International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:6, No:1, 2012

Geographic Profiling Based on Multi-point Centrography with K-means Clustering

Jiaji Zhou, Le Liang, and Long Chen

Abstract—Geographic Profiling has successfully assisted investigations for serial crimes. Considering the multi-cluster feature of serial criminal spots, we propose a Multi-point Centrography model as a natural extension of Single-point Centrography for geographic profiling. K-means clustering is first performed on the data samples and then Single-point Centrography is adopted to derive a probability distribution on each cluster. Finally, a weighted combinations of each distribution is formed to make next-crime spot prediction. Experimental study on real cases demonstrates the effectiveness of our proposed model.

Keywords—Geographic profiling, Centrography model, K-means algorithm

I. Introduction

GEOGRAPHIC profiling has been proven to be an effective method to assist in investigation of serial crimes in the past decades [1]. The essential problem of traditional geographic profiling approach is to estimate the serial offender's residential place from previous known crime sites and predict the next-crime spot. The area surrounding the criminal's residential place and the next-crime spot deserve more police attention as it is a potentially dangerous region. The estimation of this residential place (anchor point), has been extensively studied in the current literature. Traditional methods tend to employ a probability distance strategy by choosing an appropriate distance metric and a decay function to model the offender's crime behavior and then calculating a hit score [1], [2]. Different from the traditional approach, [3] proposes a model based on Bayesian analysis. The Bayesian model incorporates geographic and demographic information and greatly improves the profiling performance.

In this paper, we first introduce the Single-point Centrography model adapted from the traditional Centrography models [1] [2]. It is based on a shifted normal distribution with respect to the anchor point, which is the spatial mean of all criminal spots. Although this approach is proven to have good performance in the prediction of anchor point for some existing serial criminal cases, its predictions for next-crime locations are unsatisfactory. Taking the multi-cluster feature of criminal sites into consideration, we incorporate K-mean clustering [12] [16] and establish a new Multi-point Centrography model. The multi-point centrography model outperforms both the Bayesian and single-point centrogaphy approach is potentially a powerful tool in narrowing down the criminal searching area.

II. Model Formulation

A. Key Terms and Definition

- An **anchor point** can be a the offender's place of residence, place of work, or some other locations important to the offender [8]. However, it can also be a pseudo point like the spatial mean (defined later in the Centrography Model) which is merely of mathematical meaning.
- **Buffer zone** is an area centered around the anchor point within which targets are viewed as less desirable because of perceived risk associated with operating too close to the anchor point [1].
- The **spatial distance** metric can be Euclidean distance, the Manhattan distance, or the shortest street distance following the local road network [7] and the choice of distance metric shall be subject to the actual geographic and topological condition of the studied area. The Manhattan distance is a possible candidate for urban areas where street layouts are largely influenced by the designed rectangular blocks, while the Euclidean distance can be potentially more suitable in expansive suburban areas.

  The Manhattan distance between points $\mathbf{x} = (x^{(1)}, x^{(2)})$ and $\mathbf{y} = (y^{(1)}, y^{(2)})$ in a Cartesian coordinate system is defined as:

  $$d_1(\mathbf{x},\mathbf{y}) = \left|x^{(1)} - y^{(1)}\right| + \left|x^{(2)} - y^{(2)}\right|. \qquad (1)$$

  The Euclidean distance between points $\mathbf{x} = (x^{(1)}, x^{(2)})$ and $\mathbf{y} = (y^{(1)}, y^{(2)})$ in a Cartesian coordinate system is defined as:

  $$d_2(\mathbf{x},\mathbf{y}) = \sqrt{(x^{(1)} - y^{(1)})^2 + (x^{(2)} - y^{(2)})^2}. \qquad (2)$$
- The **hunting area** is defined as a rectangular zone containing all crime locations [1]. We assume that the anchor point of the offender and the next potential crime site are both confined to the hunting area. The area can be acquired by drawing the boundaries of the offenders hunting area from the crime sites. In the Manhattan distance system, borders are determined as follows:

$$\begin{aligned} y_{high} &= y_{max} + (y_{max} - y_{min})/2(C-1) \\ y_{low} &= y_{min} - (y_{max} - y_{min})/2(C-1) \\ x_{high} &= x_{max} + (x_{max} - x_{min})/2(C-1) \\ x_{low} &= x_{min} - (x_{max} - x_{min})/2(C-1) \end{aligned}$$

Jiaji Zhou is with the Chien-Shiung Wu Honors College, Southeast University, Nanjing, 210096 China. e-mail: jiajizhou@seu.edu.cn

Le Liang is with the Chien-Shiung Wu Honors College, Southeast University, Nanjing, 210096 China. e-mail: liangle@seu.edu.cn

Long Chen is with the Department of Computer Science, Columbia University, NY, USA. e-mail: lc2808@columbia.edu

B. Single-point Centrography Model

Centrography has been used in a variety of criminological studies and investigative contexts [1]. An investigative review team helped locate the hometown of the Yorkshire Ripper from the geographic center of the murder sites [9].

We assume that the crime series consist of n linked crimes, and that these have taken place at the locations $\mathbf{x}_1, \mathbf{x}_2$, ..., and $\mathbf{x}_n$. The offender's anchor point is denoted by $\mathbf{z}$. The geographic profiling problem is the problem of estimating the offender's anchor point and predicting the next potential crime location from the known crime locations in the series. The offender commits crimes according to an unknown probability density function (pdf) f, which can be modeled as linear, normal, lognormal, truncated negative exponential and other decay functions [3]. In this single-point centrography approach, we use the normal distribution to model the offender's behavior and obtain the pdf of next crime at a location $\mathbf{y}$ as

$$f(\mathbf{y}) = \frac{a}{2\pi\sigma} exp(-\frac{(d(\mathbf{y}, \mathbf{z}) - d_0)^2}{2\sigma^2}) \tag{3}$$

where d can be any distance metric, d_0 is the average distance between the offender's crime locations and the anchor point, and a is the normalizing constant factor. The standard distance σ is designed to model the deviation of each crime location $\mathbf{x}_i$ to the offender's anchor point $\mathbf{z}$ and can be defined as

$$\sigma = \sqrt{\frac{\sum\limits_{i=1}^{n} d^2(\mathbf{x}_i, \mathbf{z})}{n}}. \tag{4}$$

The key to solving this geographic profiling problem is determining the location of the offender's anchor point from the given n crime locations. Different from traditional maximum likelihood estimation method, we propose a simple but effective approach to approximate the estimation of the anchor point $\mathbf{z}$. We propose the spatial mean as a univariate measurement of the central tendency of a point pattern and use it to represent the actual anchor point, which will be shown accurate in later case verification. The spatial mean $\mathbf{m}$ is defined as a point to minimize the sum of the distances to the various crime locations in a certain case. If we take the Manhattan distance as the distance metric, we have

$$\mathbf{m} = \arg \min_{\mathbf{m} \in \mathcal{R}} \sum_{i=1}^{n} d_1(\mathbf{x}_i, \mathbf{y}) \tag{5}$$

where $\mathcal{R}$ is the hunting area.

With the calculated spatial mean as the estimation of the anchor point $\mathbf{z}$, we may predict the next potential dangerous area according to (3) by choosing the area with high value of $f(\mathbf{y})$. Police force shall be mainly directed to these potential dangerous area while the anchor point can also be another police attraction location.

We take the case of Peter Sutcliffe [10] to verify the accuracy of the determination of the anchor point. Peter Sutcliffe has committed 13 crimes during 1975-10-30 to 1980-11-17. We set up a coordinate system ($592 * 528$ with the original point in the lower-left corner) on the map as shown in Fig. 1. The crime locations and intervals between two crimes are given in the Table. 1.

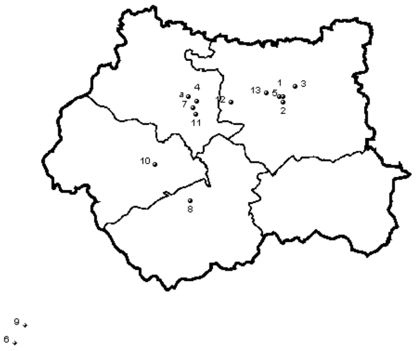

Fig. 1. The location information of the Peter Sutcliffe Serial Crime Case

We apply both Manhattan distance and Euclidean distance metrics in this case and get:

$$\mathbf{m}_1 = (276, 335) \qquad \mathbf{m}_2 = (280, 286)$$
$$\sigma_1 = 188.89 \qquad \sigma_2 = 140.995$$

where $\mathbf{m}_1$ and $\mathbf{m}_2$ denote the spatial mean, σ_1 and σ_2 represent the standard distance of Manhattan and Euclidean distance respectively.

TABLE I
PETER SCUTCLIFFE SERIAL CRIME CASE

Time interval	x coordinate	y coordinate
0	369	342
82	369	336
382	382	352
77	276	337
64	365	342
97	91	95
112	272	330
10	269	230
105	80	77
323	231	269
151	275	322
353	313	335
89	351	346

The actual location of the criminal residential place is $\mathbf{z} = (267, 342)$. It is noticeable that Manhattan distance fits better in this case. We might be more confident to say that the map generally depicts an urban area. Thus, Manhattan distance system is preferred rather than Euclidean distance system due to the specific geographic and topological conditions of the hunting area.

The traditional centrography approach has been successful in many circumstances, especially for those who have a relatively small activity radius. The FBI and ATF analyze serial arson cases by determining the spatial mean of fire sites [11]. They found that 70% of the serial arson set fires within 2 miles of their home. However, the Single-Point centrography model provides only a few piece of information about the possible location of offender's resident and can be seriously distorted by some outliers. While there exist certain levels of correlation between the spatial mean and the anchor point, it is still doubtful to say that the spatial mean is geographically close to the offender's home. In extreme cases, the offender can even be transient. So we need to recognize and specify different patterns first and solve them with different techniques.

C. Multi-point Centrography

Noticed that the criminal spots are usually distributed in a multi-clusters, and also inspired by the *Divide and Conquer* idea, we modify the traditional Centrography Model [1] to enhance robustness by dividing the points (criminal spots) into several parts and apply the Single-point model simultaneously. As for dividing the points, we apply the K-means algorithms [12] [16] to cluster points into a predefined number of partitions.

1) K-means Clustering: K-means clustering partitions n points (criminal spots) into k clusters in which each point belongs to the cluster with the nearest spatial mean (anchor points). This results into a partitioning of the data space into Voronoi cells in the sense of Manhattan distance. Denote $S = \{S_1, S_2, ..., S_k\}$ as the set of partitioned points. We could define the optimization for multiple anchor points as a natural extension of equation (5):

$$S = \arg\min_{S} \sum_{i=1}^{k} \sum_{\mathbf{x_j} \in S_i} d_1(\mathbf{x_j}, \mathbf{z_i}) \tag{6}$$

where $\mathbf{z_i}$ denotes the ith anchor point(spacial mean). Figure 2 is the dual division result of the Peter Sutcliffe case.

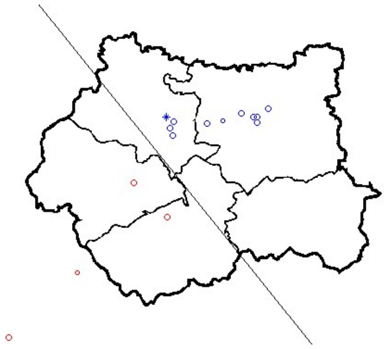

Fig. 2. Dual partition of the Peter-Sutcliffe Case

2) Weighted Probability Distribution: As we have already partitioned the points into several clusters, we can sum up the probability distribution of each one with cluster size as weighting coefficient. Therefore, we have the weighted probability distribution for crime spot as follows:

$$f(y) = \frac{\sum N_i \cdot f_i(y)}{\sum N_i} \tag{7}$$

where N_i denotes the size of cluster S_i.

III. Experimental Results and Analysis

We have conducted experiments on several serial crime dataset, including the previously introduced ”Peter Sutcliffe” [10], ”Jack the Ripper” case [13], ”Zodiac Killer” [14], ”Prostitute Killer” [15]. Comparisons are made with state-of-art algorithms, including Rossmo’s formula [1] and O’Leary’s Bayesian [3] approach.

A. Evaluation Index

If we rank the offending probability density function value evaluated at the finite number of pixel points in the hunting area, we can therefore define $N_\mathbf{x}$ as the number of pixels that has a lower offending probability density than $\mathbf{x}$. Also, denote N as the total number of pixels in the hunting area. Therefore we can define the Crime Rank Index as follows:

$$CI(\mathbf{x}) = \frac{N_\mathbf{x}}{N} \tag{8}$$

This index quantifies the crime risk, and higher $C(\mathbf{x})$ demands greater police force.

B. Case study

The models would predict the kth crime spot by previous k-1 crime spot. In all cases, we suppose that at least five historical crime spots are needed. Thus predication starts from the sixth point and successive points are predicted first, compared with its true location and then been used to predict the followers. Note that when applying Multi-point Centrography, the K-means algorithm may make a single point as a cluster, as the total number of points are relatively small. If we still use equation (4) to estimate σ, it would be zero and therefore makes equation (3) meaningless. Thus, we would assign a small initial value for σ, and the estimation for σ is changed into the following equation:

$$\sigma = \sigma_0 + \sqrt{\frac{\sum_{i=1}^{n} d^2(\mathbf{x}_i, \mathbf{z})}{n}}. \tag{9}$$

For all our studied cases, we set $\sigma_0 = 50$. Table II-V show the result of $CI(\mathbf{x})$ value on various cases. As Rossomo’s model involves too much parameter tuning, we show only its corresponding result for the ”Peter Sutcliffe” case. For the Bayesian model, we assume a uniform distributed prior and use truncated exponential as the likelihood function. Parameters are tuned to have the best performance.

TABLE II
Crime Rank Index of ”Peter Sutcliffe” Case

Number	Single-point	Multi-point	Bayesian(Trunc)	Rossomo’s
6	0.1098	0	0	0
7	0.6660	0.9448	0.9703	0.8769
8	0.9002	0.7911	0.6231	0.6348
9	0.1291	0.2341	0	0
10	0.4242	0.8424	0.7401	0.6694
11	0.5412	0.9667	0.9207	0.8798
12	0.7964	0.9528	0.8711	0.9524
13	0.9822	0.9740	0.7793	0.9923
Average	0.5658	0.7132	0.6131	0.6257

TABLE III
Crime Rank Index of ”Jack the Ripper” Case

Number	Single-point	Multi-point	Bayesian(Trunc)
6	0.7630	0.7933	0.5165
7	0.4412	0.9190	0.8415
Average	0.6021	0.8562	0.6790

TABLE IV
Crime Rank Index of ”Zodiac killer ” Case

Number	Single-point	Multi-point	Bayesian(Trunc)
6	0.8682	0.9672	0.1148
7	0.4399	0.4100	0.4576
Average	0.6541	0.6886	0.2862

TABLE V
Crime Rank Index of ”Prostitute killer” Case

Number	Single-point	Multi-point	Bayesian(Trunc)
6	0.9929	0.9843	0.1885
7	0.9021	0.9679	0.4167
Average	0.9475	0.9761	0.3026

It can be seen Multi-point Centrography outperforms the other compared methods. In addition, as Multi-point Centrography does not involve much parameter tuning as O’Leary’s Bayesian method and Rossomo’s model, it tends to be more robust and easy-to-perform.

IV. Conclusion

In this paper, we have proposed a Multi-point Centrography model as a natural extension of Single-point Centrography for geographic profiling. The model first performs a K-means clustering on the data samples and uses Single-point Centrography to derive a probability distribution on each cluster, then a weighted combinations of each distribution is formed to make next-crime spot prediction. Four real cases are studied and the result demonstrates the effectiveness of our proposed model. However, in our case studies, the number of cluster is set as 2 due to the relatively small number of criminal points. Further research could focus on automatically determine the number of clusters. And another interesting direction is to explore different types of probability functions other than normal distributions in the Single-point Centrography model.

References

[1] D. K. Rossmo, *Geographic Profiling*, CRC Press, 2000.

[2] D. Canter, T. Coffey, M. Huntley, and C. Missen, “ Predicting serial killers’ home base using a decision support system,” *J. Quantitative Criminology*, vol. 16, pp. 457–478, 2000.

[3] M. O’Leary, “The mathematics of geographic profiling,” *J. Investig. Psych. Offender Profil*, vol. 6, pp. 253–265, 2009.

[4] V. Latora, and M. Marchiori, “Efficient behavior of small-world networks,” *Phys. Rev. Lett.*, vol. 87. pp. 198701, 2001.

[5] A. Arenas, A. Diaz-Guilera, J. Kurths, Y. Moreno, and C. Zhou, “Synchronization in complex networks,” *Phys. Rep.*, vol. 469, pp. 93–153, 2008.

[6] M. Rosvall, and A. Trusina, and P. Minnhagen, and K. Sneppen, “Networks and cities: An information perspective,” *Phys, Rev. Lett.*, vol. 94, pp. 28701, 2005.

[7] C. Qian, Y. Wang, J. Cao, J. Lu, and J. Kurths, “Weighted-traffic-network based geographic profiling for serial crime location prediction,” *Europhysics Letters*, vol. 93, pp. 68006, 2011.

[8] M. O’Leary, “Determining the optimal search area for a serial criminal,” *Joint Mathematics Meetings*, Washington DC, USA, 2009.

[9] S. S. Kind, “Navigational ideas and the Yorkshire Ripper investigation,” *Journal of Navigation*, vol. 40, pp. 385–393, 1987.

[10] http://en.wikipedia.org/wiki/Peter_Sutcliffe.

[11] D. J. Icove, H. J. Crisman, “Application of pattern recognition in arson investigation,”*Fire Techonology*, 1975.

[12] Bishop, C. M., *Pattern recognition and machine learning*, Springer New York, 2006.

[13] http://en.wikipedia.org/wiki/Jack_the_Ripper.

[14] http://en.wikipedia.org/wiki/Zodiac_Killer.

[15] http://en.wikipedia.org/wiki/Milwaukee_North_Side_Strangler.

[16] Duda, R.O. and Hart, P.E. and Stork, D.G., *Pattern classification*, wiley New York, 2001.