New Approaches for the Geographic Profiling Problem

Mike O'Leary

Towson University

The NIJ Conference
Arlington, VA
June 14, 2010

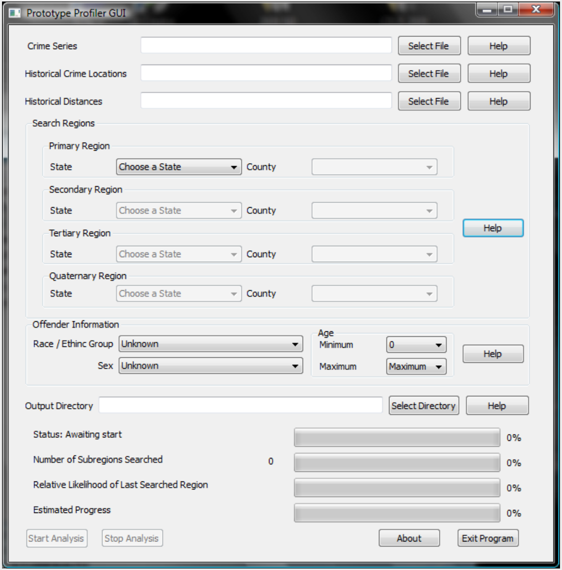
Acknowledgments

- 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

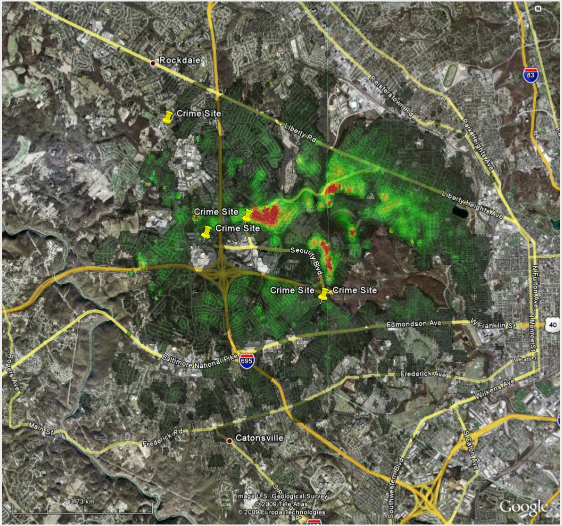
- We have developed a new tool for the geographic profiling problem.
  - It is free for download and use, and is entirely open source.
    - `http://pages.towson.edu/moleary/Profiler.html`
  - It is still in the prototype stage.
  - The tool is designed to be simple and easy to use.
- The tool is able to account for
  - Geographic features that affect the selection of the crime sites,
  - Geographic features that affect the distribution of potential anchor points,
  - Differences in the travel distances of different offenders, and
  - Any available demographic characteristics (race/ethnic group, age, and sex) of the offender.

The Tool


Prototype Profiler GUI
Crime Series
Select File
Help
Historical Crime Locations
Select File
Help
Historical Distances
Select File
Help
Search Regions
Primary Region
State
Choose a State
County
Secondary Region
State
Choose a State
County
Tertiary Region
State
Choose a State
County
Quaternary Region
State
Choose a State
County
Help
Offender Information
Race / Ethinc Group
Unknown
Sex
Unknown
Age
Minimum
0
Maximum
Maximum
Help
Output Directory
Select Directory
Help
Status: Awaiting start
0%
Number of Subregions Searched
0
0%
Relative Likelihood of Last Searched Region
0%
Estimated Progress
0%
Start Analysis
Stop Analysis
About
Exit Program

Sample Results

- When the program runs, it produces an estimate for the offender's anchor point



The Model

- The program begins by assuming that an offender with anchor point at $\mathbf{z}$ commits a crime at the location $\mathbf{x}$ according to the probability density

$$P(\mathbf{x}|\mathbf{z}, \alpha) = D(d(\mathbf{x}, \mathbf{z}), \alpha) \cdot G(\mathbf{x}) \cdot N(\mathbf{z}, \alpha).$$

- Here:
    - $d(\mathbf{x}, \mathbf{z})$ is the distance between $\mathbf{x}$ and $\mathbf{z}$
    - α is the average distance that the offender is willing to travel to offend
    - $D(d(\mathbf{x}, \mathbf{z}), \alpha)$ models the effect of distance decay
    - $G(\mathbf{x})$ represents the relative attractiveness to the offender of a target located at $\mathbf{x}$.
    - $N(\mathbf{z}, \alpha)$ is a normalization factor to ensure that P is a probability density; it is completely determined by the choices of the other factors.
- Given a series of crimes at $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, Bayesian methods let us estimate the distribution $P(\mathbf{z}|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ of $\mathbf{z}$, provided:
    - Crime sites are selected independently
    - We can construct a prior estimate for the distribution of offender anchor points $\mathbf{z}$ and offender average offense distance α.

The Model

- This approach can only be as accurate as the particular choices that were made in the model.
  - If the offender's actual behavior is very different than the model, then the profiling algorithm will be of little value.
- It would be simple to add additional choices of modeling functions to the existing code, and even let the analyst select which model is to be used.
  - How is this better than guessing?
- To improve the accuracy and validity of geographic profiling, we need to better understand its scientific foundations.
  - We need to understand the advantages, disadvantages, limitations and the accuracy of our approach.
  - In particular, we need to study and to better understand how offenders behave.

Data

- We have data for residential burglaries in Baltimore County
  - 5863 solved offenses from 1990-2008
  - Data for each offense includes:
    - Home location, offense location (geocoded)
    - Date & time range of the offense
    - Age, sex, race & DOB of the offender
  - 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 had 54 elements.

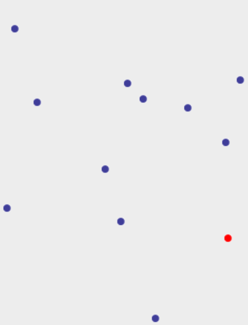
Residential Burglary

- Why study residential burglary?
  - Circle theory works relatively well on crimes like rape and arson, but less well on crimes like residential burglary.
    - Residential burglary has a broader mix of commuter and marauder offenders.
    - If there is a difference between commuter and marauder behavior, this data set may be able to resolve the difference.
  - We also need to account for the distribution of potential targets.
    - Since we are assuming that the anchor point is a residence and since the only potential targets are residences, then we expect at small scale the distribution of targets to be roughly homogeneous.
    - Similarly, at large scale, we also expect the distribution of targets to be roughly homogeneous.
    - Hopefully this will reduce the impact of geography on target selection, and let us focus more closely on the distance decay components.

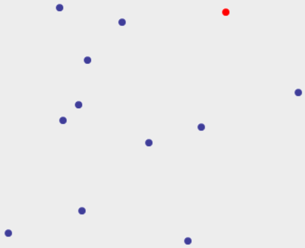
Commuters & Marauders

- Canter's definition of a marauder is an offender whose anchor point lies within the circle whose diameter is formed by the two crimes farthest apart; other offenders are 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.

Which is the Commuter?



or



- Here the crime locations are in blue, and the offender's anchor point is red

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 center of minimum distance is the point $\mathbf{y}_{\text{cmd}}$.
  - The offender's anchor point is $\mathbf{z}_{\text{anchor}}$.
- Define

$$\mu_1 = \frac{\displaystyle\sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{y}_{\text{cmd}})}{\displaystyle\sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{z}_{\text{anchor}})}$$

- Because of the definition of $\mathbf{y}_{\text{cmd}}$, we know that $0 \leqslant \mu_1 \leqslant 1$.
- Offenders with $\mu_1 \leqslant 1/2$ correspond to μ_1-commuters, while offenders with $\mu_1 \geqslant 1/2$ correspond to μ_1-marauders.

Commuters & Marauders

- We have a corresponding measure for the mean center;
- Define

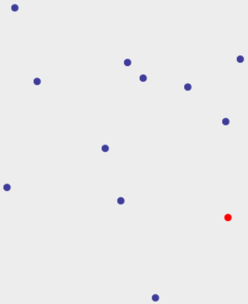
$$\mu_2 = \sqrt{\frac{\sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{y}_{\text{centroid}})^2}{\sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{z}_{\text{anchor}})^2}}$$

- Here $\mathbf{y}_{\text{centroid}}$ is the centroid of the crime series; it minimizes the value of

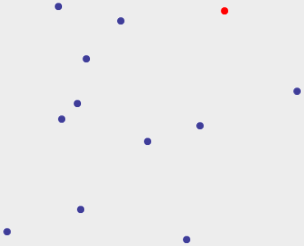
$$\mathbf{p} \mapsto \sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{p})^2$$

  (when d is Euclidean distance) ensuring that $0 \leqslant \mu_2 \leqslant 1$.
- Offenders with $\mu_2 \leqslant 1/2$ correspond to μ_2-commuters, while offenders with $\mu_2 \geqslant 1/2$ correspond to μ_2-marauders.

Which is the Commuter?



or

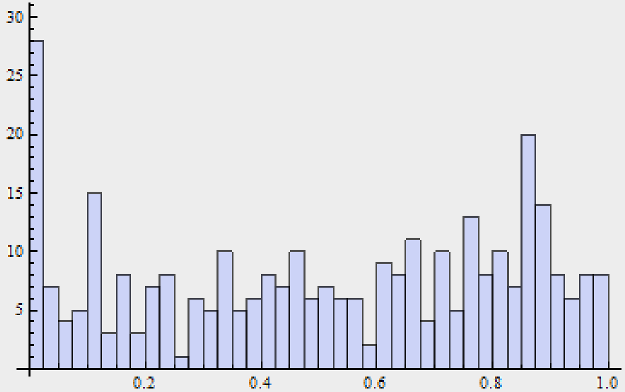


$\mu_2 = 0.58$
(Canter Marauder)

$\mu_2 = 0.56$
(Canter Commuter)

Commuters & Marauders

- What is the distribution of μ_2 in our data?



- There does not appear to be a sharp distinction between commuters and marauders in this data

Dimensional Analysis

- One of the key ideas underlying the development of the quantitative measures for commuters & marauders is that the numbers so developed would be dimensionless.
- Consider

$$\mu_1 = \frac{\sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{y}_{\text{cmd}})}{\sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{z}_{\text{anchor}})} \quad \text{or} \quad \mu_2 = \sqrt{\frac{\sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{y}_{\text{centroid}})^2}{\sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{z}_{\text{anchor}})^2}}$$

Regardless of the units used to measure distance- miles, kilometers, feet or furlongs, the values of μ_1 and μ_2 remain unchanged.

Dimensional Analysis

- 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

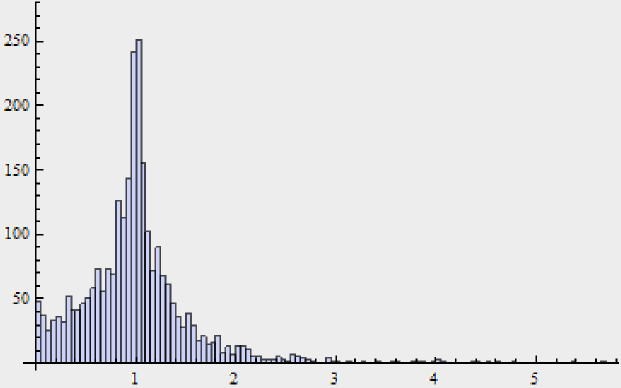
$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{z})$$

and now consider the set of scaled distances

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

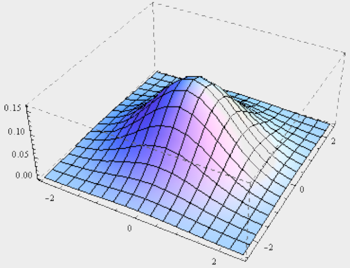
Dimensional Analysis

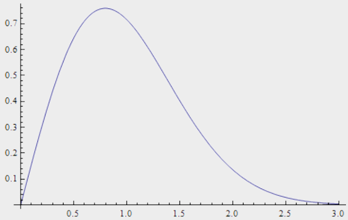
- What do we obtain when we graph not offense distance, but scaled distance?



Distance Decay

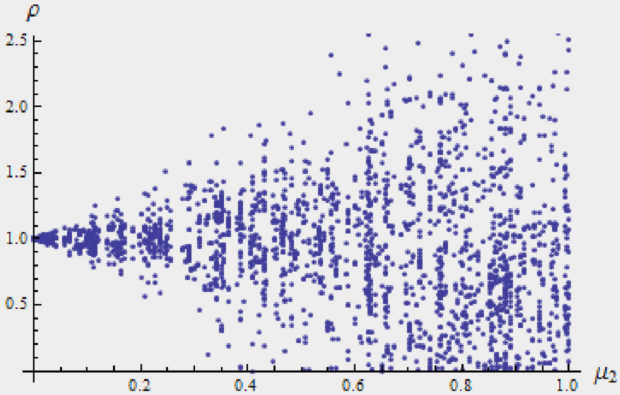
- When considering distance, it is important to realize that it is a *derived* quantity.
  - Offenders do not select a distance- they select a target.
- Suppose that the offender selects a target from a two-dimensional normal distribution; then the distribution of distances is a Rayleigh distribution.





Dimensional Analysis

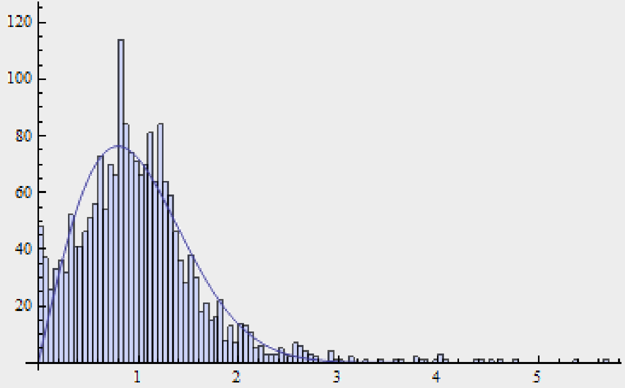
- It is useful to look at the dependence of μ_2 versus ρ



- "Commuters" ($\mu_2 \approx 0$) exhibit very different behavior than "marauders" $(\mu_2 \approx 1)$.
- Focus our attention only on "marauders"- say $\mu_2 > 0.25$.

Dimensional Analysis

- If we assume that each offender chooses targets from a two-dimensional normal distribution with their own average offense distance, then the distribution of scaled distances should follow a Rayleigh distribution with mean 1:



Dimensional Analysis

- It is important to know that no parameters were used to generate this fit.
- If the offender selects a crime site according to a bivariate normal distribution with average offense distance α, (where α may be different for different offenders) then the scaled distances must follow a Rayleigh distribution with average 1.

Dimensional Analysis

- It is possible that this fit is caused by either:
  - Happenstance, or
  - Something peculiar to the data set under investigation
- 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*.
  - In this paper, they graphed scaled distances for serial rape

Dimensional Analysis

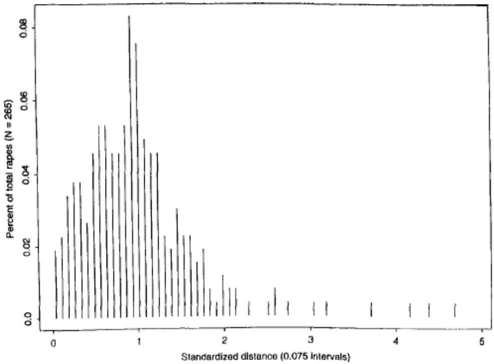
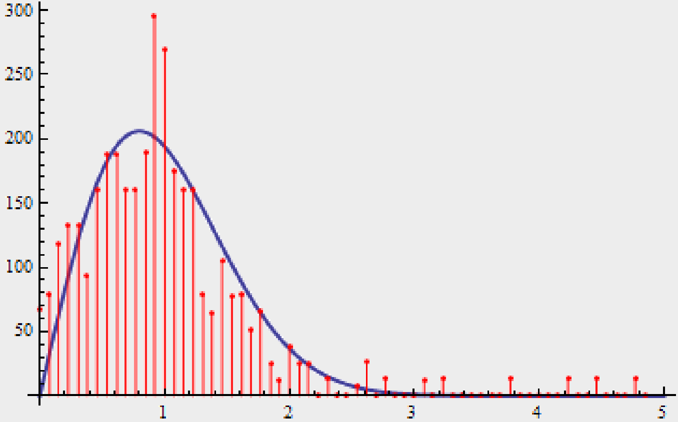


Fig. 2. Proportion of rapes by standardized distance from residence to rape location. Cases with five or more rapes.

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

Dimensional Analysis

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

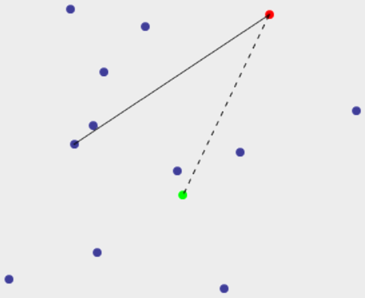


Dimensional Analysis- Caveats

- It is important to note that, though compelling, these graphs do not provide justification that offenders follow a bivariate normal distribution.
    - Agreement is necessary, but not sufficient for this conclusion.
    - There are other two dimensional distributions whose distribution of distances also is Rayleigh.
- We still do not understand the situation yet with commuters.
    - The Warren *et. al.* data is for serial rape, which is known to be well approximated by circle theory- suggesting that this data set may be weighted away from commuters, which our theory does not yet handle.
- Interestingly, the differences between the prediction and the observed data seem to show a similar pattern
    - Observed data exceeds the prediction both near the origin and near $\rho \approx 1$, while it remains below the prediction near the shoulders.
    - This suggests that this may be due to more than random chance. Perhaps a different theoretical structure would better fit the data.

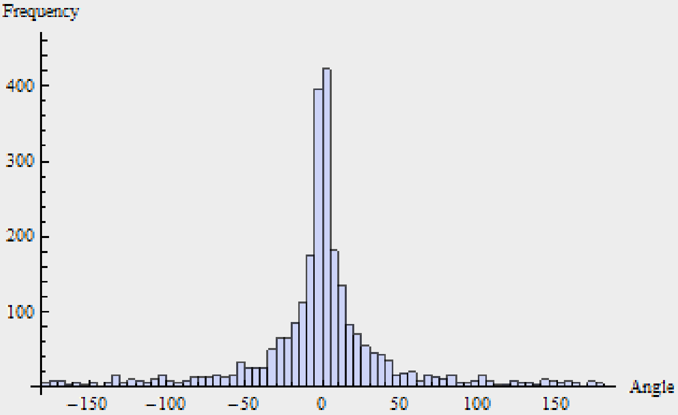
Angular Dependence

- 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 dot the anchor point, and the green dot 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.



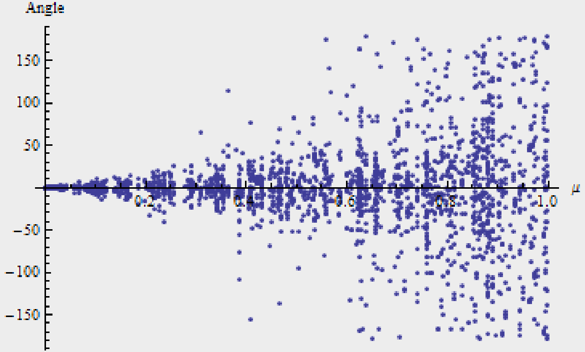
Angular Dependence

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



Angular Dependence

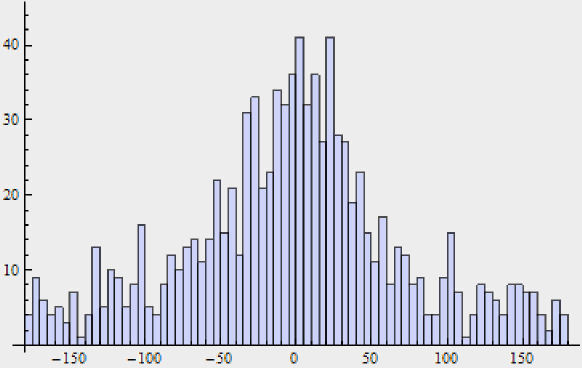
- We can again examine the angular variation as μ_2 varies.



- Even for relatively large values of μ_2, the data is clustered near the zero angle.

Angular Dependence

- The strong central peak reamains, even if we restrict our attention to series with $\mu_2 > 0.7$:



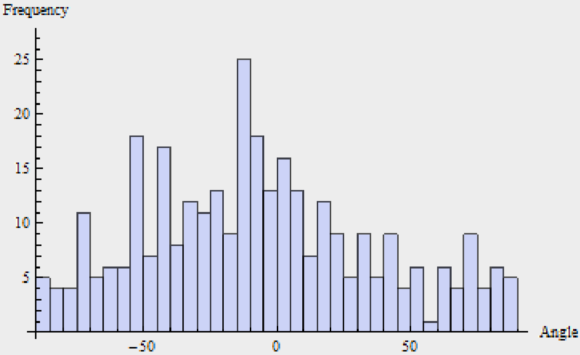
- Note the dramatic changes in the vertical scale between these images!

Angular Dependence

- Clearly there is a strong relationship between the directions the offender took to the different crime sites.
- Moreover, this relationship appears to be strong whether the offender is a commuter or a marauder.
- This suggests that weak information about direction would be more valuable than strong information about distance if one wanted to reduce the area necessary to search for the offender.

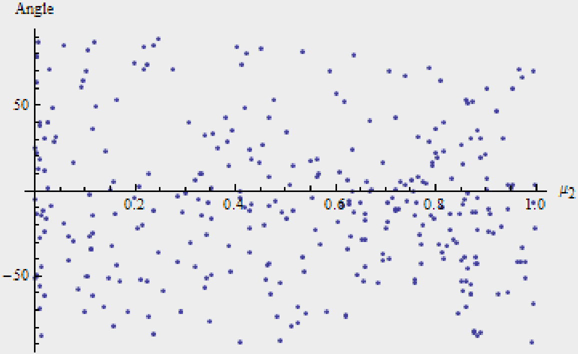
Angular Dependence

- One approach would be to calculate the principal component axes for the set of crime sites.
- It does not appear that there is a strong relationship between the principal axis and the direction to the offender's anchor point in this data.



Angular Dependence

- Moreover, this lack of a relationship persists whether or not the offender is a “commuter” or a “marauder”.



- Clearly much more work needs to be done to understand this phenomenon.

Questions?

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

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