Models for Offender Target Location Selection with Explicit Dependency Structures

Mike O'Leary with Jeremiah Tucker

Towson University

Quantitative Methods in Defense and National Security 2012
Fairfax, VA
April 30 - May 1, 2012

Acknowledgments

- This work has been supported by the National Institute of Justice, through grant 2009-SQ-B9-K014
- We thank the Baltimore County Police Department for providing the data used in this study
- We thank Phil Canter from the Baltimore County Police Department for his assistance.

The Geographic Profiling Problem

- The geographic profiling problem is to estimate the location of the home base of a serial criminal from the known locations of the elements of the offender's crimes.
  - The home base is also called the anchor point of the offender. It may be the offenders home, the home of a relative, a place of work, or even a favorite bar.
- We have developed a new tool for the geographic profiling problem.
  - It is free for download and use, and is entirely open source.
    - http://pages.towson.edu/moleary/Profiler.html
  - It is still in the prototype stage.

Offender Target Selection

- Suppose that we have a model for offender target selection $P(\mathbf{x} \mid \mathbf{z}, \alpha)$ that gives the probability density that an offender with home base $\mathbf{z}$ selects a target at the location $\mathbf{x}$;
  - Here α is the average distance the offender is willing to travel.
- One reasonable example is a bivariate normal distribution centered at the offender's home base

$$P(\mathbf{x} \mid \mathbf{z}, \alpha) = \frac{1}{4\alpha^2} \exp\left(-\frac{\pi}{4\alpha^2}|\mathbf{x} - \mathbf{z}|^2\right).$$

- This can be augmented with information about the distribution of potential targets, so that

$$P(\mathbf{x} \mid \mathbf{z}, \alpha) \propto G(\mathbf{x}) \exp\left(-\frac{\pi}{4\alpha^2}|\mathbf{x} - \mathbf{z}|^2\right).$$

  for a function $G(\mathbf{x})$ that describes the relative desirability of a target at location $\mathbf{x}$.

Offender Target Selection

- Knowledge of the location of a crime $\mathbf{x}$ can then be parlayed into information about the location of the offender's anchor point $\mathbf{z}$ through Bayes' Theorem.

$$P(\mathbf{z} \,|\, \mathbf{x}) \propto \int_0^\infty P(\mathbf{x} \,|\, \mathbf{z}, \alpha)\pi(\mathbf{z}, \alpha)\, d\alpha$$

  - Here $\pi(\mathbf{z}, \alpha)$ is the prior estimate of the joint distribution of offender home bases $\mathbf{z}$ and average offense distances α.
  - A simple choice is $\pi(\mathbf{z}, \alpha) = H(\mathbf{z})\pi(\alpha)$ where $H(\mathbf{z})$ estimates the distribution of offender home bases, while $\pi(\alpha)$ estimates the distribution of offender average offense distance.

Offender Target Selection

- This approach still works for multiple crime site locations; if the offender has committed crimes at $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ then

$$P(\mathbf{z} \,|\, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) \propto \int_0^\infty P(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \,|\, \mathbf{z}, \alpha)\pi(\mathbf{z}, \alpha)\, d\alpha$$

  for a model of offender target selection $P(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \,|\, \mathbf{z}, \alpha)$.
- This is the approach taken in the existing software prototype.

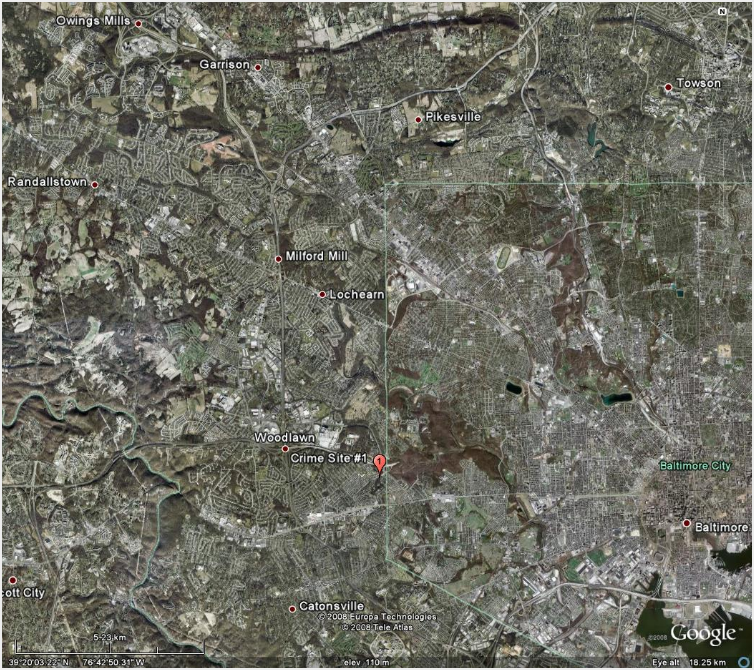
Offender Target Selection

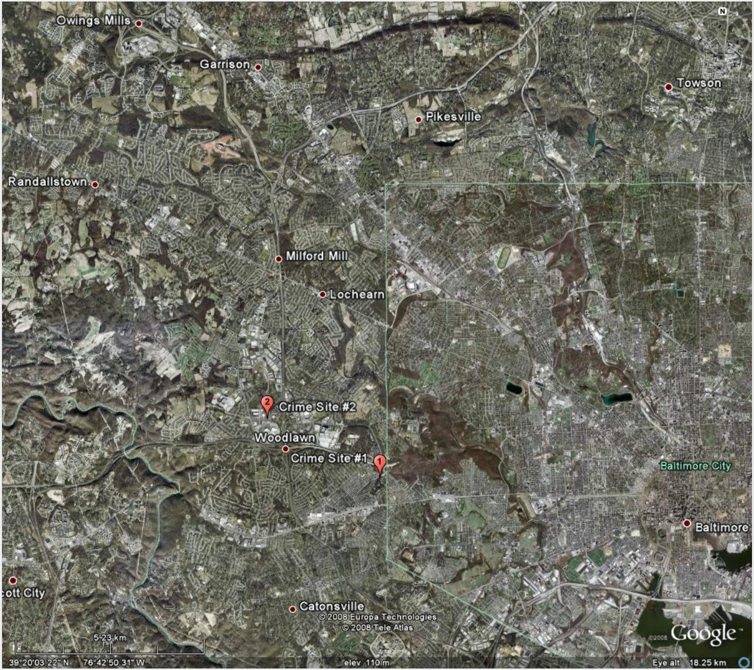
- How can we construct reasonable models for offender target selection for a crime series?
- One approach is to assume that the crime sites are selected independently, so that

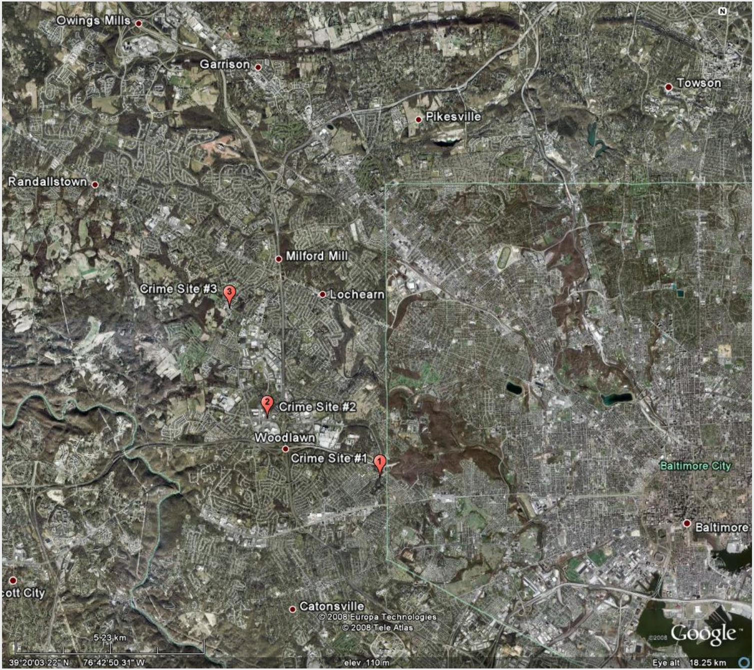
$$\begin{aligned} P(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \,|\, \mathbf{z}, \alpha) &= \prod_{i=1}^{n} P(\mathbf{x}_i \,|\, \mathbf{z}, \alpha) \\ &\propto \prod_{i=1}^{n} G(\mathbf{x}_i) \exp\left(-\frac{\pi}{4\alpha^2}|\mathbf{x}_i - \mathbf{z}|^2\right). \end{aligned}$$

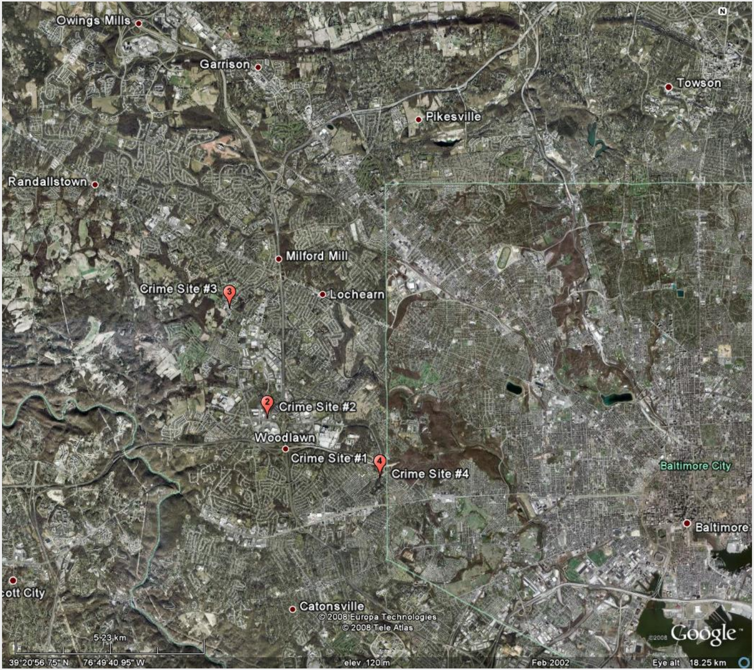
Example- Convenience Store Robberies

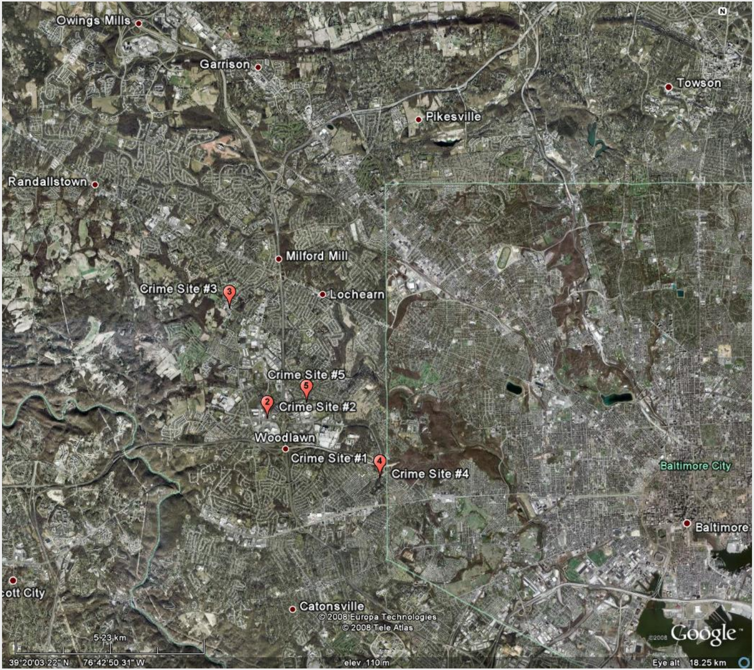
Date	Time	Location		Target
		Latitude	Longitude	
March 8	12:30 pm	-76.71350	39.29850	Speedy Mart
March 19	4:30 pm	-76.74986	39.31342	Exxon
March 21	4:00 pm	-76.76204	39.34100	Exxon
March 27	2:30 pm	-76.71350	39.29850	Speedy Mart
April 15	4:00 pm	-76.73719	39.31742	Citgo
April 28	5:00 pm	-76.71350	39.29850	Speedy Mart

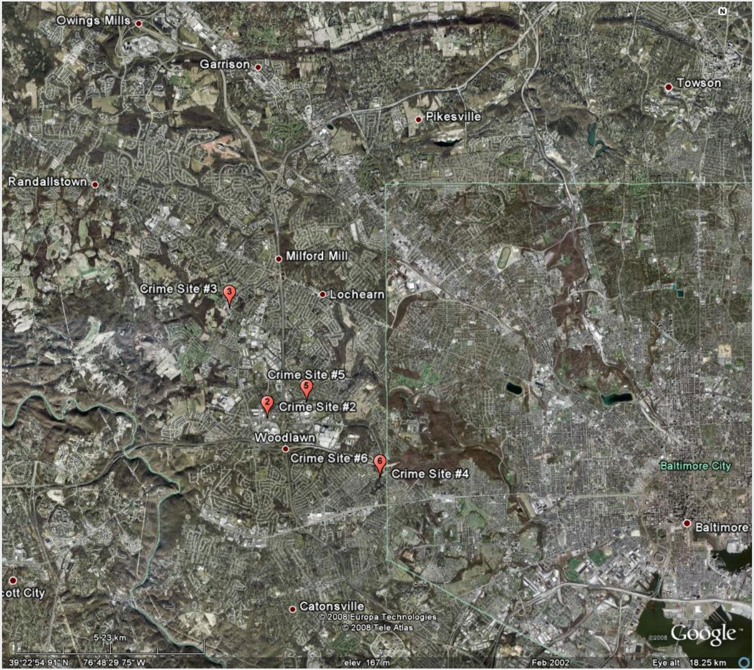

Owings Mills
Garrison
Towson
Pikesville
Randallstown
Milford Mill
Lochearn
Woodlawn
Crime Site #1
Baltimore City
Baltimore
ott City
Catonsville
© 2008 Europa Technologies
© 2008 Tele Atlas
5.23 km
39°20'03.22" N 76°42'50.31" W
elev 110 m
Eye alt 18.25 km
Google


Owings Mills
Garrison
Towson
Pikesville
Randallstown
Milford Mill
Lochearn
Crime Site #2
Woodlawn
Crime Site #1
Baltimore City
Baltimore
ott City
Catonsville
© 2008 Europa Technologies
© 2008 Tele Atlas
5.23 km
39°20'03.22" N 76°42'50.31" W
elev 110 m
Eye alt 18.25 km
Google


Owings Mills
Garrison
Towson
Pikesville
Randallstown
Milford Mill
Crime Site #3
Lochearn
Crime Site #2
Woodlawn
Crime Site #1
Baltimore City
Baltimore
ott City
Catonsville
© 2008 Europa Technologies
© 2008 Tele Atlas
5.23 km
39°20'03.22" N 76°42'50.31" W
elev 110 m
Eye alt 18.25 km
Google


Owings Mills
Garrison
Towson
Pikesville
Randallstown
Milford Mill
Crime Site #3
Lochearn
Crime Site #2
Woodlawn
Crime Site #1
Crime Site #4
Baltimore City
Baltimore
ott City
Catonsville
© 2008 Europa Technologies
© 2008 Tele Atlas
5.23 km
39°20'56.75" N 76°49'40.95" W
elev 120 m
Feb 2002
Eye alt 18.25 km
Google


Owings Mills
Garrison
Towson
Pikesville
Randallstown
Milford Mill
Crime Site #3
Lochearn
Crime Site #5
Crime Site #2
Woodlawn
Crime Site #1
Crime Site #4
Baltimore City
Baltimore
ott City
Catonsville
© 2008 Europa Technologies
© 2008 Tele Atlas
5.23 km
39°20'03.22" N 76°42'50.31" W
elev 110 m
Eye alt 18.25 km
Google



Owings Mills
Garrison
Towson
Pikesville
Randallstown
Milford Mill
Crime Site #3
Lochearn
Crime Site #5
Crime Site #2
Woodlawn
Crime Site #6
Crime Site #4
Baltimore City
Baltimore
ott City
Catonsville
© 2008 Europa Technologies
© 2008 Tele Atlas
5.23 km
39°22'54.91" N 76°48'29.75" W
elev 167 m
Feb 2002
Eye alt 18.25 km
Google

Data

- We have data for residential burglaries in Baltimore County
  - 5863 solved offenses from 1990-2008
  - We have 324 crime series with at least four crimes
    - A series is a set of crimes for which the Age, Sex, Race, DOB and home location of the offender agree.
  - The average number of elements in a series is 8.1, the largest series has 54 elements.
- We have data for non-residential burglaries in Baltimore County
  - 2643 solved offenses from 1990-2008
  - We have 167 crime series with at least three crimes.
  - The average number of elements in a series is 7.87, the largest series has 111 elements.
- We have data for bank robberies in Baltimore County
  - 602 solved offenses from 1993-2009.
  - We have 70 crime series with at least three crimes.
  - The average number of elements in a series is 4.51, the largest series has 15 elements.

Circle Theory

- Canter's Circle hypotheses[1]: Given a series of crimes, construct the circle whose diameter is the segment connecting the two crimes that are farthest apart.
  - If the offender is a *marauder*, then their anchor point will lie in this circle.
  - If the offender is a *commuter*, then their anchor point will lie outside this circle.
    - Note that all of the crimes are not necessarily within the circle.
- For crimes like rape and arson, there is evidence that most offenders are marauders; for crimes like residential burglary the evidence shows a mixture of marauders and commuters.
- This is a binary approach- either someone is a commuter or they are a marauder.
  - This binary approach may not be suitable in many cases.

[1] Canter D. & Larkin, P. (1993). The environmental range of serial rapists. Journal of Environmental Psychology, 13, 63-69.

Commuters & Marauders

- We have created a different way to differentiate between commuters and marauders.
- Suppose that:
  - The crimes are at $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$;
  - The offender's anchor point is $\mathbf{z}$.
- Define

$$\mu_2 = \left[\frac{\min_{\mathbf{y}} \sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{y})^2}{\sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{z})^2} \right]^{1/2} = \sqrt{\frac{\sum_{i=1}^{n} |\mathbf{x}_i - \mathbf{y}_{\text{centroid}}|^2}{\sum_{i=1}^{n} |\mathbf{x}_i - \mathbf{z}|^2}}$$

  where $\mathbf{y}_{\text{centroid}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$ is the centroid of the crime series.
- Note that $0 \leqslant \mu_2 \leqslant 1$.
- Offenders with small μ_2 correspond to μ_2-commuters, while offenders with large μ_2 correspond to μ_2-marauders.

Distance Decay

- Though we have data for the distance from the offenders home to the offense site for a large number of solved crimes, we cannot directly use this information to draw inferences about the behavior of any individual offender.
    - To do so is to commit the *ecological fallacy*.
- There are two sources of variation- the variation within each individual, and the variation between individuals.
    - If all of the individuals behaved in the same fashion, then the aggregate data can be used to draw inference about the (common) underlying behavior.

Distance Decay

- If the only quantity that varies between offenders is the average offense distance, then the resulting scaled distances should exhibit the same behavior regardless of the offender.
  - In particular, this will allow us to aggregate the data across offenders and draw valid inference about the (assumed) universal behavior.
- For each serial offender with crime sites $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ and home $\mathbf{z}$, estimate the average offense distance α by

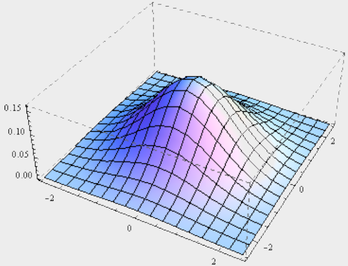
$$\hat{\alpha}_h = \frac{1}{n} \sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{z})$$

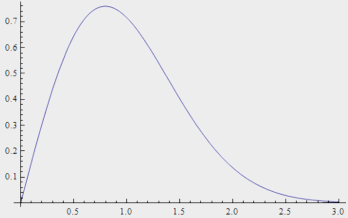
and now consider the set of scaled distances

$$\rho_i = \frac{d(\mathbf{x}_i, \mathbf{z})}{\hat{\alpha}_h}$$

Distance Decay

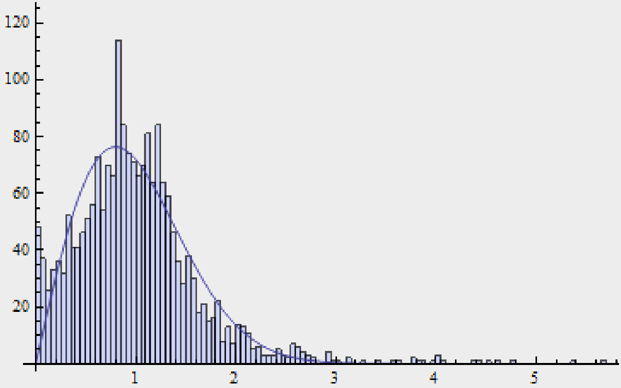
- When considering distance, it is important to realize that it is a *derived* quantity.
  - Offenders do not select a distance- they select a target.
- For example, if the offender selects a target from a two-dimensional normal distribution; then the distribution of distances is a Rayleigh distribution.





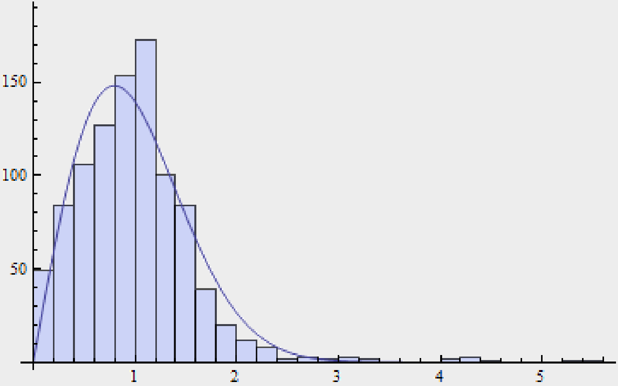
Distance Decay

- Focus our attention only on non-commuters- say $\mu_2 \geqslant 0.25$, and compare the result to a Rayleigh distribution



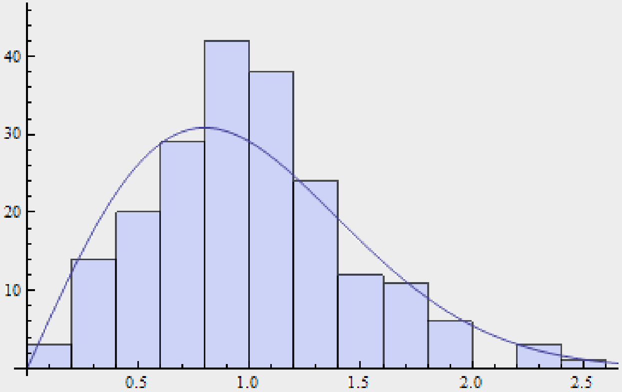
Distance Decay

- The agreement with the Rayleigh distribution does not appear to be happenstance. Here is what occurs for non-residential burglaries with $\mu_2 \geqslant 0.25$



Distance Decay

- Here is what occurs for bank robberies with $\mu_2 \geqslant 0.25$



Distance Decay

- It is possible that these fits are caused by something peculiar to the geography of Baltimore County.
- However, we are not the first to examine scaled distances.
  - Warren, Reboussin, Hazelwood, Cummings, Gibbs, and Trumbetta (1998). *Crime Scene and Distance Correlates of Serial Rape*, Journal of Quantitative Criminology **14** (1998), no. 1, 3559.
  - In that paper, they graphed scaled distances for serial rape:

Distance Decay

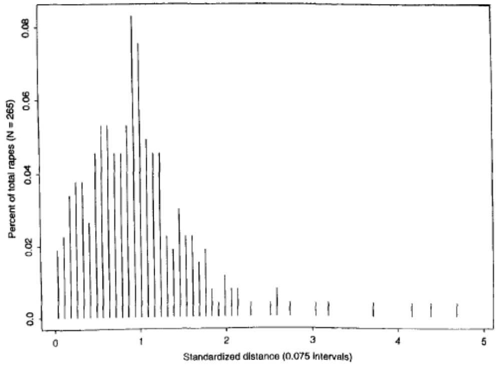
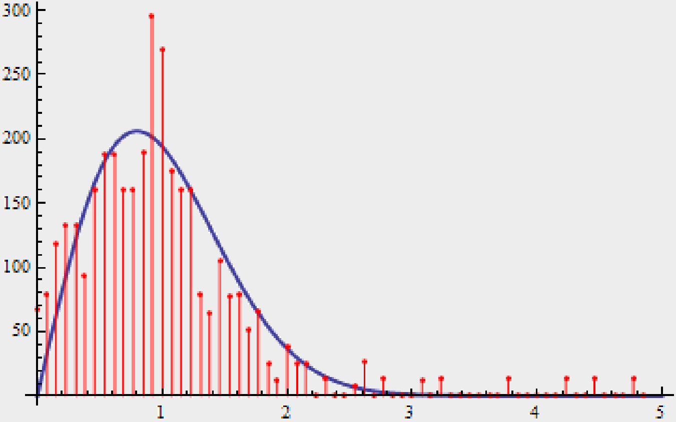


Fig. 2. Proportion of rapes by standardized distance from residence to rape location. Cases with five or more rapes.

two reasons. First, the nonrepresentative nature of the data diminishes the meaningfulness of significance levels. Second, the applied purpose of the paper heightened the need to present the data in a visually clear and practically interpretable form. Distance was found to vary with the demographic characteristics of the offender as well as certain "signature" and "modus

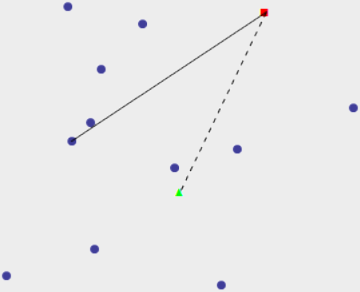
Distance Decay

- Our Rayleigh distribution with mean 1 appears to fit this data as well:



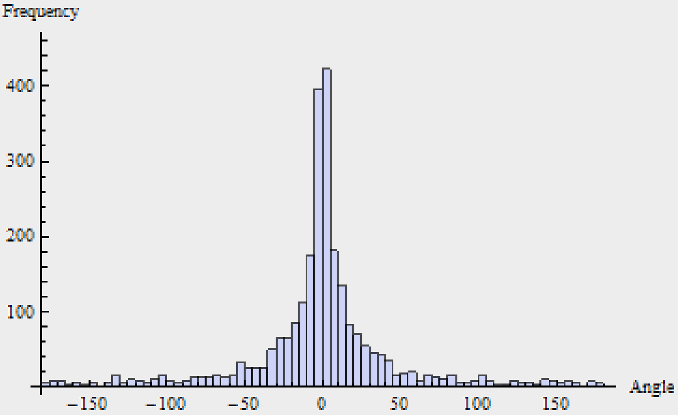
Angular Dependence

- If our idea that the underlying distribution is bivariate normal is correct, then there should be no angular dependence in the results.
- To measure angles, let the blue dots represent crime locations, the red square the anchor point, and the green triangle the centroid of the crime series.
- Then measure the angle between the ray from the anchor point to the crime site and the ray from the anchor point to the centroid.



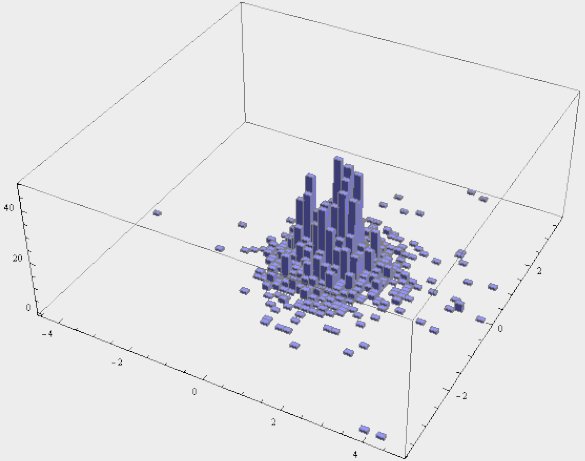
Angular Dependence

- The residential burglary data shows a striking relationship- nearly all of the crime sites lie in the same direction as the centroid.



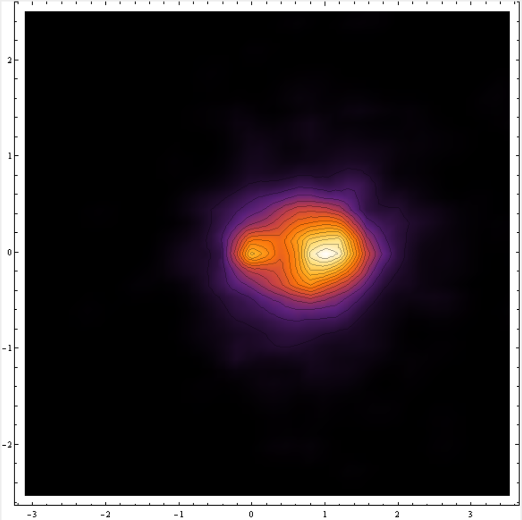
Two-dimensional Distribution

- Plot the histogram of the scaled two dimensional data set; here the offender's home is at the origin, and the centroid of the crime series is at $(x, y) = (1, 0)$.



Two-dimensional Distribution

- Here is another view as a two-dimensional density; note that it is not centered at the origin.



Conclusions

- The bivariate distribution is not bivariate normal, though the distances are Rayleigh.
- It is clear that there are significant correlations between the locations of the different crime site locations.
  - As evidence, we have the fact that the scaled bivariate distribution clusters not around the offender's home, but around the centroid of the crime series.
  - Perhaps offenders exhibit dependency in the selection of crime sites? In particular, perhaps offenders prefer to commit crimes near locations where they have already succesfully offended.

Models with Explicit Dependency Structure

Suppose that an offender with anchor point $\mathbf{z}$ and average offense distance α has committed crimes at $\mathbf{C} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. Consider the following models of offender behavior:

- The normal model:

$$P(\mathbf{x}_{n+1} \,|\, \mathbf{C}, \mathbf{z}, \alpha) = \frac{1}{4\alpha^2} \exp\left(-\frac{\pi}{4\alpha^2}|\mathbf{x}_{n+1} - \mathbf{z}|^2\right)$$

- The near repeat model:

$$P(\mathbf{x}_{n+1} \,|\, \mathbf{C}, \mathbf{z}, \alpha, \gamma) = \frac{\gamma}{n} \sum_{i=1}^{n} \frac{1}{4\epsilon_{\text{rep}}^2} \exp\left(-\frac{\pi}{4\epsilon_{\text{rep}}^2}|\mathbf{x}_{n+1} - \mathbf{x}_i|^2\right) + (1-\gamma)\frac{1}{4\alpha^2} \exp\left(-\frac{\pi}{4\alpha^2}|\mathbf{x} - \mathbf{z}|^2\right)$$

where ϵ_{rep} is fixed and small- say 0.005 mi = 26 ft.

Models with Explicit Dependency Structure

- The general model:

$$P(\mathbf{x}_{n+1} \,|\, \mathbf{C}, \mathbf{z}, \alpha, \gamma, \epsilon) = \frac{\gamma}{n} \sum_{i=1}^{n} \frac{1}{4\epsilon^2} \exp\left(-\frac{\pi}{4\epsilon^2} |\mathbf{x}_{n+1} - \mathbf{x}_i|^2\right) + (1-\gamma) \frac{1}{4\alpha^2} \exp\left(-\frac{\pi}{4\alpha^2} |\mathbf{x} - \mathbf{z}|^2\right)$$

  Here ϵ is now unknown.
- The number γ is the mixture parameter, and satisfies $0 \leqslant \gamma \leqslant 1$.
  - It represents the fraction of the crime series that can be explained by the offender committing crimes near the locations of previous crimes.

Analyzing the Dependency Structure

- To investigate which model is most useful, we employ AIC with the small-sample correction (AICc).
- Given a crime series with n elements and a model with k parameters, we use maximum likelihood to calculate the likelihood L and to estimate the parameters that appear (*e.g.* $\hat{\alpha}, \hat{\gamma}, \dots$).
- The value of $AICc$ is then

$$AICc = -2\ln(L) + 2k\left(\frac{n}{n-k-1}\right)$$

- For model selection, the weight w_i associated to model i is

$$w_i = \frac{\exp(-\frac{1}{2}AICc_i)}{\sum_{j=1}^{n}\exp(-\frac{1}{2}AICc_j)} = \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_{j=1}^{n}\exp(-\frac{1}{2}\Delta_j)}$$

where $\Delta_i = AIC_i - \min_j AIC_j$.

Analyzing the Dependency Structure

- Parameters to be estimated:

Model	Home known	Home unknown
Normal	α	$\alpha, (z_1, z_2)$
Near Repeat	α, γ	$\alpha, \gamma, (z_1, z_2)$
General	α, γ, ϵ	$\alpha, \gamma, \epsilon, (z_1, z_2)$

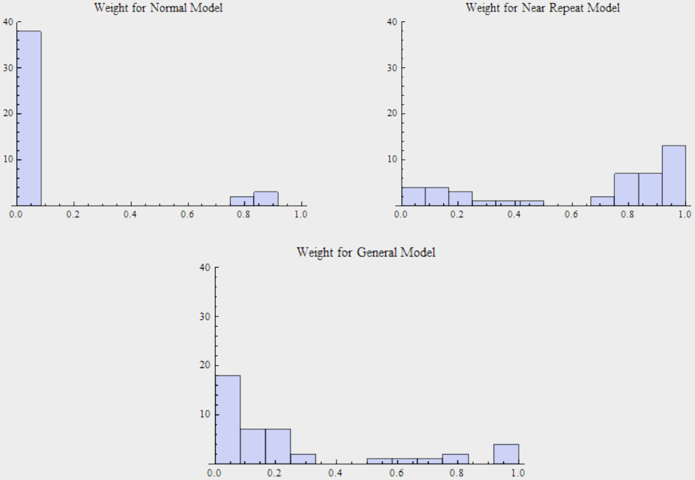
- To even calculate AICc, we need two more data points than parameters, so we will analyze only series with at least seven incidents.

Analyzing the Dependency Structure- Data

- In many crime series, the offender commits multiple crimes in the same location on the same day- *e.g.* by robbing multiple storage lockers or breaking into multiple offices in a building.
  - In this analysis, multiple crimes in the same place on the same day have been consolidated into a single incident.
- Available data:
  - We have 136 residential burglary series with at least seven crimes
  - We have 43 non-residential burglary series with at least seven crimes
  - We have 10 bank robbery series with at least seven crimes.

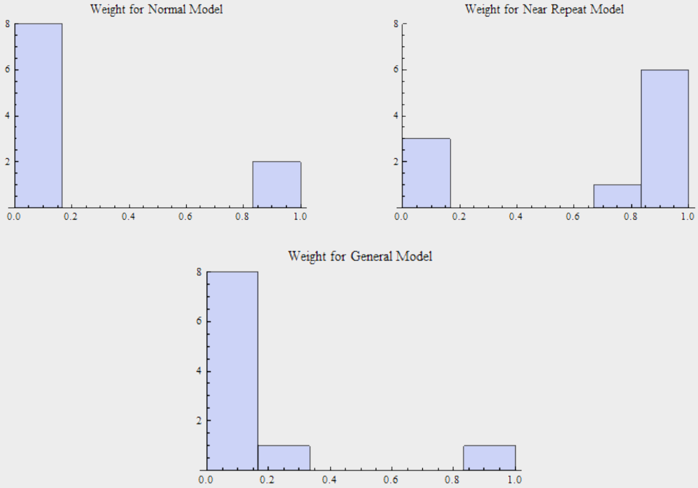
Analyzing the Dependency Structure: Home Known

- For non-residential burglaries where the home is considered known, the normal model is the least supported by the data, and the near repeat model is the most supported.



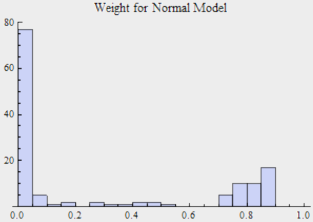
Analyzing the Dependency Structure: Home Known

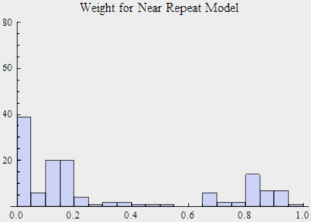
- For bank robberies with a known home, the normal model is least supported and the near repeat model most supported.

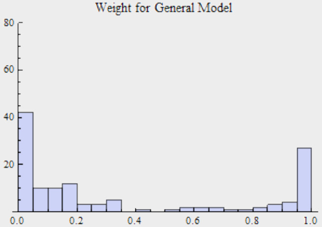


Analyzing the Dependency Structure: Home Known

- For residential burglaries with a known home, the normal model is least supported with evidence in favor of both near repeat and the mixed models.







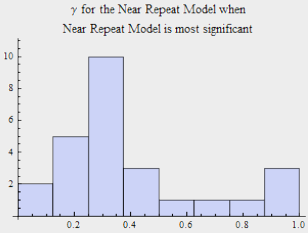
Analyzing the Dependency Structure: Home Known

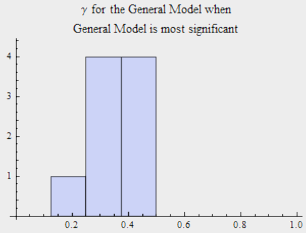
- The most supported model is

Crime Type	Normal	Near Repeat	General
Non-Residential Burglary	5	29	9
Bank Robbery	2	7	1
Residential Burglary	45	41	50

Mixture Parameter: Home Known

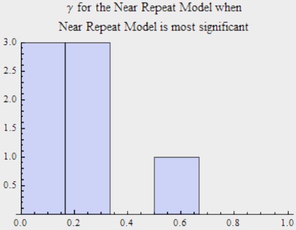
- The distribution of the mixture parameter γ for non-residential burglaries is

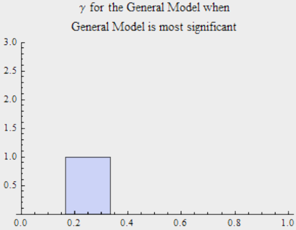




Mixture Parameter: Home Known

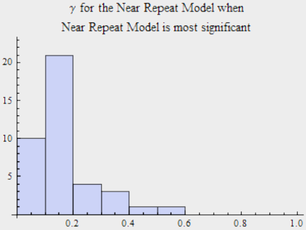
- The distribution of the mixture parameter γ for bank robberies is

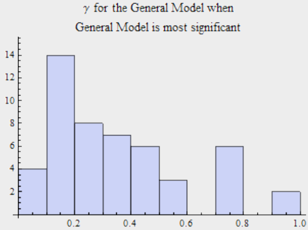




Mixture Parameter: Home Known

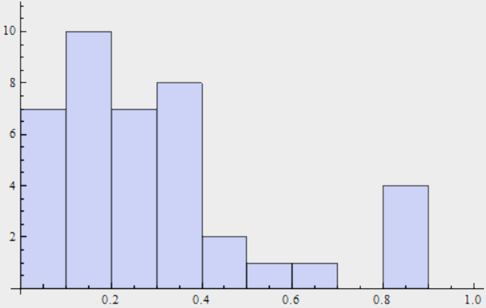
- The distribution of the mixture parameter γ for residential burglaries is





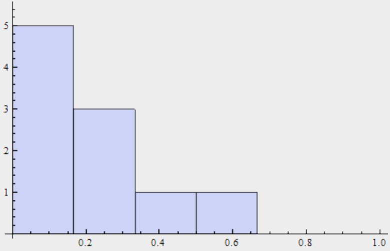
Repeat Analysis: ρ

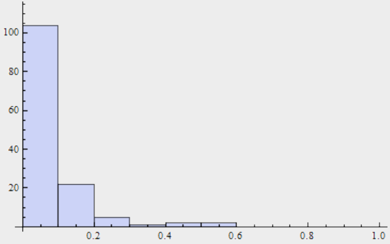
- One possible explanation for the observed behavior is the extent to which offenders return to the same location(s) for subsequent offenses.
- For a crime series, let ρ be the fraction of the crimes that are within the distance $\epsilon_{\text{rep}} = 0.005$ mi. $= 26$ ft. of a previous crime in the series.
- If we plot a histogram of ρ over all non-residential burglary series, we obtain



Repeat Analysis: ρ

- The corresponding graphs for bank robberies (left) and residential burglaries (right) are



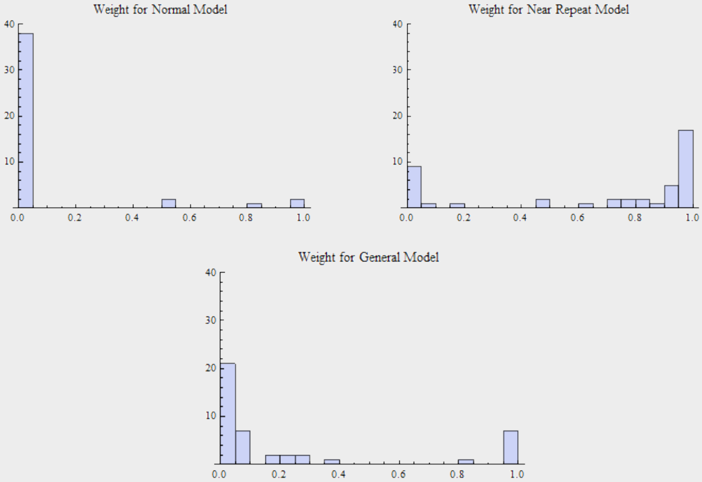


Analyzing the Dependency Structure: Home Unknown

- Since the purpose of the analysis is to aid investigators, it is important to compare the results when the home base is not *a priori* known.

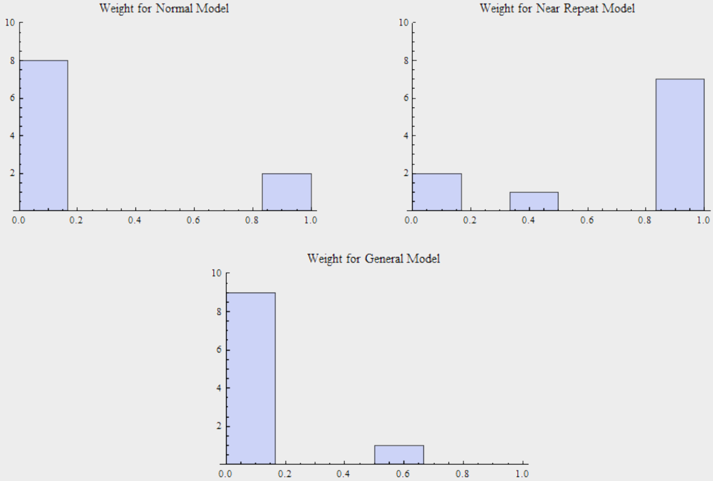
Analyzing the Dependency Structure: Home Unknown

- For non-residential burglaries, with the home unknown, there is support for both the near repeat and the general model.

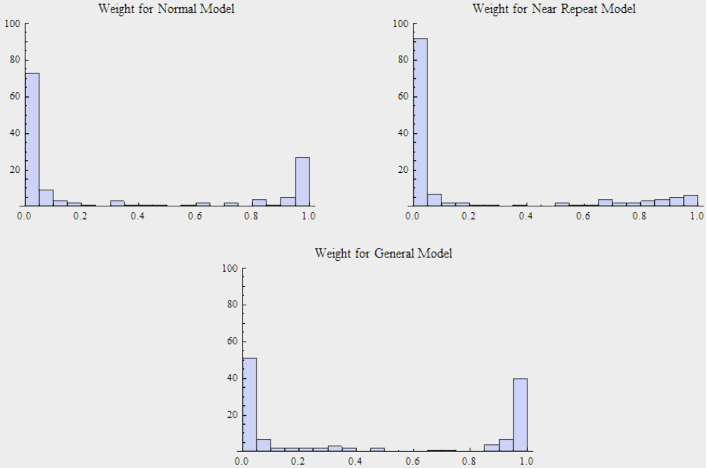


Analyzing the Dependency Structure: Home Unknown

- For bank robberies with the home unknown, the near repeat model is most supported.



- For residential burglaries with the home unknown, the general model is most supported



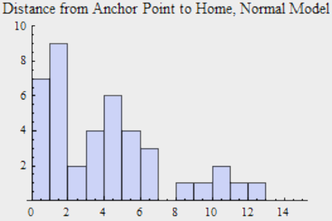
Analyzing the Dependency Structure: Home Unknown

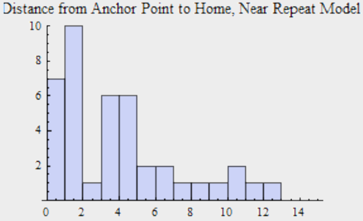
- The most supported model has essentially the same distribution as the case when the home was known.

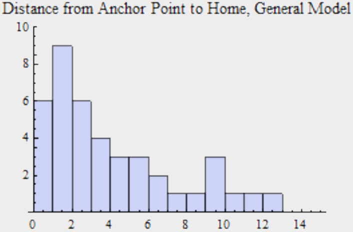
Crime Type	Normal	Near Repeat	General
Non-Residential Burglary	5	30	8
Bank Robbery	2	7	1
Residential Burglary	42	30	64

Implications for Geographic Profiling

- Compare the distance from the actual home to the predicted value of the home location for the different models. For non-residential burglary, we obtain

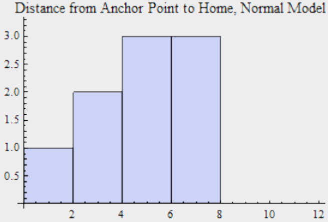


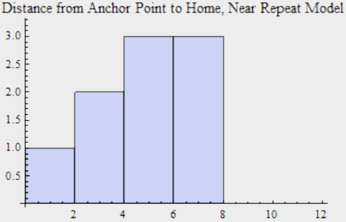


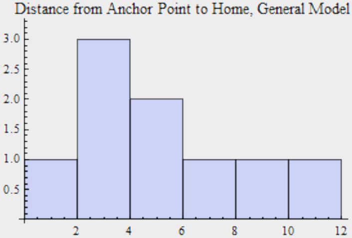


Implications for Geographic Profiling

- The distance from the actual home to the predicted home for bank robberies

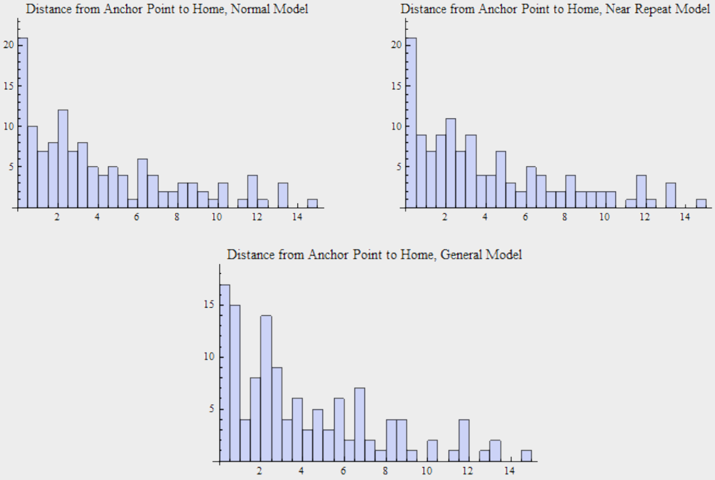






Implications for Geographic Profiling

- The distance from the actual home to the predicted home for residential burglaries



Implications for the Geographic Profiling Problem

- Which model predicts an anchor point closest to the actual anchor point?

Crime Type	Normal	Near Repeat	General
Non-Residential Burglary	17	6	14
Bank Robbery	4	2	1
Residential Burglary	49	24	51

(This excludes the case where there is a tie).

Questions?

Mike O'Leary
Department of Mathematics
Towson University
moleary@towson.edu

`http://pages.towson.edu/moleary/Profiler.html`