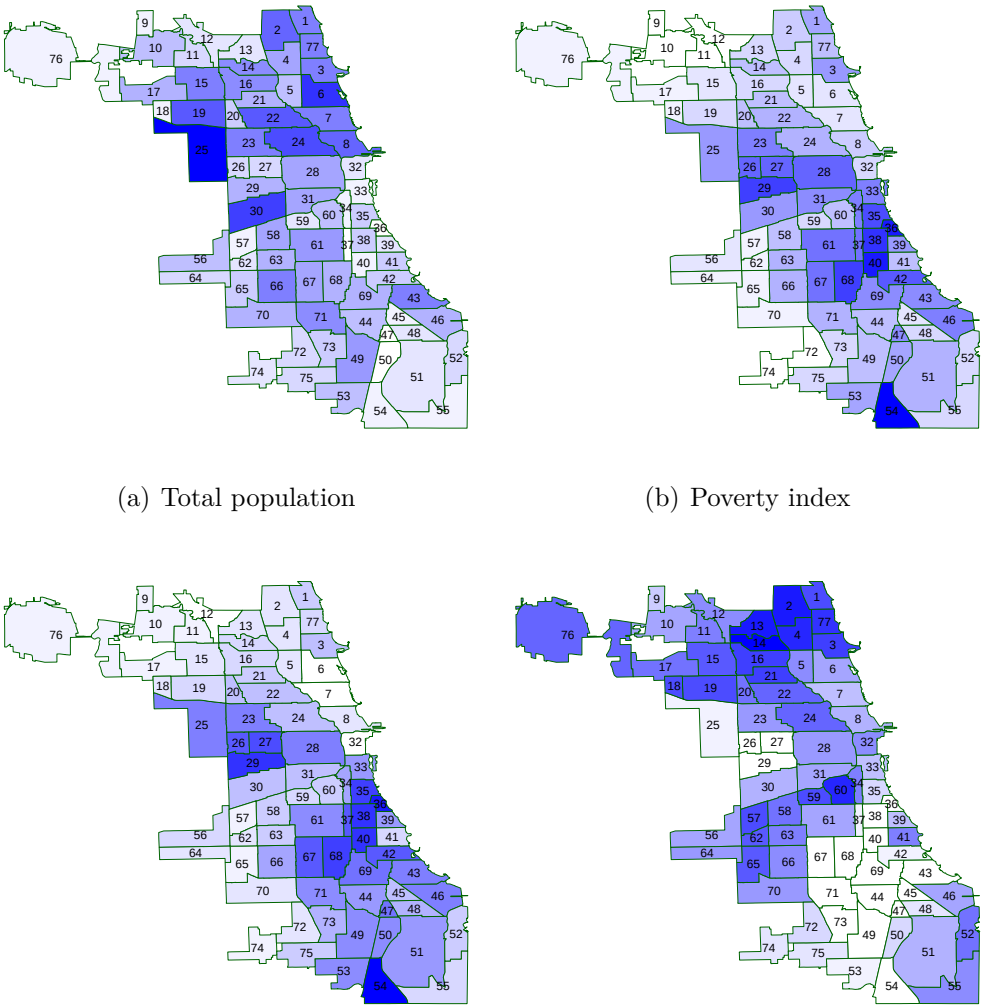
Figure 2.2. Crime rate of Chicago by community areas. The community area #32 is Chicago downtown, which has the highest crime rate.

in literature as well. To estimate the focal region, the crime rate of nearby regions are weighted according to spatial distances. (2) Hyperlink by taxi flow. Locations are connected through the frequent trips made by humans, which can be considered as the hyperlinks in space. This type of feature has never been studied in literature. We propose to use taxi trips to construct the social flow. Our hypothesis is that similarity in the crime rate of two regions should correlate with the social flow strength between these two regions.

Below we will describe the datasets used to construct features and the characteristics of these features.

Nodal Feature: Demographics

Socioeconomic and demographic features of neighborhoods have been widely used to predict crime [18, 28–30]. Previous studies have shown that crime rate correlates with certain demographics. For example, [6, 31] suggests that population diversity leads to less crime in certain neighborhoods. In our study, we include demographic



(a) Total population

(b) Poverty index

(c) Disadvantage index

(d) Ethnic diversity

Figure 2.3. (a)-(d) Demographics in Chicago by community areas. Darker colors indicate higher values.

information from the US Census Bureau's Decennial Census of 2010 [32] and American Community Survey's five-year average estimates between 2007 and 2011. We use year 2010 data because we are evaluating crime rates in 2010-2013. The demographics include the following features:

total population, population density, poverty, disadvantage index, residential stability, ethnic diversity, race distribution.

The poverty index measures the proportion of community area residents with

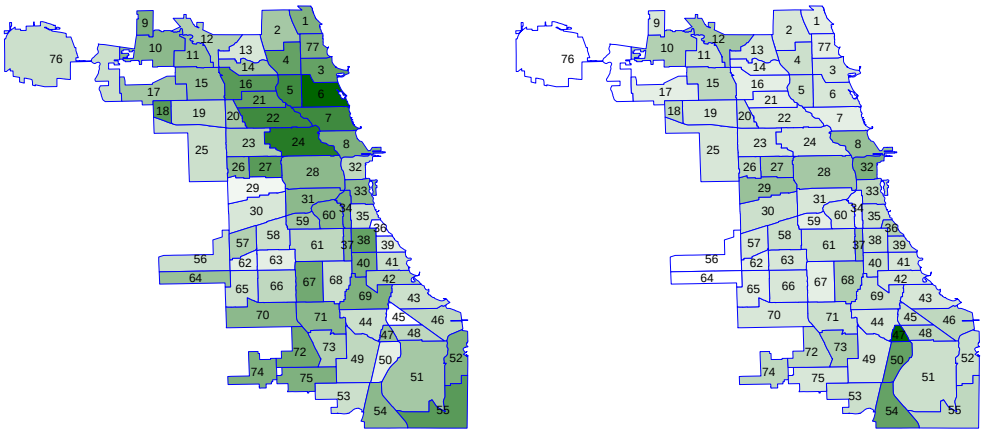
income below the poverty level. The disadvantage index is a composite scale based on prior work [33], a function of poverty, unemployment rate, proportions of families with public assistance income, and proportion of female headed households. The residential stability measures home ownership and proportion of residents who lived in the neighborhood for more than one year. Racial and ethnic diversity is an index of heterogeneity [6] based on six population groups, including: Hispanics, non-Hispanic Blacks, Whites, Asians, Pacific Islanders and others.

Figure 2.3 visualizes the crime rate and demographics features in Chicago by community areas. Comparing with Figure 2.2, it is clear that the crime rate and poverty index and disadvantage index are consistent, the ethnic diversity shows an inverse correlation, and the total population has little correlation with crime.

Table 2.1 shows the Pearson correlation coefficient between various demographics features and the crime rate at community area level. The corresponding p-value is also calculated and shown in the table to indicate the significance of the correlation coefficient. There are in total 77 community areas in Chicago. Table 2.1 shows such correlation with several most correlated features. We can see that the poverty index and disadvantage index positively and strongly correlate with crime, while the ethnic diversity negatively correlates with crime. Other features such as total population, population density, and residential stability have weaker correlations. One counter-intuitive observation is that the total population has a weak and negative correlation with crime. The reason is that we use crime rate in each community area, which is already normalized by the population, and therefore the total population and population density have less impact.

Table 2.1. Pearson correlation between demographic features and crime rate (***** indicates significant correlations with p-value less than 5%).

Feature	Correlation	p-value
Total Population	-0.1269	0.2716
Population Density	-0.1972	0.0855
Poverty Index	**0.5573***	1.403e-07
Disadvantage Index	**0.5959***	1.082e-08
Residential Stability	-0.0453	0.6965
Ethnic Diversity	**-0.5545***	1.678e-07
Percentage of Black	**0.6696***	2.779e-11
Percentage of Hispanic	**-0.3820***	0.0006



(a) Nightlife (b) Professional

Figure 2.4. Plot the POI ratio per neighborhood. The saturation of color is proportional to the ratio value. The "professional" category distribution is more consistent with the crime distribution, and therefore it is the most correlated with crime. Meanwhile, the "nightlife" category is not positively correlated with Chicago crime.

Nodal Feature: Point-of-Interest (POI)

While demographics are traditional census data, POI is a type of modern data that provide fine-grained information about locations. We collect POI from FourSquare [34]. POI data from FourSquare provide the venue information including venue name, category, number of check-ins, and number of unique visitors. We mainly use the major category information because categories can characterize the neighborhood functions. There are 10 major categories defined by FourSquare:

food, residence, travel, arts & entertainment, outdoors & recreation, college & education, nightlife, professional, shops, and event.

In total, we have crawled 112,000 POIs from FourSquare for Chicago. Most of these POIs are in downtown area of Chicago. We normalize the POIs count per category by the total POI count in a neighborhood and plot two selected category, i.e. nightlife and professional, in Figure 2.4. The darker colored neighborhoods in Figure 2.4 are the ones with a higher portion of residence POIs.

In Table 2.2 we show the Pearson correlation between POI category and crime rate. The category "professional" is most significantly correlated with the crime rate. Under the professional POI category, there are some venues with a large population concentration, such as transportation center, convention center, community center,

Table 2.2. Pearson correlation between POI category and crime rate (***** indicates significant correlations with p-value less than 5%).

POI category	Correlation	p-value
Food	-0.1543	0.1803
Residence	-0.0610	0.5984
Travel	-0.0017	0.9883
Arts & Entertainment	-0.0049	0.9661
Outdoors & Recreation	0.0668	0.5637
College & Education	-0.0078	0.9473
Nightlife	-0.1553	0.1775
Professional	**0.3221***	0.0043
Shops	-0.1676	0.1450
Event	0.2196	0.0549

and co-working space. In those venues, the population volume is high and residential stability is low, therefore the professional POI counts positively correlates with crime rate. One counter-intuitive observation is that "nightlife" category is not positively correlated with crime (-0.1553). This can be explained through Figure 2.4(a). The majority of nightlife venues in Chicago locate in northern area, while most crime incidents occur in downtown area.

Edge: Geographical Influence

Together with the US census demographics data, we also collected the boundary shape files of Chicago, which are used to calculate the geographical influence feature.

Previous studies have also shown that the crime rate at one location is highly correlated with nearby locations [35, 36]. Such geographical influence is also frequently used in the literature [37, 38], which is calculated as:

$$\vec{F^g} = W^g \cdot \vec{Y}, \tag{2.3}$$

where W^g is the spatial weight matrix. If region i and j are not geospatially adjacent, $w_{ij}^g = 0$; otherwise, $w_{ij}^g \propto distance(i,j)^{-1}$.

In Figure 2.5, we plot crime rate with respect to geographical influence calculated in Eq. 2.3. We observe an obvious positive correlation, which means if nearby neighborhoods have a high crime rate, the focal neighborhood is more likely to have a high crime rate. We also do observe a few outliers in Figure 2.5. These neighborhoods show different crime rate in their nearby neighborhoods compared

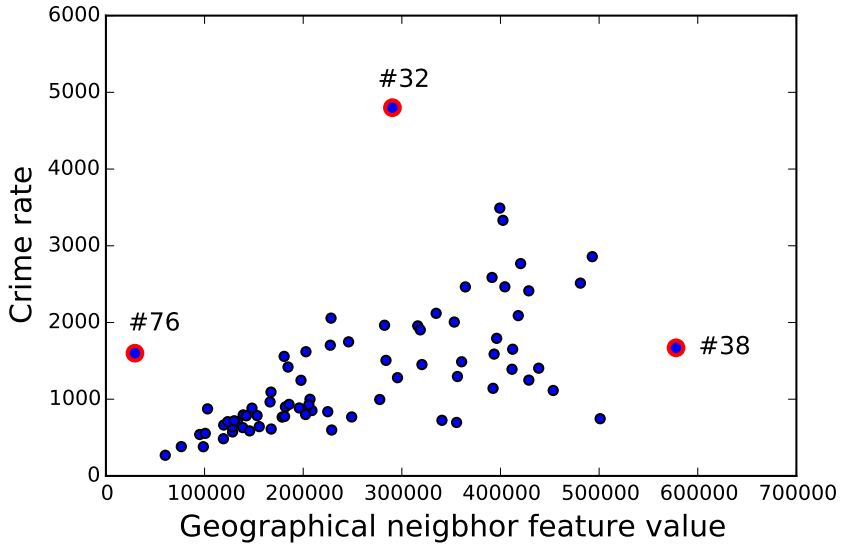


Figure 2.5. The geographical influence feature correlation with crime. In the plot we marked out three outliers and their corresponding community area ID.

to their own. For example, as we can also see in Figure 2.2, community area #38 locates in an area where the the neighbors have high crime rates but its crime rate is relatively low; in contrast, neighborhood #32 has a high crime rate even though its neighbors have relatively low crime. The community area #76 home of the O'Hare International Airport is far from most of other community areas, however its own crime rate is relative high.

Edge: Hyperlinks by Taxi Flow

In our Chicago taxi dataset, there are $1,048,576$ taxi trips in total during the October to December in 2013. For each trip the following information are available: pickup/dropoff time, pickup/dropoff location, operation time, and total amount paid. We requested the taxi trip records from Chicago taxi commission pursuant of the Freedom of Information Law. Figure 2.6 shows a visualization of the major flows at community level.

One of our hypothesis is that the social interaction among two community areas propagates crime from one region to another. The Chicago taxi data captures the social interactions among various community areas. To calculate this first, we first map all taxi trips to community areas to get the taxi flow w_{ij} $\forall i, j \in \{1, 2, \cdots n\}$. Then the taxi flow lag is constructed by the product of social flow and the crime

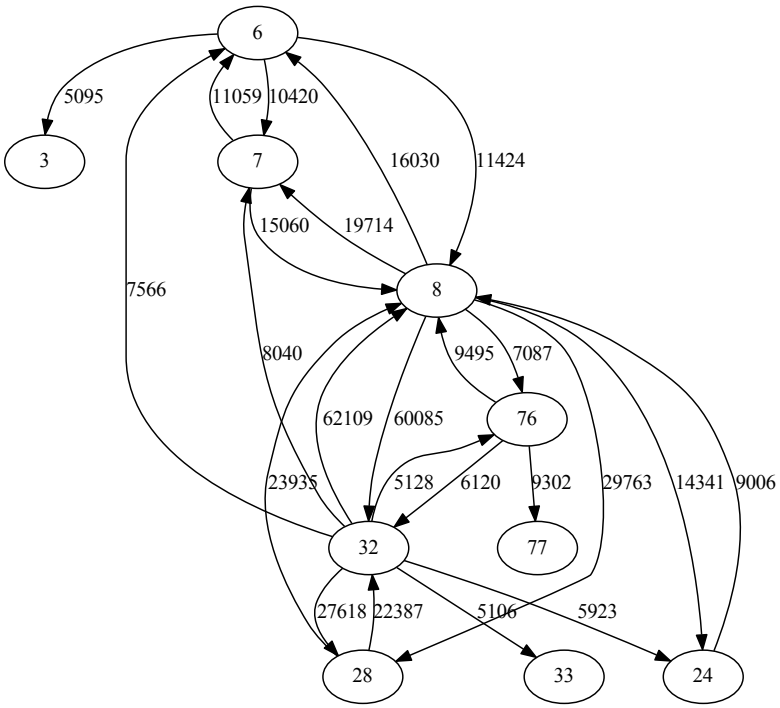


Figure 2.6. Major taxi flows between neighborhoods. We set a threshold ($> 5,000$) on the flow and only plot the high volume flow. The label on the node is the ID of the corresponding community areas. We can see that there are several hub community areas, such as #6, #8, #32, which are all in the downtown areas. The label on the edge shows how many taxi trips are commuting through the two community areas for three months in 2013.

rate of neighboring regions as follows

$$\vec{F^t} = W^t \cdot \vec{Y}. \tag{2.4}$$

The taxi flow W^t is a matrix with entry w_{ij} denoting the taxi flow from i to j. Note that $\forall i$, $w^s_{ii} = 0$ in matrix W^t, because we have to exclude the crime in focal area from its own predictor. The semantic of this taxi flow feature is how many crime in the focal area is contributed by its neighboring areas through social interaction.

The correlation between taxi flow and crime rate is shown in Figure 2.7. From the scatter plot, we can see that overall the crime rate is positively correlate with the taxi flow. There are two outliers clearly shown in Figure 2.7. The community

area #32 is the downtown Loop, which has the highest crime rate and is hard to predict by taxi flow. Another anomalous community area #47 has relatively low crime rate by itself. However, this area has a lot of in flows from high-crime communities.

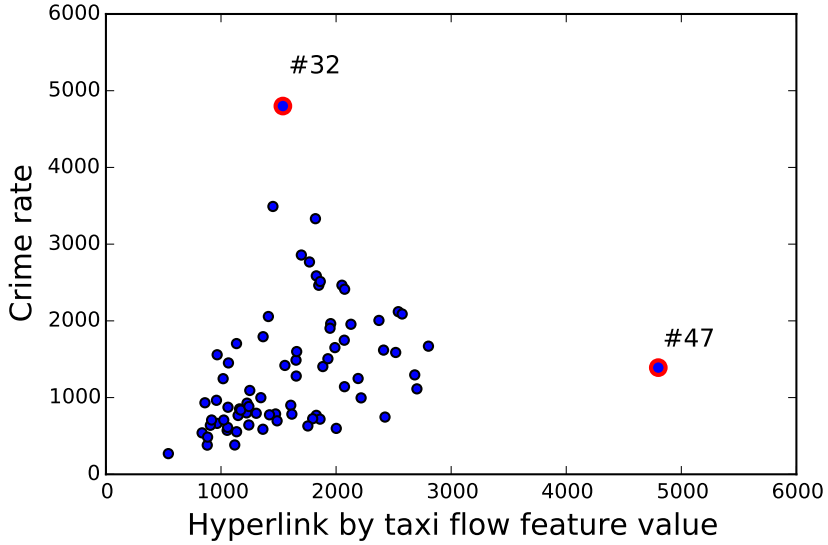


Figure 2.7. Correlation between taxi flow and crime rate. In the plot, we marked out three outliers and their corresponding community area ID.

2.2.3 Spatial Autoregressive Inference Model

Linear Regression

The most straightforward prediction is linear regression model. This model assumes the error terms follow a Gaussian distribution $\epsilon \sim \mathcal{N}(0, \sigma^2)$. As a result the parameter distribution also follows a Gaussian distribution. This assumption makes the model less generative, since in real applications, there is no way to ensure the dependent variable has a Gaussian error term.

Equation 2.5 gives the linear regression formulation of our problem.

$$\vec{y} = \vec{\alpha}^T \vec{x} + \beta^f W^f \vec{y} + \beta^g W^g \vec{y} + \vec{\epsilon}, \tag{2.5}$$

where $\vec{x}$ represents the nodal features including demographics and POI distribution, W^f is the flow matrix of taxi flow, and W^g is the spatial matrix representing the

geographical adjacency. On the right-hand side, ϵ is the only stochastic variables, and all other terms are fixed observation values. Therefore, we incorporate all the fixed observations into one term X, and we get the standard regression problem

$$E(y) = Xw + \epsilon.$$

In order to learn the regression parameter w, we can use a maximum likelihood estimator. Since $\epsilon = y - Xw$, the joint probability of error term is

$$P(\epsilon|w) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-Xw)^2}{2\sigma^2}}. \tag{2.6}$$

Maximizing the joint probability gives us the optimal solution.

Linear regression gives negative prediction

One obvious drawback is that the linear regression is not a count prediction model, since it will give negative number as prediction. For example, we find the suspicious community #32 case. Refer to Figure 2.2, we know this community is the downtown. The crime count for community #32 is $7,709$ in 2010. However, the linear regression base model gives $-1,448$ as prediction. This result is not acceptable in a count prediction model. We further look into the features of community #32 to figure out why it has negative crime count estimation. It turns out that community #32 has $12,175$ venues in total, which is 10 times more than the average of other communities $(1,011)$. In our learned model, the venue count feature has a negative coefficients, which indicates that popular places tend to have less crime incidents. The big difference in the venue count feature lead to a negative prediction for community # 32.

Poisson regression as a count prediction model

To address this issue, a count prediction model is a natural selection. The *Poisson regression* is another form of regression, more appropriate for count data than linear regression [39] [40]. With shortened notation X, Poisson regression model has the exponential function as link function

$$E(y) = e^{Xw}. \tag{2.7}$$

This comes from the assumption that y follows Poisson distribution with mean λ. Additionally, the mean *lambda* is determined by observed independent variables X,

with the link function $\lambda = e^{Xw}$. Adding all together, the joint probability of y is

$$P(y|w) = \frac{e^{-e^{Xw}}(e^{Xw})^y}{y!}. \tag{2.8}$$

Compared with the linear regression, the negative log-likelihood function of Poisson regression is derived from the dependent variable itself, unlike linear regression, which is derived from the joint distribution of error term.

However, Poisson regression enforces the mean and variance of dependent variable y to be equal. This restriction leads to the "over-dispersion" issue for some real problems, that is the presence of larger variability in data set than the statistical model expected. In our crime dataset, the mean of crime count for all communities is $4,787$, while the variance is 1.6×10^7. The variance is almost the square of the mean, which significantly violate the Poisson distribution assumption. Therefore, we should look for other count prediction model.

Negative binomial regression addresses over-dispersion

To allow larger variance in the predicted value, we introduce the Poisson-Gamma mixture model, which is also known as *negative binomial regression*. The negative binomial regression has been used in similar work [41].

Given that the crime rate y follows Poisson distribution with mean λ. In order to allow for larger variance, now the λ itself is a random variable, distributed as a Gamma distribution with shape $k = r$ and scale $\theta = \frac{1-p}{p}$. The probability function of y becomes

$$\begin{aligned} P(y|r,p) &= \int_0^\infty P_{Poisson}(y|\lambda) \cdot P_{Gamma}(\lambda|r,p) d\lambda \\ &= \int_0^\infty \frac{\lambda^y}{y!} e^{-\lambda} \cdot \lambda^{r-1} \frac{e^{-\lambda(1-p)/p}}{(\frac{p}{1-p})^y \Gamma(r)} d\lambda \\ &= \frac{\Gamma(r+y)}{y!\Gamma(r)} p^k (1-p)^y \end{aligned} \tag{2.9}$$

This is exactly the probability density function of negative binomial distribution.

In negative binomial regression, the link function is

$$E(y) = e^{Xw+\epsilon}. \tag{2.10}$$

The error term e^ϵ is the mixture prior, and we assume it follows Gamma distribution

with shape parameter $k = \frac{1}{\theta}$, so that it has mean $E(e^{\epsilon}) = k\theta = 1$ and variance $Var(e^{\epsilon}) = k\theta^2 = \theta$. This setting ensures the $E(y) = e^{Xw} \cdot e^{\epsilon} = e^{Xw}$.

2.2.4 Evaluation of negative binomial regression model

Evaluation Settings

We adopt the leave-one-out evaluation to estimate the crime rate of one geographic region given all the information of all the other regions. When we construct the spatial/social lag variable for the training data, the effect of testing region is completely removed. For example, if region y_t is the testing region, the remaining $\{y_i\} \backslash y_t$ become the training set. For any y_j in the training set, its geographical influence feature and taxi flow feature are constructed from $\{y_i\} \backslash \{y_t, y_j\}$.

In the evaluation, we estimate the crime rate for testing community area. The accuracy of estimation is evaluated by mean absolute error (MAE) and mean relative error (MRE).

$$MAE = \frac{\sum_i^n |y_i - \hat{y}_i|}{n} \tag{2.11}$$

$$MRE = \frac{\sum_i^n |y_i - \hat{y}_i|}{\sum_i^n y_i} \tag{2.12}$$

Performance Study: Negative Binomial Regression vs. Linear Regression

We evaluate the estimation accuracy under various feature combinations. The leave-one-out evaluation results are shown in Table 2.3. We run both linear regression model and negative binomial model on five consecutive years, 2010 – 2014. Both MAE and MRE are shown in the table. We have four types of features, demographics, POI, geographical influence and taxi flow. We test the various settings of feature combinations.

We compare the estimation error of negative binomial model with the linear regression base model. We use a incremental settings, where new features are added on top of previous one. The results are shown in Figure 2.8.

It is clear that the negative binomial model significantly outperform the linear regression model. Meanwhile, the negative binomial model captures our intuition well. Namely, adding new features will effectively improve the estimation accuracy.

Table 2.3. Performance evaluation. Various feature combinations are shown in each column. The linear regression model and negative binomial results are compared by year group.

			Settings							
Column ID			1	2	3	4	5	6	7	8
Features[1]		D	✓	✓	✓	✓	✓	✓	✓	✓
		G					✓	✓	✓	✓
		P		✓		✓		✓		✓
		T			✓	✓			✓	✓
Year	Model[2]	Error								
2010	LR	MAE	394.41	416.98	408.09	406.93	394.78	432.45	402.25	416.41
		MRE	0.294	0.311	0.304	0.304	0.295	0.323	0.300	0.310
	NB	MAE	391.53	333.14	395.64	323.47	389.55	350.06	387.43	**320.75**
		MRE	0.292	0.249	0.295	0.241	0.290	0.261	0.289	**0.239**
2011	LR	MAE	380.22	409.30	396.97	401.11	379.61	422.94	389.39	408.91
		MRE	0.295	0.318	0.309	0.312	0.295	0.328	0.302	0.320
	NB	MAE	381.11	332.62	388.81	328.94	378.84	345.24	381.33	**335.97**
		MRE	0.296	0.259	0.302	0.256	0.294	0.268	0.296	**0.253**
2012	LR	MAE	378.91	412.95	401.54	412.20	376.53	423.88	399.25	419.93
		MRE	0.306	0.334	0.325	0.333	0.304	0.343	0.322	0.339
	NB	MAE	386.31	337.24	389.58	331.41	384.23	352.22	381.67	**345.49**
		MRE	0.312	0.273	0.315	0.268	0.310	0.284	0.308	**0.279**
2013	LR	MAE	367.89	420.81	390.75	402.75	369.24	433.48	388.92	412.31
		MRE	0.324	0.370	0.344	0.354	0.325	0.381	0.342	0.362
	NB	MAE	376.08	333.92	373.08	312.63	377.57	350.33	368.49	**319.86**
		MRE	0.331	0.294	0.328	0.275	0.332	0.308	0.324	**0.281**
2014	LR	MAE	331.28	375.53	349.00	350.31	329.93	386.90	345.79	361.28
		MRE	0.326	0.369	0.343	0.345	0.324	0.380	0.340	0.355
	NB	MAE	340.73	293.52	339.17	274.45	336.09	308.18	326.07	**273.27**
		MRE	0.335	0.289	0.334	0.270	0.331	0.303	0.321	**0.269**

[1] D – demographic features, G – geographical influence, P – POI features, T – taxi flow feature.
[2] LR – Linear Regression, NB – Negative Binomial Regression.

In Table 2.3, we can see that in different years and under most settings, the negative binomial regression significantly outperforms the linear regression (with only a few exceptions when using only demographic feature). When using all the features, NB is significantly better than LR with at least 6% improvement in relative error. One reason is that the negative binomial is a count prediction model, which assumes some distribution for the predicted variable and guarantee its positivity. Another reason is that it is difficult to get very precise estimation of crime rate,

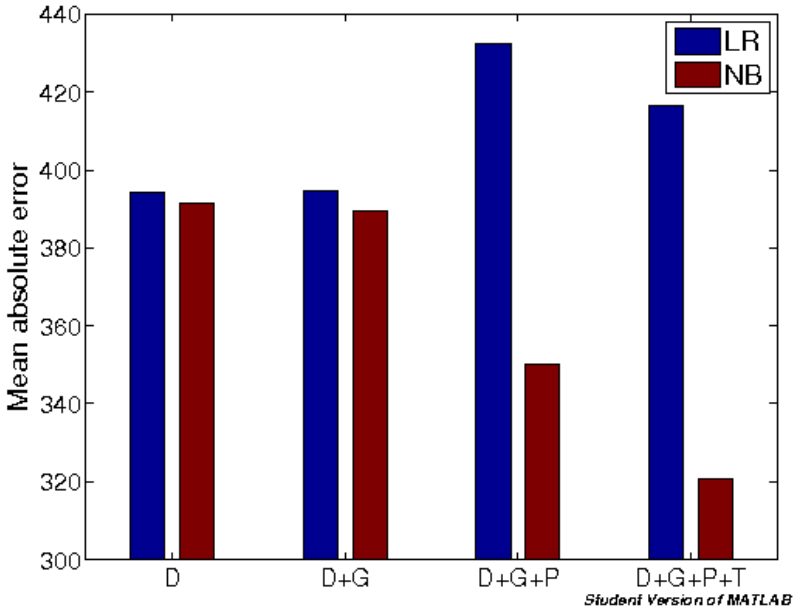


Figure 2.8. The inference error for linear regression base model. *D – demographic features, G – geographical influence, P – POI features, T – taxi flow feature

and negative binomial model allows a large variance in the estimated crime rate. Therefore negative binomial is more appropriate for crime rate estimation than linear regression.

2.3 Conclusion

In this chapter, we discuss the a general inference problem that using information on the links to understand the nodes.

We extend the spatial autoregressive model in the literature by adding new types of flow into the model. We call this model flow augumented spatial autoregressive model. This modification makes it impossible to use standard T-test to get the significance of the model coefficients. Therefore, we use the Monte-Carlo tests, where a repeated leave-one-out schemes is designed.

The flow augumented spatial autoregressive model has its own weakness, which will be discussed in next chapter. In the next chapter, we presents the unified graphical model to model the interactions.

Chapter 3
A Unified Graphical Model

In this chapter, we propose to use a unified graphical model to model the interactions among regions. We start from a discussion on the major flaw of previous flow augmented spatial autoregressive model, that as a global spatial model is assumes spatial stationarity. However, in the real problem there is usually a spatial non-stationarity over some properties. For example, we observe that the relationship between crime count and total population in each community presents such spatial non-stationarity.

To fix the spatial non-stationarity issue, the most widely used method is called *geographically weighted regression* (GWR) in the literature. In the next section, we will discuss the pros and cons to employ GWR to capture the spatial non-stationarity. Since our problem has multiple "hyperlink" flows in addition to space continuity, the idea of GWR must be generalized over a high dimension space consisted of multiple interactions. However, this generalization is non-trivial, and thus we propose a graphical model to address this challenge.

The graphical model is a natural way to capture the complicated interactions among regions. With the same graphical model, we further show that other data mining tasks in urban space can be solved as well. Therefore, we believe this graphical model is superior than other model in capturing the interactions, which consequently could be a united model.

3.1 Capture Spatial Non-stationarity

3.1.1 Global Model vs. Local Model

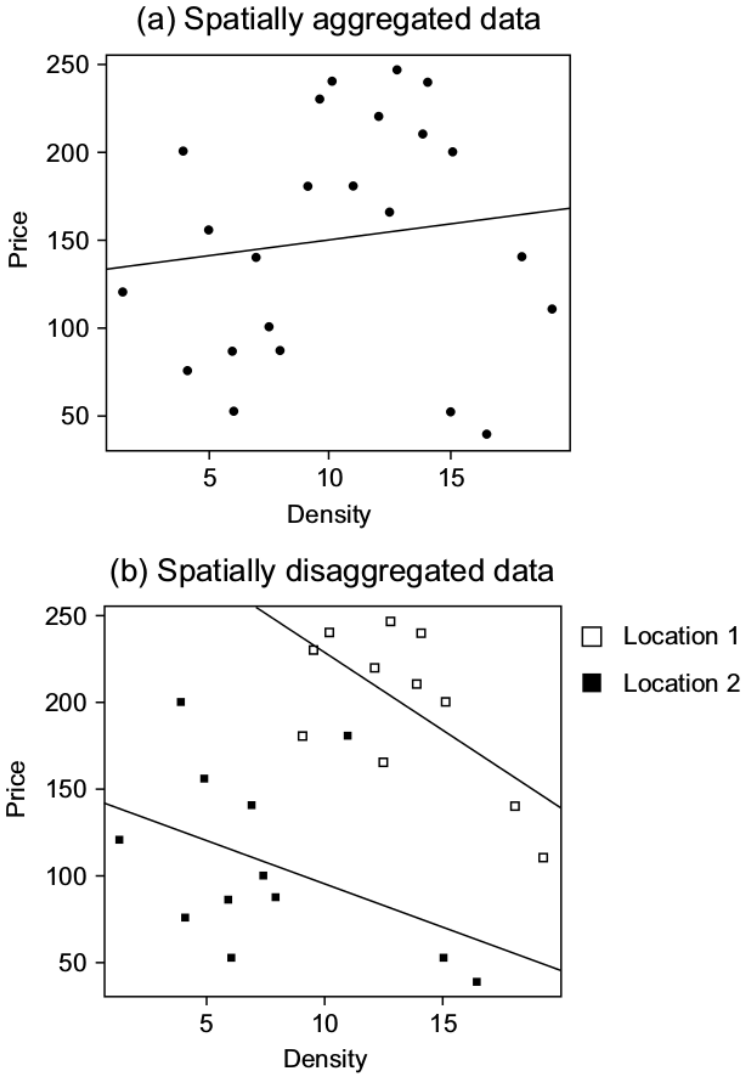


Figure 3.1. A spatial example of Simpson's Paradox (from [**?**]).

Global model assumes the statistical relationships does not change over space. However, some statistical relationship is not stationary over space, which requries specific treatment for different locations, i.e. a local model at different places.

According to [42], using global estimates of statistics can present misleading interpretations of local models. This is shown as an example in Figure 3.1, known

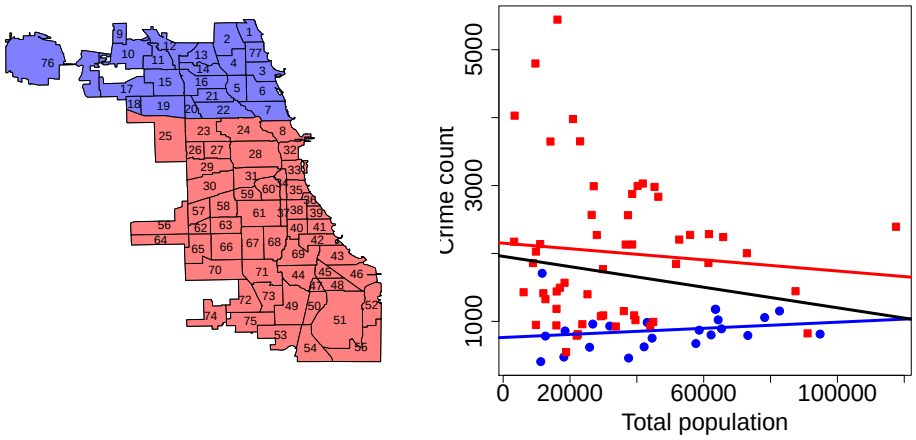


Figure 3.2. The crime count vs. total population relationship shows spatial non-stationary property.

as Simpson's Paradox [43]. The reason of Simpson's Paradox is that there is a hidden distribution of properties over space, which leads to opposite results in global model when aggregate subgroups over space.

In the Chicago crime inference example, we also observed similar phonomenon, as shown in Figure 3.2. We divide Chicago into two half: i) downtown north, ii) downtown and downtown south. The intuition is that the Chicago south usually have more crime than north, therefore we want to split Chicago into low-crime half and high-crime half. The splition of Chicago communites are visualized in left figure of Figure 3.2. On the right of Figure 3.2, we present the total crime plotted against total population. It is clear that in the high-crime half of Chicago the relation is almost neutral (coefficient as 0), and in the low-crime half of Chicago the relation is postive. However, the global model (black line) presents a negative correlation.

3.1.2 An Existing Solution: Geographically Weighted Regression

Geographically weighted regression (GWR) is a term introduced in [42] to describe a family of regression method in which the coefficients β are allowed to vary spatially. GWR uses the coordinates of each sample or zone centroid t_i, as a target point for

a form of spatially weighted least squares regression. The model is of the form:

$$y = X\beta(t) + \epsilon \tag{3.1}$$

The coefficient $\beta(t)$ are determined at different regression point locally by its neighboring sample points. A weighting schema or spatial kernel W is employed to weight neighboring sample points accordint to its distance to regression point. The Figure 3.3 is an example of the weighting function.

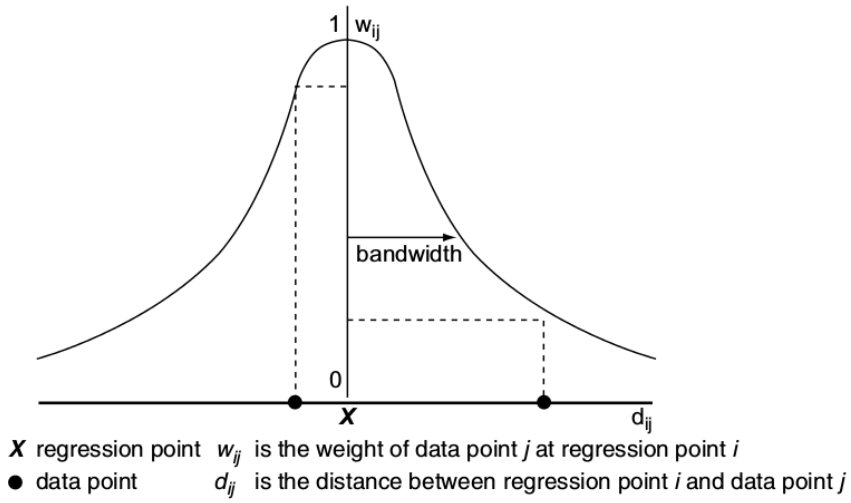


Figure 3.3. A spatial kernel. Example from [42]

The GWR idea is able to be applied on top of spatial autoregressive (SAR) model, so that SAR model deals with spatial non-stationarity. In the next we discuss another similar model called adaptive spatial model. After that, we discuss a common drawbacks of this two models.

3.1.3 An Alternative Solution: Adaptive Spatial Model

We use various types of data to estimate the crime count in community area (CA) of Chicago. For each CA we have observations on crime count and demographics. For each pair of CAs we also have observations on the taxi flow and spatial distance. One straightforward method is to build a regression model from all the features we observed to the crime counts.

We have one interesting observations is that in the south and north part of

Chicago, the significance of different features are different. Therefore, the idea is to learn a dynamic weights for different spatial region.

Suppose we have n regions in total, $R = \{r_1, r_2, \cdots, r_n\}$. The following notations are used

crime count at r_i	y_i
demographics at r_i	$\mathbf{d}_i$
taxi flow between r_i and r_j	f_{ij}
taxi flow weight matrix for r_i	$\mathbf{f_i}$
spatial weight matrix for r_i	$\mathbf{g_i}$
social flow lag variable for r_i	$s_i = \mathbf{f}_i^T \mathbf{y}$
spatial flow lag variable for r_i	$p_i = \mathbf{g}_i^T \mathbf{y}$

Table 3.1. Symbols for the dynamic coefficient model.

Dynamic linear regression model

For simplicity we use linear regression model

$$y_i = \mathbf{w}_1^T \mathbf{d}_i + w_2 s_i + w_3 p_i + w_4,$$

where $\{w\}$ are the coefficients.

To simplify notations, we use $\mathbf{x}_i$ denote all the available predictors for region r_i,

$$\mathbf{x}_i = [\mathbf{d}_i, s_i, p_i, 1].$$

Then the model becomes

$$y_i = \mathbf{w}^T \mathbf{x}_i.$$

Now we use a dynamic model, where $\mathbf{w}$ is different for various regions. This leads to

$$y_i = \mathbf{w}_i^T \mathbf{x}_i.$$

The problem with formulation is that there are too many parameters to learn. To address this issue, we use the constraint that **spatially adjacent regions share similar coefficients**.

We use S_{ij} to denote the adjacency of r_i and r_j. And the aforementioned constraint is formulated as

$$\min \sum_{i,j} S_{ij} ||\mathbf{w}_i^T - \mathbf{w}_j^T||_2^2$$

The several choice of S_{ij}

- Binary indicator. $S_{ij} = 1$ if two regions are contiguous, otherwise $S_{ij} = 0$.
- The reverse distance between r_i and r_j.

The overall objective is

$$\min_{\mathbf{W}} \sum_i ||y_i - \mathbf{w}_i^T \mathbf{x}_i||_2^2 + \eta \sum_{i,j} S_{ij} ||\mathbf{w}_i^T - \mathbf{w}_j^T||_2^2 + \theta ||\mathbf{W}||_F^2 \tag{3.2}$$

Optimization

Rewrite the Frobenius norm in the last term

$$||\mathbf{W}||_F^2 = \sum_i ||\mathbf{w}_i - \mathbf{0}||_2^2.$$

Therefore the Equation 3.2 is rewritten as

$$\min_{\mathbf{W}} \sum_i ||y_i - \mathbf{w}_i^T \mathbf{x}_i||_2^2 + \eta \sum_{i,j \in 0,\cdots,N} S_{ij} ||\mathbf{w}_i - \mathbf{w}_j||_2^2, \tag{3.3}$$

where $\mathbf{w}_0 = \mathbf{0}$ and $S_{0i} = 1$ for $\forall i$.

To solve the objective in Equation 3.3, we use variable splitting. Namely, when optimizing for $\mathbf{w}_i$, we assume all other $\mathbf{w}_{j,j \neq i}$ are fixed. The sub-problem is

$$\min_{\mathbf{w}_i} ||y_i - \mathbf{w}_i^T \mathbf{x}_i||_2^2 + \eta \sum_j S_{ij} ||\mathbf{w}_i - \mathbf{w}_j||_2^2. \tag{3.4}$$

The update on $\mathbf{w}_i$ is

$$\mathbf{w}_i = \min_{\mathbf{w}_i} ||y_i - \mathbf{w}_i^T \mathbf{x}_i||_2^2 + \eta \sum_j S_{ij} ||\mathbf{w}_i - \mathbf{w}_j^{(t)}||_2^2.$$

The closed-form solution is

$$\mathbf{w}_i = (\mathbf{x}_i^T \mathbf{x}_i + \eta \sum_j S_{ij} \mathbf{I})^{-1} (y_i \mathbf{x}_i + \eta \sum_j S_{ij} \mathbf{w}_j) \tag{3.5}$$

Inference

We use the **leave-one-out** setting to infer and evaluate the crime rate of new community area.

Suppose the we want to estimate the crime rate y_i of CA_i. During the training process, we hold everything about CA_i out (including y_i, flow coming in and leaving from CA_i). Then training the model on CA_j, $\forall j \neq i$, which gives us w_j, $\forall j \neq i$. To infer the y_i, we need estimate the model coefficient w_i first. Follow the same intuition that model on CA_i is only similar to all its neighboring models, we have

$$\min_{\mathbf{w}_i} \sum_{j,\forall j \neq i} S_{ij} ||\mathbf{w}_i - \mathbf{w}_j||_2^2 + ||\mathbf{w}_i||_2^2 \tag{3.6}$$

After getting $\mathbf{w}_i$, we infer y_i by

$$\hat{y}_i = \mathbf{w}_i^T * \mathbf{x}_i \tag{3.7}$$

3.1.4 Comparison of GWR and Adaptive Model

Both the GWR and Adaptive model address the spatial non-stationarity by building local models. Two models use different approaches to model the spatial continuity.

GWR implicitly model the spatial continuity by using similar sample points for two nearby regression models. Meanwhile, the adaptive model explicitly put a constrait to make two nearby models similar.

Both mehtods can be extened with newer type of interactions. We use W^0 denote the spatial adjacency matrix, and use W^k, $k = 1, 2, \cdots$ to denote the interactions matrix of data type k. Therefore, we are extending the two dimensional distance weighting kernel into a high dimension distance measure.

However, this high dimension distance measure is non-trivial to calculate. Suppose now we have taxi flow and LEHD flow in addition to geospatial adjacency. The tricky question to ask is which interaction is more important to connect two regions? We can give weight coefficient to different interactions, so that

$$W = \alpha_0 W^0 + \alpha_1 W^1 + \cdots + \alpha_k W^k \tag{3.8}$$

where $\sum_0^k \alpha_i = 1$.

However, the W in both GWR and adaptive model is pre-given. Therefore, incorporating heterogeneous flows is the major challenge of both GWR and adaptive model.

3.2 Graphical Model to Capture Complicated Interactions

The methods in the literature is not originally designed to handle heterogenous interactions among regions, and therefore have the major challenge mentioned above. The heterogenous interactions generate different network structure for us. Therefore, we propose to model the interactions of two regions as a latent variables with a graphical model.

3.2.1 Problem Formulation

We want to predict the crime rate y_i of each geographical grid (tract/community area) g_i. The available observations are demographics features $\vec{\mathbf{x}}_i$ of each g_i from census, and the interactions among grids. We denote the interactions as $\vec{\mathbf{f}}_{ij}$ for grid pair g_i, g_j, and examples of such interactions are social flow and geospatial distance.

3.2.2 Conditional random field model

Potential Function Each grid is a node, and its crime rate y_i is the hidden variable that we want to estimate. Two kinds of fixed parameters are observed for each grid g_i. The first one is the demographic features $\vec{\mathbf{x}}_i$. The second is the interactions among grids, such as social flow and geospatial distance, denoted by $\vec{\mathbf{f}}ij$.

We use Conditional Random Field (CRF) shown in Figure A.1 to model the dependency of nodal features. The learning goal is to estimate the conditional probability of y given $\vec{\mathbf{x}}$ and $\vec{\mathbf{f}}$

$$P(y|\vec{\mathbf{x}}, \vec{\mathbf{f}}) \tag{3.9}$$

This model can handle the spatial non-stationary issue, since we can learn the conditional probability separately at different locations.

In the CRF model, we factorize the probability distribution of y to a series of potential functions ψ on the clique.

$$P(Y) = \frac{1}{Z} \prod_{c \in C} \psi(c) \tag{3.10}$$

Use C_1 to denote the set of cliques of size 1 with the form $\langle y_i \rangle$, and C_2 to denote size-2 clique. We define the potential function as follows:

$$\psi_{C_1} = exp(-|y_i - \vec{\alpha}^T \cdot \vec{\mathbf{x}}_i|) \forall i \in [1, n], g_i \in C_1, \tag{3.11}$$

$$\psi_{C_2} = exp(-|y_i - y_j - \vec{\mathbf{w}}^T \cdot \vec{\mathbf{f}}_{i,j}|) \forall i, j \in [1, n], g_i, g_j \in C_2, \tag{3.12}$$

where $\vec{\alpha}$ and $\vec{\mathbf{w}}$ are all positive coefficients.

The distribution of Y is given by

$$P(Y) = \frac{1}{Z} \left[\prod_{i=1}^{n} \psi_{C_1}(y_i) \times \prod_{i=1}^{n} \prod_{j=i}^{n} \psi_{C_2}(y_i, y_j) \right] \tag{3.13}$$

$$P(Y) = \frac{1}{Z} exp \left(- \sum_{i=1}^{n} |y_i - \vec{\alpha}^T \cdot \vec{\mathbf{x}}_i| - \sum_{i=1}^{n} \sum_{j=i}^{n} |y_i - y_j - \vec{\mathbf{w}}^T \cdot \vec{\mathbf{f}}_{i,j}| \right) \tag{3.14}$$

The inference of CRF model is given in the Appendix A, and we refer interested reader there.

3.3 Graphcial Model Solve Other Proposed Problems

The graphical model is not only useful to solve inference problem on nodes using links, but also able to solve other problems.

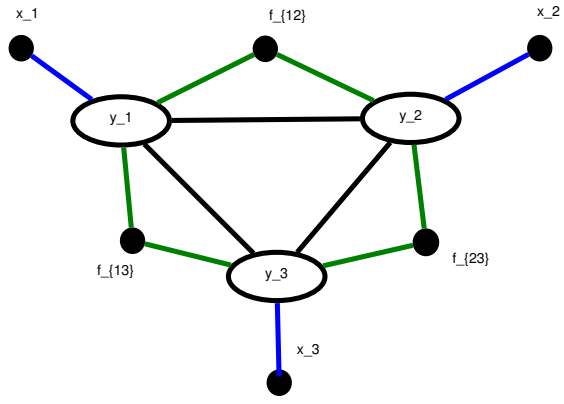


Figure 3.4. The CRF model of the crime rate y_i for each grid g_i.

3.3.1 Understand links using nodes.

In a graphical model shown in Figure A.1, we can estimate the flow f_{ij} using the nodal attributes y_i and $\vec{\mathbf{x}}_i$ as well.

This problem answers important questions about the link. For example, we can answer when two regions are becoming very dissimilar in $\vec{\mathbf{x}}$, what will happend to the taxi interaction between them? What about the LEHD flow.

3.3.2 Understand the causal structures

Graphical model also helps us understand the real dependency structure of node property and flow between nodes. The assumptions that crime is neighboring community will influence crime in focal community is over-simplified. Actually, we do not know what impacts crime in focal community. It is very likely to be nodal properties in neighboring communites that impacts crime in focal community. There are various explanations behind, as shown in Figure 3.5.

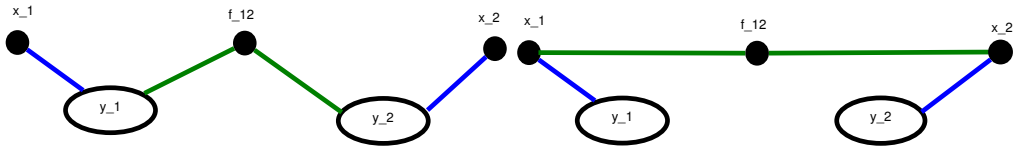


Figure 3.5. Various assumptions on interations behind the inference problem.

Chapter 4
Research Plan and Schedule

The first part of improvement on the base model is already finished and under review now. Solve two proposed problems: 1) develop a dynamic spatial model and 2) develop a unified probabilistic graphical model in the next two years. Specifically,

Fall 2016

- Work on the dynamic spatial model on top of negative binomial regression.
- Look for other geospatial local model to compete.
- By the end of semester have a paper ready to submit.

Spring 2017

- Submit the dynamic spatial model paper to KDD. Prepare for revision.
- Develop a unified model starting from CRF for the crime inference.
- Look into other urban problems to see whether the model can be easily generalized.

Fall 2017

- Evaluate the unified CRF model on crime inference problem, and hopefully another real problem.
- Start drafting the paper.

Spring 2018

- Submit the unified model to KDD. Prepare for revision.
- Write my thesis and prepare for final defense.

Appendix A
Inference of Conditional Random Field Model

The first model comes to mind.

A.1 Potential Function

Each grid is a node, and its crime rate y_i is the hidden variable that we want to estimate. Two kinds of fixed parameters are observed for each grid g_i. The first one is the demographic features $\vec{\mathbf{x}}_i$. The second is the interactions among grids, such as social flow and geospatial distance, denoted by $\vec{\mathbf{f}}ij$.

We use Conditional Random Field (CRF) shown in Figure A.1 to model the dependency of nodal features. The learning goal is to estimate the conditional probability of y given $\vec{\mathbf{x}}$ and $\vec{\mathbf{f}}$

$$P(y|\vec{\mathbf{x}}, \vec{\mathbf{f}}) \tag{A.1}$$

In the CRF model, we factorize the probability distribution of y to a series of potential functions ψ on the clique.

$$P(Y) = \frac{1}{Z} \prod_{c \in C} \psi(c) \tag{A.2}$$

Use C_1 to denote the set of cliques of size 1 with the form $\langle y_i \rangle$, and C_2 to denote size-2 clique. We define the potential function as follows:

$$\psi_{C_1} = exp(-|y_i - \vec{\alpha}^T \cdot \vec{\mathbf{x}}_i|) \forall i \in [1, n], g_i \in C_1, \tag{A.3}$$

$$\psi_{C_2} = exp(-|y_i - y_j - \vec{\mathbf{w}}^T \cdot \vec{\mathbf{f}}_{i,j}|) \forall i, j \in [1, n], g_i, g_j \in C_2, \tag{A.4}$$

where $\vec{\alpha}$ and $\vec{\mathbf{w}}$ are all positive coefficients.

The distribution of Y is given by

$$P(Y) = \frac{1}{Z} \left[\prod_{i=1}^{n} \psi_{C_1}(y_i) \times \prod_{i=1}^{n} \prod_{j=i}^{n} \psi_{C_2}(y_i, y_j) \right] \tag{A.5}$$

$$P(Y) = \frac{1}{Z} exp \left(- \sum_{i=1}^{n} |y_i - \vec{\alpha}^T \cdot \vec{\mathbf{x}}_i| - \sum_{i=1}^{n} \sum_{j=i}^{n} |y_i - y_j - \vec{\mathbf{w}}^T \cdot \vec{\mathbf{f}}_{i,j}| \right) \tag{A.6}$$

A large value of potential function ψ implies the high probability of $P(Y)$. The goal is to find a set of Y maximizing $P(Y)$.

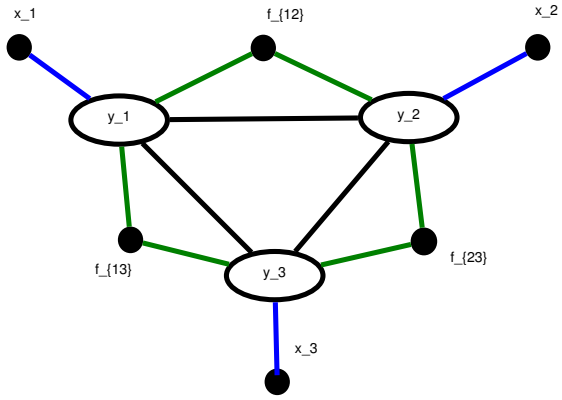


Figure A.1. The CRF model of the crime rate y_i for each grid g_i.

A.2 Inference

A.2.1 Estimate CRF Parameters

We solve the Equation (A.6) by minimizing the negative log-likelihood function

$$\min_{\vec{\alpha},\vec{\mathbf{w}}} -\log P(Y|\vec{\alpha},\vec{\mathbf{w}}) = \min_{\vec{\alpha},\vec{\mathbf{w}}} \Bigg[\log Z + \sum_{i=1}^{n} |y_i - \vec{\alpha}^T \cdot \vec{\mathbf{x}}_i| + \sum_{i=1}^{n}\sum_{j=i}^{n} |y_i - y_j - \vec{\mathbf{w}}^T \cdot \vec{\mathbf{f}}_{i,j}|\Bigg] \quad \text{(A.7)}$$

Use matrix form

$$\min_{\vec{\alpha},\vec{\mathbf{w}}} ||X\vec{\alpha} - \vec{\mathbf{y}}||_1 + ||F \cdot \vec{\mathbf{w}} - \vec{\mathbf{y}_p}||_1,$$

where X is demographics matrix with n rows, $\vec{\mathbf{y}}$ is the n-dimension crime rate vector, F is the pairwise features matrix with $(n^2+n)/2$ rows, and $\vec{\mathbf{y}_p}$ is the $(n^2+n)/2$-dimension pairwise crime rate difference vector $\{y_i - y_j\}$.

We can minimize separably.

Minimize $\vec{\alpha}$

$$\min_{\vec{\alpha}} ||X\vec{\alpha} - \vec{\mathbf{y}}||_1$$

Take $X\vec{\alpha} - \vec{\mathbf{y}} = \vec{\mathbf{z}}$, and use ADMM.

$$\min_{\vec{\mathbf{z}},\vec{\alpha}} ||\vec{\mathbf{z}}||_1 + \rho/2||\vec{\mathbf{z}} - X\vec{\alpha} + \vec{\mathbf{y}}||_2^2, \quad s.t. \quad \vec{\mathbf{z}} - X\vec{\alpha} + \vec{\mathbf{y}} = 0 \quad \text{(A.8)}$$

$$L(\vec{\theta_1}, \vec{\mathbf{z}}, \vec{\alpha}) = \max_{\vec{\theta_1}} \min_{\vec{\mathbf{z}},\vec{\alpha}} ||\vec{\mathbf{z}}||_1 + \rho/2||\vec{\mathbf{z}} - X\vec{\alpha} + \vec{\mathbf{y}}||_2^2 + \vec{\theta_1}(\vec{\mathbf{z}} - X\vec{\alpha} + \vec{\mathbf{y}}) \quad \text{(A.9)}$$

$\vec{\alpha}$ **update**

$$\vec{\alpha}^{k+1} \leftarrow \arg\min_{\vec{\alpha}} \rho/2||\vec{\mathbf{z}}^{k+1} - X\vec{\alpha} + \vec{\mathbf{y}} + \vec{\theta_1}^{k+1}||_2^2$$

Take derivative we have

$$\frac{\partial}{\partial \vec{\alpha}} = \rho X^T(X\vec{\alpha} - \vec{\mathbf{z}}^{k+1} - \vec{\mathbf{y}} - \vec{\theta}_1^{k+1})$$

Make it 0, $\vec{\alpha}^{k+1} = (X^TX)^{-1}X^T(\vec{\mathbf{z}}^{k+1} + \vec{\mathbf{y}} + \vec{\theta}_1^{k+1})$.

$\vec{\mathbf{z}}$ update

$$\vec{\mathbf{z}}^{k+1} \leftarrow \arg\min_{\vec{\mathbf{z}}} ||\vec{\mathbf{z}}||_1 + \rho/2||\vec{\mathbf{z}} - X\vec{\alpha}^{k+1} + \vec{\mathbf{y}} + \vec{\theta_1}^{k+1}||_2^2$$

So, $\vec{\mathbf{z}}^{k+1} = S_{1/\rho}(X\vec{\alpha}^{k+1} - \vec{\mathbf{y}} - \vec{\theta}_1^{k+1})$.

$\vec{\theta}_1$ update

$$\vec{\theta}_1^{k+1} = \vec{\theta}_1^k + \vec{\mathbf{z}}^{k+1} - X\vec{\alpha}^{k+1} + \vec{\mathbf{y}}$$

Minimize $\vec{\mathbf{w}}$

$$\min_{\vec{\mathbf{w}}} ||F \cdot \vec{\mathbf{w}} - \vec{\mathbf{y}_p}||_1$$

It has exactly the same form as previously. Therefore,

$$L(\vec{\theta_2}, \vec{\mathbf{z}'}, \vec{\mathbf{w}}) = \max_{\vec{\theta}_2} \min_{\vec{\mathbf{z}'}, \vec{\mathbf{w}}} ||\vec{\mathbf{z}'}||_1 + \rho/2||\vec{\mathbf{z}'} - F\vec{\mathbf{w}} + \vec{\mathbf{y}_p}||_2^2 + \vec{\theta_2}(\vec{\mathbf{z}'} - F\vec{\mathbf{w}} + \vec{\mathbf{y}_p}) \quad \text{(A.10)}$$

$\vec{\mathbf{w}}$ update

$$\vec{\mathbf{w}}^{k+1} = (F^TF)^{-1}F^T(\vec{\mathbf{z}'}^{k+1} + \vec{\mathbf{y}_p} + \vec{\theta}_2^{k+1})$$

$\vec{\mathbf{z}'}$ update

$$\vec{\mathbf{z}'}^{k+1} = S_{1/\rho}(F\vec{\mathbf{w}}^{k+1} - \vec{\mathbf{y}_p} - \vec{\theta}_2^{k+1})$$

$\vec{\theta}_1$ update

$$\vec{\theta}_2^{k+1} = \vec{\theta}_2^k + \vec{\mathbf{z}'}^{k+1} - F\vec{\mathbf{w}}^{k+1} + \vec{\mathbf{y}_p}$$

A.2.2 Infer New y_i

To infer new y_i, we want to maximize the following probability

$$P(y_i|\vec{\mathbf{x}}_i, F, \vec{\mathbf{y}}, \vec{\alpha}, \vec{\mathbf{w}}),$$

where $\vec{\alpha}$ and $\vec{\mathbf{w}}$ are estimated using previous section, $\vec{\mathbf{y}}$ denote the crime rates of other observed geographical units, and F is the pairwise feature matrix.

Take negative log of the probability, we have

$$\begin{aligned} &\min_{y_i} \left[\log Z + |y_i - \vec{\alpha}^T \cdot \vec{\mathbf{x}}_i| + \sum_{j \neq i}^{n} |y_i - y_j - \vec{\mathbf{w}}^T \cdot \vec{\mathbf{f}}_{i,j}| \right] \\ &\min_{y_i} \left[|y_i - \vec{\alpha}^T \cdot \vec{\mathbf{x}}_i| + \sum_{j \neq i}^{n} |y_i - y_j - \vec{\mathbf{w}}^T \cdot \vec{\mathbf{f}}_{i,j}| \right] \end{aligned} \tag{A.11}$$

In Equation A.11, we have $n+1$ ℓ_1-norm terms, which have exactly the same form $|y_i - b_j|$. The objective becomes

$$\min_{y_i} \sum_{j}^{n} |y_i - b_j|, \tag{A.12}$$

where $b_1 = \vec{\alpha}^T \cdot \vec{\mathbf{x}}_i$ and $b_j = y_{j-1} + \vec{\mathbf{w}}^T \cdot \vec{\mathbf{f}}_{i,j-1}$ for $j > 1$.

Notice that the optimal solution must take value on $\{b_j\}$. We solve this by sorting all the b_j and then calculate the result segment by segment.

Bibliography

[1] Zheng, Y., L. Capra, O. Wolfson, and H. Yang (2014) "Urban computing: concepts, methodologies, and applications," *ACM TIST*, **5**(3), p. 38.

[2] Baum, K. (2005) *Juvenile victimization and offending, 1993-2003*, US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

[3] Finkelhor, D. (2008) *Childhood victimization: violence, crime, and abuse in the lives of young people: violence, crime, and abuse in the lives of young people*, Oxford University Press, USA.

[4] for Disease Control, N. C. and P. (CDC) (2015) "Leading Causes of Nonfatal Injury, United States 2001 - 2013." *Injury Prevention and Control: data and statistics*.

[5] Graif, C. (2015) "Toward a Geographically Extended Perspective of Neighborhood Effects on Children's Victimization," *American Society of Criminology Annual Meeting*.

[6] Graif, C. and R. J. Sampson (2009) "Spatial heterogeneity in the effects of immigration and diversity on neighborhood homicide rates," *Homicide Studies*.

[7] Anselin, L. (2002) "Under the hood: issues in the specification and interpretation of spatial regression models," *Agricultural economics*, **27**(3), pp. 247–267.

[8] Graif, C., A. S. Gladfelter, and S. A. Matthews (2014) "Urban poverty and neighborhood effects on crime: Incorporating spatial and network perspectives," *Sociology Compass*, **8**(9), pp. 1140–1155.

[9] Mohler, G. O., M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita (2012) "Self-exciting point process modeling of crime," *Journal of the American Statistical Association*.

[10] Wang, T., C. Rudin, D. Wagner, and R. Sevieri (2013) "Learning to detect patterns of crime," in *Machine Learning and Knowledge Discovery in Databases*, Springer.

[11] Ehrlich, I. (1975) "On the relation between education and crime," in *Education, income, and human behavior*, NBER, pp. 313–338.

[12] Braithwaite, J. (1989) *Crime, shame and reintegration*, Cambridge University Press.

[13] Patterson, E. B. (1991) "Poverty, income inequality, and community crime rates," *Criminology*, **29**(4), pp. 755–776.

[14] Freeman, R. B. (1999) "The economics of crime," *Handbook of labor economics*, **3**, pp. 3529–3571.

[15] Wang, X., M. S. Gerber, and D. E. Brown (2012) "Automatic crime prediction using events extracted from twitter posts," in *Social Computing, Behavioral-Cultural Modeling and Prediction*, Springer, pp. 231–238.

[16] Gerber, M. S. (2014) "Predicting crime using Twitter and kernel density estimation," *Decision Support Systems*, **61**.

[17] Traunmueller, M., G. Quattrone, and L. Capra (2014) "Mining mobile phone data to investigate urban crime theories at scale," in *Social Informatics*, Springer, pp. 396–411.

[18] Bogomolov, A., B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland (2014) "Once upon a crime: towards crime prediction from demographics and mobile data," in *Proceedings of the 16th international conference on multimodal interaction*, ACM, pp. 427–434.

[19] Ratcliffe, J. H. (2006) "A temporal constraint theory to explain opportunity-based spatial offending patterns," *Journal of Research in Crime and Delinquency*, **43**(3), pp. 261–291.

[20] Toole, J. L., N. Eagle, and J. B. Plotkin (2011) "Spatiotemporal correlations in criminal offense records," *ACM Transactions on Intelligent Systems and Technology (TIST)*, **2**(4), p. 38.

[21] Short, M. B., M. R. D'ORSOGNA, V. B. Pasour, G. E. Tita, P. J. Brantingham, A. L. Bertozzi, and L. B. Chayes (2008) "A statistical model of criminal behavior," *Mathematical Models and Methods in Applied Sciences*, **18**(supp01), pp. 1249–1267.

[22] Chainey, S., L. Tompson, and S. Uhlig (2008) "The utility of hotspot mapping for predicting spatial patterns of crime," *Security Journal*, **21**(1), pp. 4–28.

[23] Eck, J., S. Chainey, J. Cameron, and R. Wilson (2005) "Mapping crime: Understanding hotspots," .

[24] Nakaya, T. and K. Yano (2010) "Visualising Crime Clusters in a Space-time Cube: An Exploratory Data-analysis Approach Using Space-time Kernel Density Estimation and Scan Statistics," *Transactions in GIS*, **14**(3), pp. 223–239.

[25] Buczak, A. L. and C. M. Gifford (2010) "Fuzzy association rule mining for community crime pattern discovery," in *ACM SIGKDD Workshop on Intelligence and Security Informatics*, ACM, p. 2.

[26] Wikipedia (2015), "Community areas in Chicago — Wikipedia, The Free Encyclopedia," .
URL `https://en.wikipedia.org/w/index.php?title=Community_areas_in_Chicago&oldid=696795849`

[27] (2015), "City of Chicago data portal," https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2.

[28] Hsieh, C.-C. and M. D. Pugh (1993) "Poverty, income inequality, and violent crime: a meta-analysis of recent aggregate data studies," *Criminal Justice Review*, **18**(2), pp. 182–202.

[29] Wolfe, M. K. and J. Mennis (2012) "Does vegetation encourage or suppress urban crime? Evidence from Philadelphia, PA," *Landscape and Urban Planning*, **108**(2), pp. 112–122.

[30] Sahbaz, O. and B. Hillier (2007) "The story of the crime: functional, temporal and spatial tendencies in street robbery," in *Proc of 6th International Space Syntax Symposium, Istanbul*, pp. 4–14.

[31] Jacobs, J. (1961) *The death and life of great American cities*, Vintage.

[32] "United States Census Bureau," http://www.census.gov.

[33] Sampson, R. J., S. W. Raudenbush, and F. Earls (1997) "Neighborhoods and violent crime: A multilevel study of collective efficacy," *Science*, **277**(5328), pp. 918–924.

[34] "Foursquare Venues Service," https://developer.foursquare.com/overview/venues.html.

[35] Gorman, D. M., P. W. Speer, P. J. Gruenewald, and E. W. Labouvie (2001) "Spatial dynamics of alcohol availability, neighborhood structure and violent crime." *Journal of studies on alcohol*, **62**(5), pp. 628–636.

[36] Burnell, J. D. (1988) "Crime and racial composition in contiguous communities as negative externalities: prejudiced household's evaluation of crime rate and segregation nearby reduces housing values and tax revenues," *American Journal of Economics and Sociology*, **47**(2), pp. 177–193.

[37] Anselin, L., J. Cohen, D. Cook, W. Gorr, and G. Tita (2000) "Spatial analyses of crime," *Criminal justice*, **4**(2), pp. 213–262.

[38] Morenoff, J. D. and R. J. Sampson (1997) "Violent crime and the spatial dynamics of neighborhood transition: Chicago, 1970–1990," *Social forces*, **76**(1), pp. 31–64.

[39] Gardner, W., E. P. Mulvey, and E. C. Shaw (1995) "Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models." *Psychological bulletin*, **118**(3), p. 392.

[40] Lambert, D. (1992) "Zero-inflated Poisson regression, with an application to defects in manufacturing," *Technometrics*, **34**(1), pp. 1–14.

[41] Osgood, D. W. (2000) "Poisson-based regression analysis of aggregate crime rates," *Journal of quantitative criminology*, **16**(1), pp. 21–43.

[42] Fotheringham, A. S., C. Brunsdon, and M. Charlton (2003) *Geographically weighted regression*, John Wiley & Sons, Limited.

[43] Wikipedia (2016), "Simpson's paradox — Wikipedia, The Free Encyclopedia," [Online; accessed 3-May-2016].
URL `https://en.wikipedia.org/w/index.php?title=Simpson%27s_paradox&oldid=717258021`

Vita

Hongjian Wang

I am a third year PhD students at College of IST, PSU. My advisor is Prof. Zhenhui (Jessie) Li. My research Interest is data mining, especially mining the spatiotemporal data.

Publications

- **Hongjian Wang**, Zhenhui Li, Daniel Kifer, Corina Graif. Crime Rate Inference with Big Data. ACM SIGKDD, 2016. (Under review)
- **Hongjian Wang**, Zhenhui Li, Yu-Hsuan Kuo, Daniel Kifer. A Simple Baseline for Travel Time Estimation using Large-Scale Trip Data. ACM SIGKDD, 2016. (Under review)
- **Hongjian Wang**, Zhenhui Li, Wang-Chien Lee. PGT: Measuring Mobility Relationship Using Personal, Global and Temporal Factors. Proc. ICDM, 2014.
- **Hongjian Wang**, Yamin Zhu, Qian Zhang. Compressive Sensing based Monitoring with Vehicular Networks. IEEE Proc. of INFOCOM, 2013
- Fei Wu, **Hongjian Wang**, Zhenhui Li. Interpreting Traffic Dynamics using Ubiquitous Urban Data. ACM Ubicomp, 2016. (Under review)
- Fei Wu, Zhenhui Li, Wang-Chien Lee, **Hongjian Wang**, Zhuojie Huang. Semantic Annotation of Mobility Data using Social Media, Proc. WWW, 2015.