Improving Mathematical Approaches to Geographic Profiling

Mike O'Leary
Towson University

Abstract

This report describes a nearly three-year long effort to improve the science of geographic profiling. Work proceeded along two lines: a theoretical line to construct improved models of offender behavior, and a practical line to improve an existing software prototype that had been released in 2009.

New models of offender behavior based on scaled bivariate normal distributions have been developed that show agreement with observed offender behavior from Baltimore County and other jurisdictions. However the deviations of the predictions with the observed data imply that there is still much more that needs to be understood.

A new and updated software prototype, including source code, has been released. It incorporates new usability features, like the ability to natively use shapefiles; it also has been extensively tested on solved residential and non-residential burglaries in Baltimore County. The prototype's search area contained the home of the offender for 74% of the non-residential burglaries and 70% of the residential burglaries, including 66% of the commuters for non-residential burglaries and 47% of the commuters for the residential burglaries. These search areas are comparable in size to Canter circles.

Contents

Executive Summary

Introduction

The primary question of geographic profiling is, given the locations of a series of crimes committed by a single serial criminal, to estimate the location of that offender's anchor point. A number of approaches have been developed to solve this problem; the most well known methods are those of Rossmo (2000), Levine (2010), and Canter, Coffey, Huntley, and Missen (2000).

In previously funded work, O'Leary (2009a, 2010) developed a new approach to the geographic profiling problem that began by showing how a model of offender behavior can be combined with the locations of the crime locations and Bayesian inference to produce estimates of the offender's anchor point. That work then continued by developing some reasonable models for offender behavior, and applying the technique to actual crime series. A software prototype that implemented the resulting algorithm was developed and released to police agencies.

The currently funded project is a continuation of that project. There were two primary goals: to improve the software prototype and to improve the mathematical modeling process, and to validate some of the choices made in the development of the model. Both goals were accomplished.

Summary of Major Accomplishments

Effectiveness Testing. The effectiveness of the new and updated prototype has been tested by calculating the geoprofile for 237 solved residential burglary series and 74 non-residential burglary series from Baltimore County from 1986 through 2009. The prototype's search area contained the anchor point of the offender 70% of the time for the residential burglary series and 74% of the time for non-residential burglary series.

Simply containing the offender's anchor point however is no guarantee of the effectiveness of the prototype; after all one could obtain 100% accuracy by simply selecting a search area that is sufficiently large. However, the search area produced by the prototype, while larger than those studied by Paulsen (2006), is comparable in size to the size of the search area produced by Canter's circle theory, namely the circle whose diameter is formed by the segment connecting the two crimes that are farthest apart. The prototype is much more accurate than circle theory however, as only 48% of the residential burglaries and 49% of the non-residential burglaries were marauders.

Software Prototype Round Off Error Correction. The original version of the software prototype was released in Summer 2010. Soon after, we received comments from practitioners who have been using the prototype, including analysts from the Pinellas County Sheriff's Office, the New Haven Police Department, as well as and the Baltimore County Police Department. All of these users have provided valuable feedback on the strengths as weaknesses of the tool.

It became apparent that the tool could return erroneous results; in particular at points far away from both the crime sites as well as the jurisdiction, the prototype was returning progressively larger likelihoods that the location would contain the offender's residence. Further analysis showed that the problem was in a core component of the mathematical model used in the algorithm.

To correct this error, a new mathematical approach to the calculation of one of the model's components was required. This correction has been implemented in the prototype, tested, and I believe has solved the round-off error problem.

Additional Software Features. A number of new featured have been added to the software prototype. The first request from users was for the prototype to natively support ESRI ArcGIS Shapefiles. That support has now been built in to the tool.

The second feature request was to allow the analyst to specify the bandwidth in the various estimated distributions. In particular, the prototype uses the locations of historical crimes to create an estimated distribution of offender likely targets via kernel density parameter estimation. In the original prototype, the bandwidth for this estimate was calculated automatically. The new version of the software now allows the analyst the option to manually specify the bandwidth.

Another new feature is that the tool provides additional methods to estimate the prior distribution of the offender's anchor point. This prior distribution represents the analyst's knowledge of the offender's anchor point before any information from the crime series is used. In the originally released prototype, this distribution was estimated by using demographic data about the offender coupled with block level data from the 2000 US Census. Two major changes have now been made. First, the prototype now uses the data from the 2010 Census rather than the older 2000 data. Second, it gives the analyst the option to use the anchor points of known offenders to generate the distribution rather than relying on the Census data.

Another new feature is a completely new graphical user interface. The original user interface assumed that the user correctly entered the data in the required format in the required files. However, some officers were unable to get the tool to work; it turned out that one of the issues an officer had was that the data files were specified in the form latitude, longitude; the program requires that the data come in the opposite order. To make the tool more usable, the user interface now reads the file specified by the user, and reports back some information back to the user; for example for the crime series it reports back the number of crimes in the series as well as the minimum and maximum latitude and longitude of the elements of the series. Though this will not prevent the user from providing the data in the incorrect order, it will at least provide feedback so that this kind of problem could be detected by the analyst.

Theory: Marauders and Commuters. In addition to work on the software prototype, work continued on the underlying theory. The first contribution was to create a more refined approach to classifying offenders and commuters or marauders. The classical definition of marauder and commuter (Canter & Larkin, 1993) is the following: Given a crime series, take the two elements of the series that are farthest apart, and consider the circle whose diameter is the segment between these two elements. If the offender's anchor point lies in this circle, the offender is a marauder; otherwise the offender is a commuter.

Though valuable, this definition is a binary characterization that treats similar offenders quite differently. Indeed, if an offender's anchor point is quite close to the boundary circle, then they would be classified solely on the basis of which side of the boundary line the anchor point lies, even if that distance we simply across a street. We have developed a more nuanced approach that associates a number μ between zero and one to a series; offenders with values near zero are considered to behave like commuters while offenders with values near one are considered to behave like marauders.

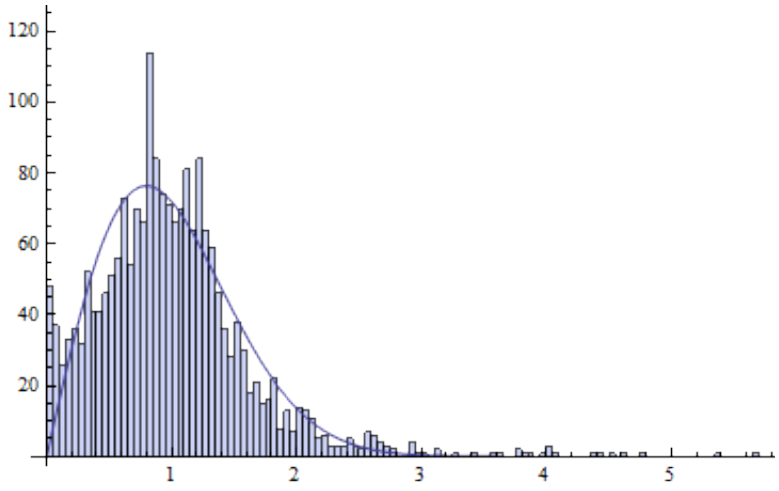
Theoretical Model Validation. One of the underlying assumptions made in the development of the original software prototype was that the distance decay patterns of offenders could be well-modeled by a bivariate normal distribution. There is now some significant evidence for something close to this, but it is also just as clear that this is not the complete story.

It is well-known that different offenders behave differently. One immediate consequence of this fact is that one must be very careful when using aggregate data to make inferences about the behavior of individuals; this essential problem is called the ecological fallacy. However, if we assume that the only quantity that varies between offenders is the average distance that the offender

is willing to travel, then the scaled distances ρ, defined as the ratio of the distance from crime site to offender anchor point with the average distance that offender travels, should exhibit the same behavior regardless of the offender. In particular, this allows us to aggregate data across offenders and draw valid inferences about the assumed universal behavior. If we start with the assumption that individual offenders select targets via a bivariate normal distribution centered at the anchor point, then the distances from crime site to anchor point should follow a Rayleigh distribution, and the scaled distances should follow a Rayleigh with average distance one.

We analyzed a data set of 5863 residential burglaries for Baltimore County Maryland, which contained 322 crime series with at least four crimes. Examining the scaled distances ρ for marauders, defined as those offenders for which $\mu \geq 0.25$ and comparing it with the theoretically predicted Rayleigh distribution with average 1, we found a good fit; see Figure 1.

Figure 1. Scaled distances for residential burglary marauders in Baltimore County and a Rayleigh distribution with average 1.



This agreement with the Rayleigh distribution does not appear to be happenstance. We obtained similar good fits for non-residential burglary and for bank robberies in Baltimore County. Moreover, by examining the work of Warren et al. (1998) who also graphed scaled distances for serial rapes, we observed the same fit with a Rayleigh distribution.

Despite the fits of the data to the theoretical Rayleigh distribution, this is not sufficient to conclude that the two-dimensional distribution of offense sites is necessarily a bivariate normal distribution; indeed there are an infinite number of bivariate distributions whose distribution of distances is Rayleigh.

If the underlying distribution is bivariate normal, then there should be no angular dependence in the results. However the same Baltimore County data show that nearly all crimes lie in the same direction as the centroid. This behavior is observed even if attention is restricted solely to marauders. As a consequence, it is clear that the underlying bivariate distribution is not bivariate normal.

The clear agreement of the scaled distance data with a Rayleigh distribution with average distance one suggests that there is some identifiable pattern in the data; at the same time the clear disagreement of the data with the bivariate normal distribution tells us that the true situation is more

complex.

A number of different approaches have been tried to construct better models for the offender's behavior. The most promising approach found so far has been a model than explicitly accounts for dependence between the locations of the offender's crimes. This is recent work done jointly with my Master's student, Jeremiah Tucker.

Introduction

The primary question of geographic profiling is, given the locations of a series of crimes committed by a single serial criminal, to estimate the location of that offender's anchor point. A number of approaches have been developed to solve this problem; the most well known methods are those of Rossmo (2000), Levine (2010), and Canter et al. (2000).

In previously funded work O'Leary (2009a, 2010) developed a new approach to the geographic profiling problem that began by showing how a model of offender behavior can be combined with the locations of the crime locations and Bayesian inference could produce estimates of the offender's anchor point. That work then continued by developing some reasonable models for offender behavior, and applying the technique to actual crime series. A software prototype that implemented the resulting algorithm was developed and released to police agencies.

The currently funded project is a continuation of that project. There were two primary goals: to improve the software prototype and to improve the mathematical modeling process, and to validate some of the choices made in the development of the model. Both goals were accomplished.

Report organization. This report begins with an introduction to the geographic profiling problem, then summarizes the main results of our investigation, including the improvements made to our software prototype and the new mathematical models developed. The prototype is then discussed in detail, beginning with complete instructions and a case study demonstration for officers and analysts who wish to use the tool. The report goes on to examine the effectiveness of the tool on residential burglary series and non-residential burglary series in Baltimore County. The discussion of the prototype concludes with details of the prototype's internal structure.

The report then discusses the underlying mathematical models that were implemented in the prototype, beginning with an analysis of offender distance decay behavior. It is shown that the Rayleigh distribution is a good model for scaled offender distance decay, but that the corresponding bivariate normal model has significant weaknesses. Other models for offender behavior are then considered, including models where crime site locations explicitly depend on the locations of previous crimes in the series. The report concludes with implications for policy and practice, as well as a discussion of fruitful areas for future research.

The Geographic Profiling Problem

Geographic profiling is the problem of creating an estimate of the anchor point of a serial offender based on knowledge of the locations of the offender's crimes. The anchor point can be the offender's home; it can also be the offender's place of work, favorite bar, or other place of particular importance to the offender.

One approach to this problem is to use a centrographic approach, where a point estimate of the offender's anchor point is made from the location of the crime series. Common centrographic techniques include estimating the offender's anchor point from the centroid of the crime series or from the center of minimum distance of the elements of the crime series. These estimates can be supplemented with other information, like the size of the standard deviational ellipse (*c.f.* LeBeau (1987)), however at their core they remain simple point estimates.

An approach that uses search regions as the fundamental building block was proposed by Canter and Larkin (1993). They distinguish two types of offenders- "marauders" who are assumed to move from their home base to the crime site and then return, and "commuters" who first move

from their home base to another area before committing the crime and then returning. For a marauder, the criminal range and the offender's home range overlap while for a commuter these ranges are essentially separate.

In the circle hypothesis of Canter and Larkin, begin with a crime series and select the two crimes farthest apart; then form the circle whose diameter is the segment that connects these two crimes. The offender's anchor point and all of the offender's crimes should lie in the circle so constructed.

There is evidence for the validity of this hypotheses for certain classes of crimes. In their original paper, Canter and Larkin (1993) examined a collection of 45 male sexual assaults in Britain. In 41 of the 45 cases the circle correctly encompassed all of the crime sites and in 39 of the 45 cases the circle correctly contained a base for the offender. Kocsis and Irwin (1997) examined 24 rape series, 22 arson series, and 27 burglary series in Australia. The circle contained all of the crimes for 79% of the rape series, and 82% of the arson series, while the circle correctly contained the home base of the offender for 71% of the rape cases, and 82% of the arson cases.

On the other hand, circle theory has been seen to be much less effective on burglary series. Kocsis and Irwin (1997) also examined 27 burglary series, and found that the circle contained the home location of the offender only for 48% of the series. In a follow up paper, Kocsis, Cooksey, Irwin, and Allen (2002) analyzed 58 burglary series committed in one of four rural Australian towns; they found the circle contained the offender just half the time. Similar results were obtained by Sarangi and Youngs (2006), who analyzed 30 burglary series from the Indian state of Orissa, and found that the circle contained the offender's home only 57% of the time. Laukkanen and Santtila (2006) examined 76 commercial robbery series in Greater Helsinki from 1992 to 2001 and found that the circle contained the offender's home only 39% of the time.

The three major computer systems used for geographic profiling (CrimeStat, Dragnet, and Rigel) have each implemented an alternative approach that Snook, Zito, Bennell, and Taylor (2005) call a probability distance strategy. In this approach a hit score is calculated by placing a distance decay function on the site of each crime and summing the results. Regions with higher hit scores are considered more likely to contain the offender's anchor point than areas with lower scores. The different programs use different decay functions and different methods of measuring distance. For example, while Rossmo (2000) (Rigel) uses the Manhattan distance and an algebraic form for distance decay with a buffer zone, Canter et al. (2000) (Dragnet) uses Euclidean distance and an exponential distance decay with and without a buffer and/or plateau. Levine (2010) (CrimeStat) allows the use of different distance metrics and different distance decay functions, including user specified functions.

CrimeStat is also capable of using a Bayesian based algorithm to estimate the anchor point of a serial offender. As described by Levine and Lee (2009), Block and Bernasco (2009), Leitner and Kent (2009), and Levine and Block (2011), a region surrounding the crimes is subdivided into zones. Historical data is then used to determine how many offenders in one zone committed a crime in another zone to create a conditional origin-destination matrix. This data is incorporated into a Bayesian framework to estimate the location of the anchor point of the serial offender. Kent and Leitner (2012) show how to modify this Bayesian model by incorporating prior data for land use.

A more mathematical approach has been taken by Mohler and Short (2012). They modeled the movement of individual offenders via a stochastic differential equation; this allows them to incorporate geographic features directly into their model. The corresponding Fokker-Planck forward equation can then be used to determine the probability density that a particular offense site is chosen

given a known anchor point, while the corresponding backward equation then gives the probability density that the offender has a particular anchor point given a known crime site. Multiple crimes are handled by assuming that crime sites are selected independently. They calibrated the parameters in their model by examining 221 solved burglary series from the Los Angeles Police Department.

Project Foundation

Our approach, begun in our previously funded work (O'Leary, 2009a, 2010), is to start by specifying a model for offender behavior. This model gives the probability density that an offender with a particular anchor point would commit an offense at a specified location. One important feature of this approach is that it makes explicit the underlying assumptions that are being made on offender behavior. All geographic profiling methods necessarily make assumptions on offender behavior; one of the reasons it is difficult to compare the different approaches is that it is unclear what these underlying assumptions are.

Beginning with a model of offender behavior turns the question from a problem in criminology to one in mathematics. Indeed, if the assumed model of offender behavior is correct, then the geographic profiling problem reduces to a classical statistical problem of parameter estimation where the parameter of interest is now the location of the offender's anchor point.

Our model assumes that the offender chooses a target based on two factors- the distance from the anchor point to the location of the offense, and the relative attractiveness of the chosen target location. To be more precise, the model assumes that the distance decay component can be modeled by a bivariate normal distribution. It is important to use a bivariate distribution (either normal or some other distribution) here rather than a simple model of distance, as offenders do not select distances, they select targets, and targets are distributed in a two-dimensional space. Using a bivariate normal distribution requires that the distribution of distances follows a Rayleigh distribution; this is in contrast to the typical models used in other approaches that are typically negative exponential or polynomial in form. Our model does not assume that all offenders have the same distance decay function, but rather that different offenders have different average travel distances to offend and that this parameter can vary by offender. This average offense distance is an additional parameter that will be estimated at the same time as the anchor point by the data in the crime series.

In addition to hypothesizing a distance decay effect, the model also assumes that some locations are more likely to be the site of an offense than others due to geographic factors particular to that location. As the simplest example, it is known that if a location does not contain a liquor store, then it cannot be the location of a liquor store robbery, so any offender who wanted to commit such a crime there would need to continue their journey to crime. Many instances however, are more subtle. For example, one region may be more likely than another to be the site of a street robbery, even though street robberies can occur in either location. The cause of this variation may be due to any number of geographic features- perhaps one area is better lit, or better patrolled, or has fewer potential targets. The model does not speculate on the underlying cause of the variation but simply tries to account for its presence.

With the model specified, Bayesian methods then allow the estimation of the parameters in the model provided prior distributions for these parameters are specified. In this case, there are two parameters- the location of the offender's anchor point and the average distance the offender is willing to travel. The prior distribution of the offender's anchor point is our estimate before we take into account any information from the crime locations themselves. Meaningful priors can be

constructed in a variety of ways, including by examining the distribution of anchor points of past offenders or by examining the distribution of population in the area by using Census data. The prior distribution of the offender's average offense distance can be calculated by looking at historical patterns of offender travel behavior.

Though the situation with a single crime is handled in this fashion, the problem of multiple crimes is somewhat more subtle. The model we have constructed gives the probability density that the offender commits a single crime. If the crime locations are all independent, then the joint probability density that the offender commits crimes in two different locations is simply the product of the individual probability densities. Mathematically this is the simplest approach and this is what was implemented in the prototype. However, subsequent analysis has shown this independence assumption to be problematic, and we have proposed and analyzed some additional models that explicitly allow for non-independence.

When complete, this process yields a joint probability distribution for both the offender's anchor point as well as for the offender's average offense travel distance. To complete the analysis, simply marginalize across the average offense distance distribution, by adding up (integrating, actually) the distribution of the offender's anchor point for each average offense distance multiplied by the probability density that the offender had that average offense distance.

This technique had been developed as part of the previously funded project, and a software prototype released. That prototype had two components- an analysis engine that performed the scientific calculations described above, and a graphical user interface for the analyst to use. This prototype was released to the public, and source code provided. Together with the mathematical algorithm described above, this forms the foundation of the current project.

Project Accomplishments

Effectiveness Testing. The effectiveness of the updated prototype has been tested by calculating the geoprofile for 237 solved residential burglary series and 74 non-residential burglary series from Baltimore County from 1986 through 2009. The prototype's search area contained the anchor point of the offender 70% of the time for the residential burglary series and 74% of the non-residential burglary series.

Simply containing the offender's anchor point however is no guarantee of the effectiveness of the prototype; after all one could obtain 100% accuracy by simply selecting a search area that is sufficiently large. However, the search area produced by the prototype, while larger than those studied by Paulsen (2006), is comparable in size to the size of the search area produced by Canter's circle theory, namely the circle whose diameter is formed by the segment connecting the two crimes that are farthest apart. The prototype is much more accurate than circle theory, as only 48% of the residential burglaries and 49% of the non-residential burglaries were marauders.

Software Prototype Round Off Error Correction. The original version of the software prototype was released in Summer 2010. Soon after, we received comments from practitioners who were using the prototype, including analysts from the Pinellas County Sheriff's Office, the New Haven Police Department, as well as the Baltimore County Police Department. After their comments came in, it became apparent that the original tool could return erroneous results in a very precise collection of circumstances. In particular, if the crime sites were all close together and the jurisdiction small, then at points far away from both the crime sites and the jurisdiction the prototype returned progressively larger likelihoods that the location would contain the offender's residence. The farther away the proposed anchor point became the larger these erroneous values

became, until they dwarfed the (correctly calculated) values near the crime series and jurisdiction. Further analysis showed that the problem was in a core component of the mathematical model used in the algorithm.

The problem turned out to be quite difficult to fix, as the underlying problem was very subtle. The implemented model of offender behavior is that the probability density that an offense is committed at a particular location depends on the distance from that point to the anchor point as well as on the geographic features of that target location. To make this approach mathematically rigorous, the sum (integral actually) of the resulting probability density must equal precisely 1; normally this is handled simply by scaling. The first issue to note though, is that the scaling factor required is going to depend on the location of the anchor point itself. Both the probability density and the scaling factor decay as the anchor point moves farther and farther from the crime. Unfortunately, these quantities need to be divided, and when the distance from the anchor point to the crimes is sufficiently large, then both numbers become so small that the round-off error when the calculation is performed on the computer becomes so large as to cause significant errors. Even though the underlying mathematics is correct, its implementation on the computer was flawed.

To solve this problem, significant changes needed to be made to the code; in particular the way this quantity is evaluated needed to be entirely rewritten so that none of the intermediate calculations would result in significant round off error. This has been done, and the updated prototype has been extensively tested.

An unanticipated consequence of this change is that the required time to complete the computation has increased; on a fast computer calculations typically take the better part of a day, and can run longer.

Additional Software Features. After listening to comments from the users of the original prototype, a number of new features were added to the tool. The first and most common request was for the tool to natively support ESRI ArcGIS Shapefiles. That support has now been built in to the tool.

The second feature request was to allow the analyst to specify the bandwidth in the various estimated distributions. In particular, the prototype uses the locations of historical crimes to create an estimated distribution of offender likely targets via kernel density parameter estimation. In the original prototype, the bandwidth for this estimate was calculated automatically. The new version of the software now allows the analyst the option to manually specify the bandwidth.

Another new feature is that the tool provides additional methods to estimate the prior distribution of the offender's anchor point. This prior distribution represents the analyst's knowledge of the anchor point before any information from the crime series is used. In the originally released prototype, this distribution was estimated by using demographic data about the offender coupled with block level data from the 2000 US Census. Two major changes have now been made. First, the prototype now uses the data from the 2010 Census rather than the older 2000 data. Second, it gives the analyst the option to use the anchor points of known offenders to generate the distribution rather than relying on the Census data.

In this new approach, the analyst provides the prototype with the anchor point locations of known offenders; the analyst can then either manually specify a bandwidth or rely on the tool to automatically choose a bandwidth. The tool then generates a map of the prior anchor point distribution using kernel density estimation. The advantage of this approach is it explicitly accounts for the fact that some neighborhoods are more likely to contain offenders that others. The disadvantage is a loss in spatial resolution. Indeed, Census data is available to the block level. However, when using

kernel density parameter estimation on offender anchor points, you will make the assumption that the offender does not live at a distance greater than the bandwidth from another offender. As such, it is important to be sure that the bandwidth is sufficiently large that this is a reasonable assumption, which puts a lower limit on the bandwidth. In testing, a resulting bandwidth has been on the order of a mile, which means that this approach cannot resolve geographic features smaller than this size. In contrast, Census data can resolve features at the block level. Of course, the bandwidth would depend on the quality of the data; the larger the data set, then the smaller the bandwidth could potentially be set, however it is unlikely to ever be sufficiently small as to resolve features on the block level.

Another new feature is a completely new graphical user interface. The original user interface assumed that the user correctly entered the data in the required format in the required files. However, some officers were unable to get the tool to work; it turned out that one of the issues an officer had was that the data files were specified in the form latitude, longitude; the program requires that the data come in the opposite order. To make the tool more usable, the user interface now reads the file specified by the user and reports back some information back to the user; for example for the crime series it reports back the number of crimes in the series as well as the minimum and maximum latitude and longitude of the elements of the series. Though this will not prevent the user from providing the data in the incorrect order, it will at least provide feedback so that this kind of problem could be detected by the analyst.

Improved Theory for Commuters and Marauders. One weakness of the the usual circle theory notion of commuter and marauder from Canter and Larkin (1993) is that it is a binary characterization of the offender. Either the offender is a commuter or the offender is a marauder; there is no option to indicate that the offender's behavior is demonstrates a blend of these characteristics. To provide a more nuanced approach, I developed a mathematical score for an offender to measure how much the observed pattern is like that of a marauder and how much it is that of a commuter.

Model Validation. As already noted in the discussion of foundational issues, the prototype assumes that the offender's behavior has two components, a distance decay component and a target attractiveness component. Moreover, it assumed that the distance decay component can be well modeled by a bivariate normal distribution. This assumption then implies that the one-dimensional distance decay factor should be modeled by a Rayleigh distribution. However, the Rayleigh distribution is rarely used as a model for distance decay behavior; negative exponential and piecewise rational functions have been much more common.

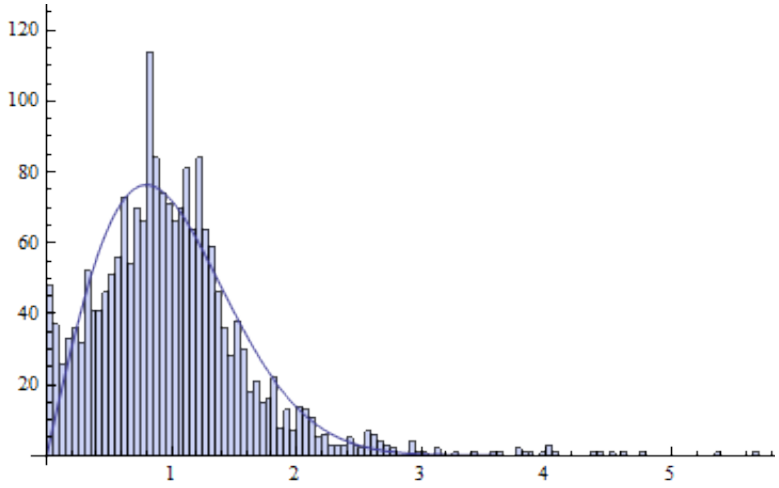
Given this difference, and the fundamental role played by the distance decay factor in the both the prototype and the general analysis of offender travel patterns, it is essential to validate this assumption. Doing so requires more than simply comparing a Rayleigh distribution to the distance decay curve found by aggregating many offenders and many crimes. The prototype explicitly assumes that different offenders may behave differently. As a consequence, the behavior of the aggregate may be significantly different than the behavior of any individual; this is the essential core of the ecological fallacy.

To account for the variation between offenders, for each offender the average distance that offender traveled to crimes was calculated. Then for each crime, the ratio of the distance the offender traveled to their average offense distance was calculated. This is a dimensionless quantity, and its use was motivated by the effectiveness of dimensional analysis in a range of mathematical problems, including fluid mechanics. The key fact is that, if individual offenders actually do commit crimes with a one-dimensional Rayleigh distance decay function with the only difference being the average

offense distance, then the dimensionless ratios would satisfy a Rayleigh distribution with an average distance of precisely one. Moreover, all offenders would be expected to demonstrate the identical pattern; thus it would be appropriate to aggregate data across multiple offenders.

This analysis was performed on a set of 5,863 residential burglaries in Baltimore County. Once obvious commuters were excluded, using the numerical method of distinguishing commuters and marauders already described, agreement is observed between the data and the theoretical curve (Figure 2). Note that this fit is not caused by carefully selecting a parameter to make the theoretical

Figure 2. Scaled distances for residential burglary marauders in Baltimore County and a Rayleigh distribution with average 1.



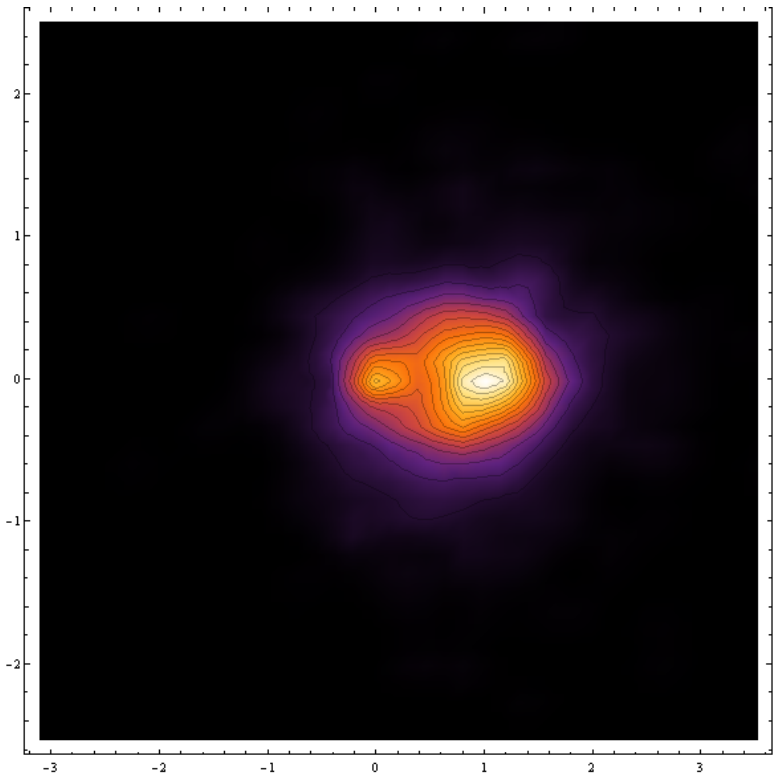
curve fit the data; the hypothesis was that the Rayleigh distribution with average 1 would fit the data, and this is what was observed.

The hypothesis was checked again on non-residential burglaries and again on bank robberies within Baltimore county, and the same agreement with the theoretical prediction is observed in both cases. Further, Warren et al. (1998) graphed scaled distance for serial rape; their observed distances also appear to match the theoretical predictions quite closely.

Though the one-dimensional distance decay patterns do fit the predicted Rayleigh distribution quite well this is not, by itself, sufficient evidence to conclude that the two-dimensional distribution is necessarily bivariate normal, as there are many possible two-dimensional distributions whose distance decay patterns are also bivariate normal. To better understand the full bivariate normal distribution, one needs to understand the directional dependence in the crime series. Taking the ray from the offender's anchor point to the centroid of the crime series as a referent, one can calculate the angles between the referent and the ray from the offender's anchor point to a crime site. If the underlying distribution is truly bivariate normal then these directions should be roughly uniformly distributed; in actual fact they are not.

To better see what is occurring, plot both the scaled distance and the angles together on the same graph scaled so that the centroid of the each crime series is at the same point; this is shown in Figure 3. If the underlying distribution were bivariate normal, we would expect to see a unimodal distribution centered at the origin. The observation however, is quite different. Not only is

Figure 3. Colored contour plot of smoothed histogram of the scaled crime locations for Baltimore County residential burglary; the anchor point is the origin and the centroid of the crime series is $(1, 0)$



it bimodal, the peak at the centroid of the crime series is much larger than the peak at the offender's anchor point. This suggests that crimes, rather than being centered around the offender's anchor point are instead centered around each other.

Offender Models with Explicit Dependency Structure. To investigate the question of the inter-dependence of the locations of the elements of the crime series, my Master's student (Jeremiah Tucker) and I examined different models that explicitly account for possible dependency structures. We considered three types of models of offender distance decay behavior, beginning the bivariate normal as the base case. The second model was that the offender would either choose a crime site that is a near repeat of a previous crime, or they would choose a new crime site according to a bivariate normal distribution. In the third general model, the crime site is chosen either from a bivariate normal distribution centered at a previous crime site or from a bivariate normal distribution centered at the offender's anchor point.

These different models were analyzed with the Akaike Information Criterion using the small sample correction (AICc); this method is well described by Burnham and Anderson (2002). The analysis was performed on 136 solved residential burglary series, 43 solved non-residential burglary series and 10 solved bank robbery series with at least seven crimes in Baltimore County. Because the series were solved, two approaches could be taken with each model- either the anchor point of

the offender could be considered known, or it could be considered as unknown and then calculated from the data. All six cases were then analyzed and compared.

Our analysis showed that for non-residential burglary with the home considered known, the normal model is by far the least supported by the data; the near repeat model was most supported followed by the general model. This same behavior was observed in bank robberies. The situation for residential burglaries is more nuanced; in 75 of the 136 series the normal model was the least supported but there were roughly 50 cases where the normal model received significant support. It is still the case though that overall the near repeat and general model significantly out-performed the simpler normal model. When the home is considered unknown, essentially the same behavior is still observed.

This completes our introduction to the project's main accomplishments. In the next section we will discuss the software prototype in detail, beginning with complete instructions for its use. The section continues with an analysis of its effectiveness, looking at both how often the provided search area contained the offender's actual home, as well as the overall size of the search area. The section concludes with a brief summary of the prorotype's internal structure. The source code for the prototype is available for others to use and modify.

Following the discussion of the prorotype, we turn our attention to the theoretical underpinnings of our approach. In contrast to the introduction, complete mathematical details will be provided. The report concludes with conclusions and a list of open problems worth further analysis.

The Software Prototype

Using the Prototype

The software prototype has been developed and released, and it is available online at `http://pages.towson.edu/moleary/Profiler.html`. Source code is also available, and will be provided to all who ask for it. Copies of both the final executable code as well as the underlying source code are provided along with this report.

Usage Instructions. To use the tool, the analyst simply downloads the tool and uncompresses it into a convenient directory. There is no installation process and the tool does not need administrator privileges to use. From the directory, run the program named `ProfilerGUI.exe`; when this occurs the program will display its main screen as shown in Figure 4.

To use the tool, the analyst first specifies the elements of the crime series under consideration. From the Crime Series box in the main dialog, select the Provide Data Button. The Crime Series box and other boxes in the main dialog contain help buttons that describe the purpose of that data component. For example, when the help button from the Crime Series box is pressed, the tool displays the dialog box in Figure 5 that describes the purpose of the data.

When the analyst selects the Provide Data button from the Crime Series box, a new dialog is presented to allow the analyst to specify the file that contains the locations of the crime series under consideration; this is shown in Figure 6.

The crime site locations need to be contained in a single file; three formats are supported:

- Plain text files. In this case, the file needs to contain the location of one crime site on each line. The line should start with the longitude of the crime site, then either one or more spaces or a comma, then the latitude of the crime site.
- Shapefile .dbf files. In this case, the analyst will be asked to select which fields contain the longitude and latitude of the crime site locations.

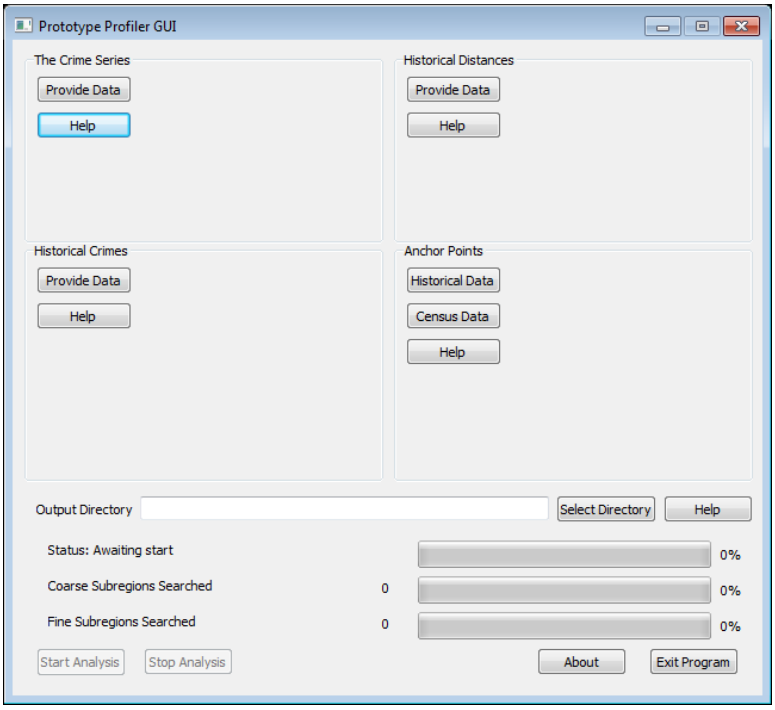


Figure 4. Main window for software prototype

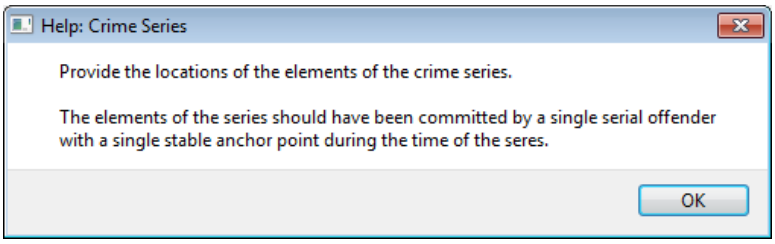


Figure 5. Help dialog box for crime series

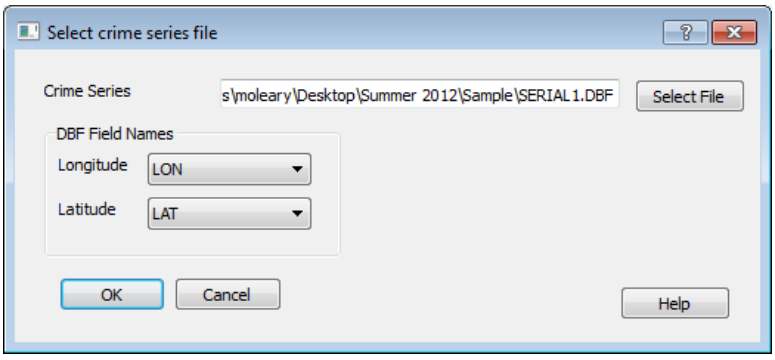


Figure 6. Entering the locations of a crime series contained in a shapefile .dbf file.

• Comma separated value .csv files. Like a plain text file, the .csv needs to contain the location of one crime site on each line. The first entry should be the longitude of the crime site while the second the latitude of that crime site.

Regardless of the file type, locations are specified in decimal degrees using the usual WGS84 datum.

Pressing the help button in the Crime Series Provide Data dialog box in Figure 6 tells the analyst the supported file types and the required file structures; see Figure 7 for an illustration of the provided help.

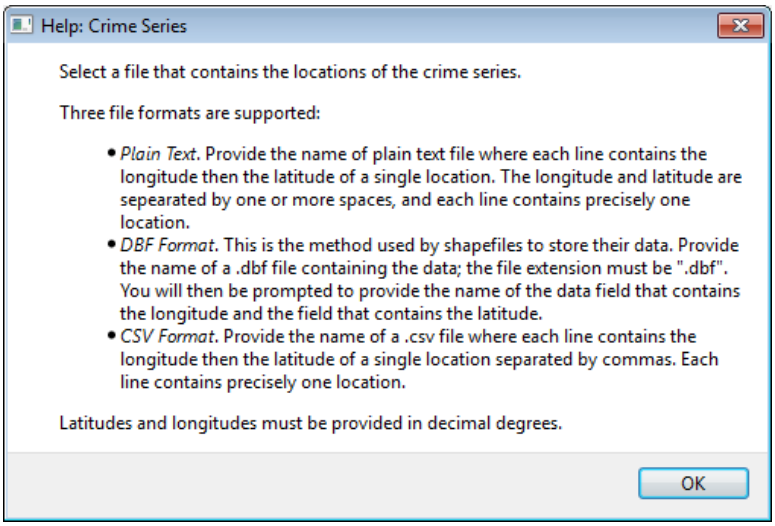


Figure 7. Help provided to the analyst after selecting help from the Select crime series file dialog box.

To understand the travel patterns of offenders in the jurisdiction under consideration, the program requires a collection of solved crimes so that it can calculate the distances offenders travel. The analyst selects the Provide Data button in the Historical distances box from Figure 4; then they must provide a file that contains both the location of the offense site and of the offender anchor point for a collection of solved offenses of the same general type as the series under consideration. These crimes do not need to be series crimes, and they do not need to be committed by the same offender. This data is used to estimate a prior distribution for the average distance an offender in this jurisdiction is willing to travel to offend.

Like the Crime Series box, there is a help button that tells the analyst the purpose of the required data. Once the Provide Data button is pressed, another dialog box is presented to let the analyst specify the name of the file that contains the required data. Like the Crime Series, the data can be a plain text file, a shapefile .dbf file, or a comma separated value .csv file. If a shapefile .dbf file is provided, then the analyst will need to provide the names of the fields that contain the longitude and latitude of the crime site location and the fields that contain the longitude and latitude of the corresponding home location. For a plain text file or a .csv file, each line contains the longitude then latitude of the crime site followed on the same line by the longitude and the latitude of the anchor point. The corresponding help buttons provide the user with this same information.

To understand the distribution of attractive crime targets in the jurisdiction, the program requires a collection of crime sites similar to the elements of the series. This is then used to generate

a map like a hot spot map of the geographical regions that are more or less likely to be the site of a crime of the same general type as the series under investigation. The analyst provides a data file that contains the locations of these crimes. These crimes need not consist of solved crimes and they need not necessarily be series crimes. Instead the collection should be sufficiently large and robust that it is representative of of the distribution of crimes similar to the series crimes throughout the jurisdiction.

Like the Crime Series box, there is a help button that tells the analyst the purpose of the required data. Once the Provide Data button is pressed, another dialog box is presented to let the analyst specify the name of the file that contains the required data. Like the previous dialog boxes, it asks for the location of a file that now contains the longitude and latitude of the historical crime locations, either as a plain text file, a shapefile .dbf file, or a comma separated values .csv file. Because this information is used to generate a hot spot map, the analyst has another choice to make, namely the bandwidth of the kernel density parameter used to make the estimate. By default this is calculated automatically by the program as the mean nearest neighbor distance between crime sites; however the analyst can override this and manually select the bandwidth. Figure 8 illustrates the dialog box presented to the analyst.

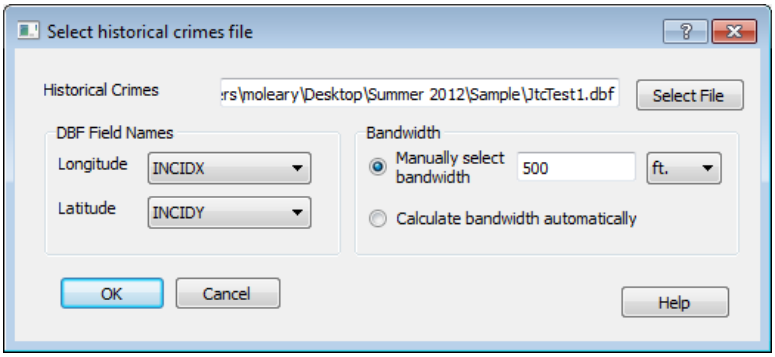


Figure 8. Selecting the historical crime file and the bandwidth

Next, the analyst needs to choose a prior distribution for the location of the offender's anchor point before any information from the crime series itself is used. The tool provides two methods to make this determination, either to use data from the US Census or to use the locations of known anchor points from prior offenders. The purpose of this information is to account for the fact that offenders are not distributed uniformly through the jurisdiction, but rather that there are places where offenders are more or less likely to be located.

If the analyst wants to use historical data to estimate the prior, they must select the Historical Data button; they will be presented with a dialog box like the one for the locations of historical crimes. The file should contain the locations of known offender anchor points either as a plain text file, a shapefile .dbf file, or a comma separated values .csv file as longitude/latitude pairs. The tool then uses kernel density parameter estimation in the same fashion as it would to generate a hot spot map to create the resulting density surface. As with the historical crimes, the analyst can accept the automatically specified bandwidth equal to the mean nearest neighbor distance or can manually select a bandwidth. One important feature though, is the fact that regions where the density is zero are assumed *a priori* by the analyst not to contain the anchor point of the offender. As a consequence, it is important that the number of past offenders in the sample and the selected bandwidth are both

sufficiently large that this is a reasonable assumption.

The analyst can instead use the 2010 US Census data to estimate the prior. To do so, they first select the Census Data button; they will then be presented with the dialog box shown in Figure 9. Up to four county-sized regions can be selected; these can be from the same or from different

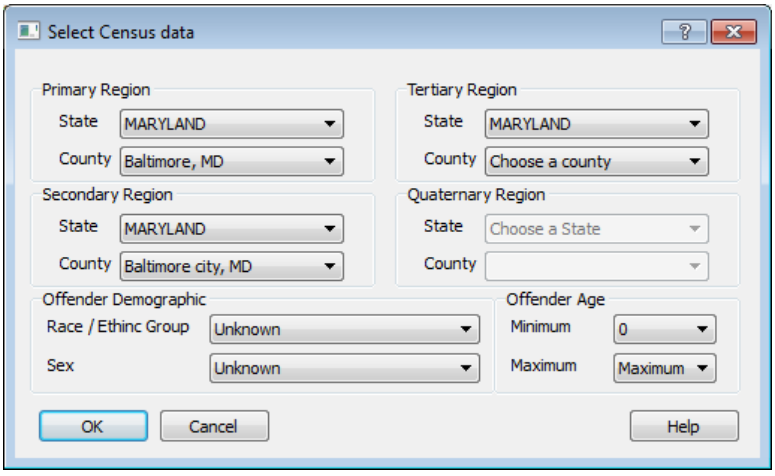


Figure 9. Using Census data to determine the prior offender distribution

states. The required Census data is not included with the program, primarily because of the very large size of the data set. If the analyst selects a state for which they do not yet have the required data, they will be presented with a dialog box like Figure 10. This dialog box specifies the precise

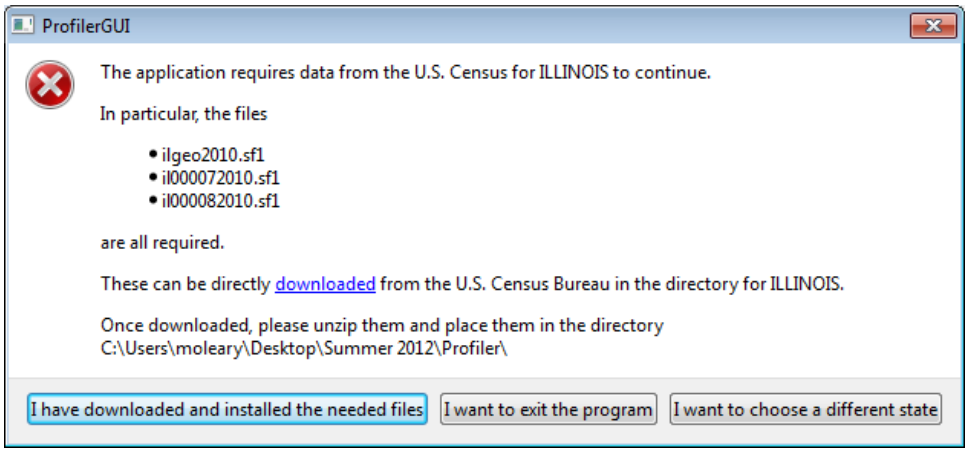


Figure 10. Error dialog for missing Census data files

files that are required and includes a link to the US Census web site where they can be downloaded. Once the files are downloaded, they simply need to be put into the proper directory, and the name of the directory is also included in the dialog box.

If demographic information is available about the offender, it too can be used. Census data is available at the block level disaggregated by age, sex, and race or ethnic group. If these options are selected, then only the matching census data will be used to generate the prior.

The analyst then needs to select the location that will contain the output of the program. The prototype generates a great deal of information and will store all of these results in the directory provided by the analyst when the Select Directory button is pressed.

The prototype will produce the following files:

The parameter file. This contains all of the parameters selected when the program was run in a plain text file. The name of the program has the form yyy-mm-dd-hh-mm that gives the year, month, day, hour, and minute the program was started.

The prior distribution of anchor points. These show the assumed prior distribution of anchor points, estimated either from the 2010 Census data or from the provided locations of the anchor points of historical offenders. This is returned as a .kml file, as a shapefile, and as a comma separated values .csv file.

The historical crime patterns. These show the assumed distribution of attractive targets in the jurisdiction, created as a hot spot map from the locations of the historical crimes. This is returned as a .kml file, as a shapefile, and as a comma separated values .csv file.

The offense distance patterns. Two sets of distance data are returned. The first is the calculated prior distribution of offense distances, before information from the crime series is taken into account; the second is the estimated average offense distance of the series offender after accounting for the crime locations and the prior distribution. Both sets of data use distance measured in decimal degrees along the surface of the earth and provide the results as a plain text file and as a comma separated values .csv file.

Estimated offender location. The prototype makes two passes through the geographic region. First it calculates a coarse approximation of a map of the offender's likely anchor point, and returns this. It then does a more thorough analysis and refines the coarse map to provide a set of fine results. Both of these maps are returned as .kml files, as shapefiles, and as comma separated value .csv files.

As the data is entered into the program, the choices made by the analyst will be reflected in the main dialog box. Compare Figure 11 which shows the tool ready to start its analysis with the original Figure 4. From this, we can see that the series under study has seven crimes; 2000 solved crimes were used to provide the offender distance calibration and the average offender in the data set traveled roughly 4.2 miles to offend. We also see that 2000 crimes were used to generate the hot spot map of historical crime locations and that the bandwidth in that map has been set to 500 feet. The prior distribution of offender anchor points was made through Census data using Baltimore City and Baltimore County in Maryland, and that no demographic information about the offender has been provided.

It is important that the analyst check all of the data provided by the tool for reasonableness. If the analyst, for example, interchanges latitude and longitude in a provided data set, there is no way for the prototype to detect this automatically. Though error handling is much improved from earlier versions of the tool, it is likely that such an error would unceremoniously crash the program; the best that could be hoped is that the tool simply returns erroneous results.

Once all of the data is loaded into the program, the Start Analysis button will be enabled, and the program can begin its analysis. It will first initialize itself; this process will take up to two

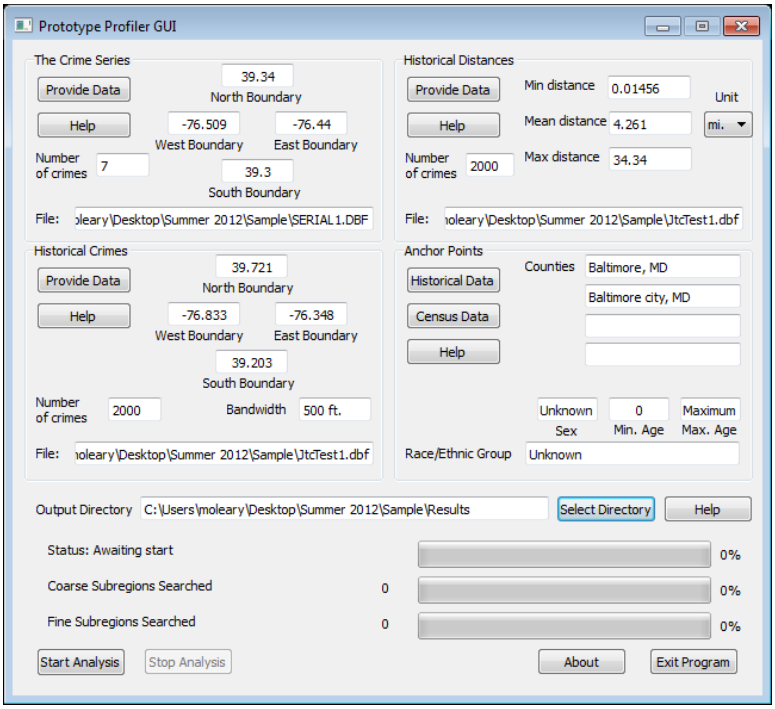


Figure 11. Software prototype ready to start

or three minutes and will be reflected in the first status bar. Once that process is complete, the prototype will begin searching through each of a number of coarse subtriangles and estimate the probability that the offender is located in that region. After the coarse analysis is complete, those coarse regions where the offender is most likely to reside will be re-analyzed at a much finer set of resolution. Progress will be recorded in the remaining status bars; see Figure 12 for an example.

Using the Prototype: A Case Study. To demonstrate the use of the prototype, let us apply it to a known data set. Ned Levine has provided sample data sets for use with the CrimeStat program; they are available at `http://www.icpsr.umich.edu/CrimeStat/download.html`. In this case study, we will use the provided sample data set for the journey to crime module (`http://www.icpsr.umich.edu/CrimeStat/files/Jtc%20Sample%20Data.zip`).

That archive contains three files. The file SERIAL.DBF contains the longitude and latitude of the crime site locations of a serial offender who has committed seven crimes in Baltimore County; we will suppose that these are the locations of the crime sites of our offender.

The file JtcTest1.dbf contains the locations of the offense site and the home location for 2000 robberies committed in Baltimore County. Because these are solved crimes, we will use this data set to calibrate our historical distances. We also use the same data set to determine the pattern of historical crimes; we set the bandwidth in the kernel density process to 500 feet.

For the prior distribution of anchor points, we will use the data from the US Census and select Baltimore County and Baltimore City as the two jurisdictions of interest. We have no demographic information about this offender, so we include none.

We enter all of this data into the program as we have already described (obtaining Figure 11) and run the analysis. The tool does take some time to run. On my (fast) workstation, it completes

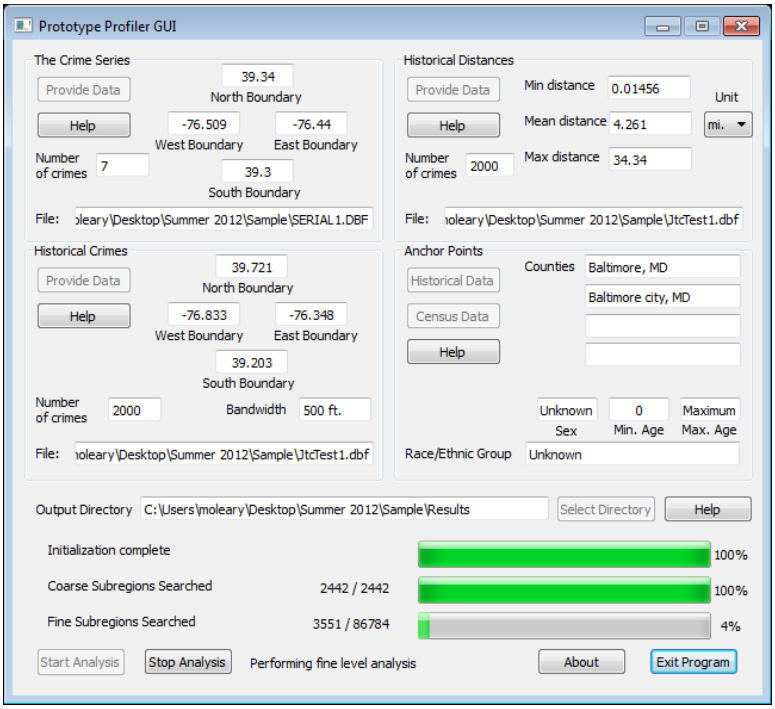


Figure 12. Analysis in progress

the initialization process in under two minutes. The coarse analysis of the data took longer at some 20 minutes; during this time it analyzed 2,442 separate subregions. Once the coarse analysis was complete, the tool refined its analysis and searched 86,784 separate fine subregions before completing its analysis, some sixteen and a half hours later.

Looking at the output directory, shown in Figure 13, we see 31 different files.

The parameter file for this run is named 2012-06-25-11-25 because I ran the program on June 25, 2012 at 11:25 in the morning. The contents of that file are:

```
Triangle Circumradius = 0.01
Crime Series Data File Name = C:\Users\moleary\Desktop\Summer 2012\Sample\SERIAL1.DBF
Crime Series Longitude Field Name = LON
Crime Series Latitude Field Name = LAT
Historical Data File Name = C:\Users\moleary\Desktop\Summer 2012\Sample\JtcTest1.dbf
Historical Crimes Longitude Field Name = INCIDX
Historical Crimes Latitude Field Name = INCIDY
Target Density Manual Bandwidth = yes
Target Density Bandwidth = 500 ft.
Historical Distances File Name = C:\Users\moleary\Desktop\Summer 2012\Sample\JtcTest1.dbf
Historical Distances Point 1 Longitude Field Name = INCIDX
Historical Distances Point 1 Latitude Field Name = INCIDY
Historical Distances Point 2 Longitude Field Name = HOMEX
Historical Distances Point 2 Latitude Field Name = HOMEY
Anchor Point Prior Distribution Data Set = census
Number of regions = 2
State = MD
County Code = 005
State = MD
County Code = 510
Race / Ethnic group = Unknown
Sex = Unknown
```

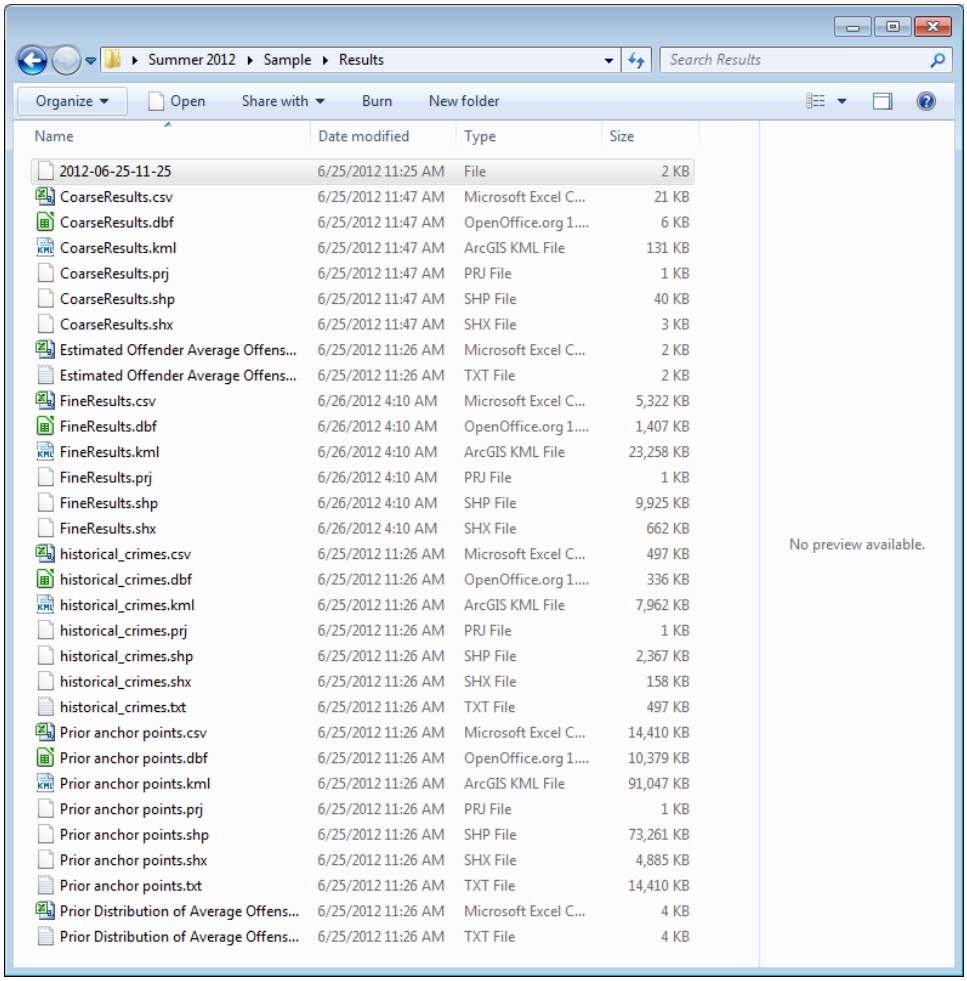


Figure 13. Program output

```
Minimum Age = 0
Maximum Age = Maximum
Results Directory  = C:\Users\moleary\Desktop\Summer 2012\Sample\Results
```

Here we can see the different choices that were made when the program was run. This record may be useful to the analyst as an archive of the settings that were used when the analysis was performed.

We also remark that it is possible to run the program without using the graphical interface by building appropriate parameter files; this feature may be of value to analysts who want to automate the process and use a scripting language to launch the analysis engine. This process will be discussed later along with other details of the prototype's internal structure.

In Figure 14 we see the file historical_crimes.kml as displayed by ArcGIS Explorer. Regions shaded red are the areas where past offenders have been most likely to offend; yellow regions are less likely, green regions less likely still, followed by unshaded regions.

Examining the map, we see a number of interesting features. The historical crime patterns are roughly what would be expected for robberies; they concentrate near the main commercial areas

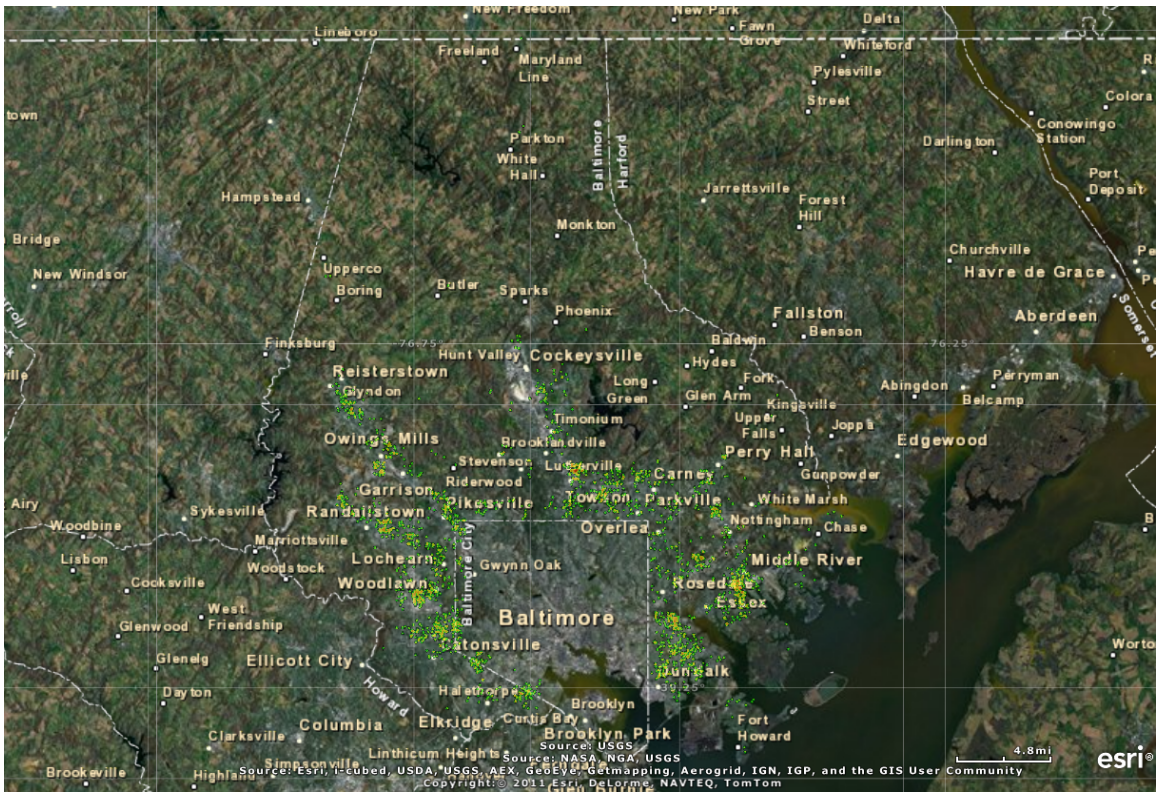


Figure 14. Case Study: historical_crimes.kml

of the county. A closer look however, shows some anomalies. Though the data is purportedly only from Baltimore county, some of the crimes appear to have taken place in Baltimore city; even worse, some of the crimes appear to have occurred in the bay. This is not a flaw with the program, but rather with the underlying data. Indeed to protect confidentiality, the locations in these data sets were randomly moved up to a quarter mile. Since we are only using this data set to demonstrate the use of the prototype, this is not a significant problem for this demonstration, however if we were actually using the program to perform an analysis, then these changes are likely to be problematic.

The same distribution of historical crimes is also returned as a shapefile, in four separate files:

- historical_crimes.dbf
- historical_crimes.prj
- historical_crimes.shp
- historical_crimes.shx

The (scaled) data are stored in the .dbf file as 16 digit numbers with eight digit precision; this should avoid both overflow and underflow. The projection for the shapefile is in the corresponding .prj file with the contents

```
GEOGCS["GCS_WGS_1984",
DATUM["D_WGS_1984",SPHEROID["WGS_1984",6378137,298.257223563]],
PRIMEM["Greenwich",0],
UNIT["Degree",0.0174532925199433295]]
```

From this we see that the output used the WGS84 datum; it should be noted that .kml files only support the WGS84 datum.

In both the .kml file and the shapefile, the underlying geographic regions are triangles. The triangles themselves are equilateral when measured in decimal degrees; when projected onto the earth's surface this results in triangles that are stretched along the north-south axis. These triangles form the fundamental spatial data structure in the internal working of the prototype; the motivation for that choice is described later, together with other elements of the program's internal structure.

In addition to the .kml file and the shapefile, the prototype also returns the files historical_crimes.csv and historical_crimes.txt. These contain the same data in a plain text format; the former separates the data elements with commas while the .txt file separates them with tabs. Each line gives the longitude then the latitude of the triangle, followed by the scaled value.

The distribution of prior anchor points is returned in a similar collection of files, including a .kml file, a shapefile, a plain text file, and a .csv file. Figure 15 shows the Prior Anchor Points.kml rendered in ArcGIS explorer. This is shaded using the same scheme as the distribution of crime sites;

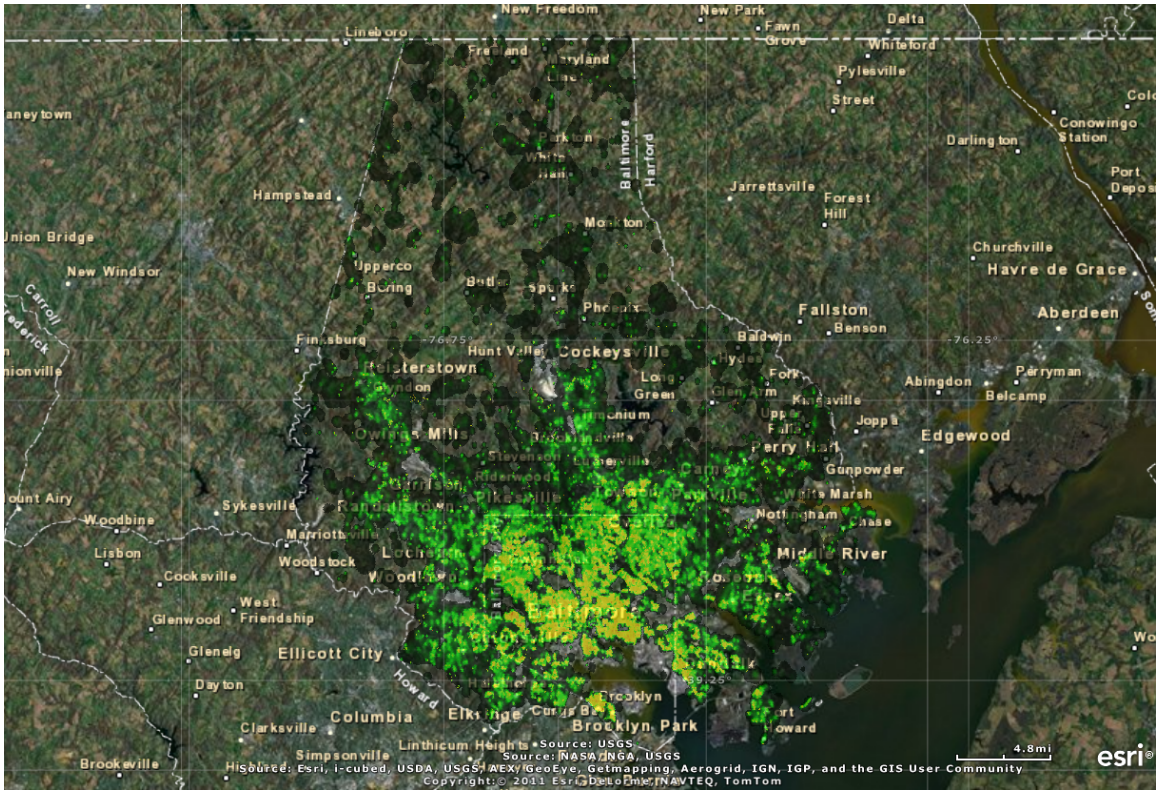


Figure 15. Case Study: Prior Anchor Points.kml (Using 2010 Census data)

regions colored red have the highest density or prior anchor points, followed by yellow, followed by green, followed by unshaded regions.

In this case study, we used the data from the 2010 US Census, and the map represents the distribution of population in Baltimore City and Baltimore County. It is important to note however, that the prototype will only look for offenders in areas where the prior anchor point density is nonzero. In particular, the tool explicitly assumed that the offender is not located in one of the adjacent counties.

We could have used the anchor points of the offenders who committed the 2000 robberies in

our data set to generate the prior distribution of anchor points. Simply select the Historical Data button from the Anchor Points box instead of the Census Data button. If we do so and specify a 1000 ft. bandwidth, we obtain the prior anchor point distribution shown in Figure 16.

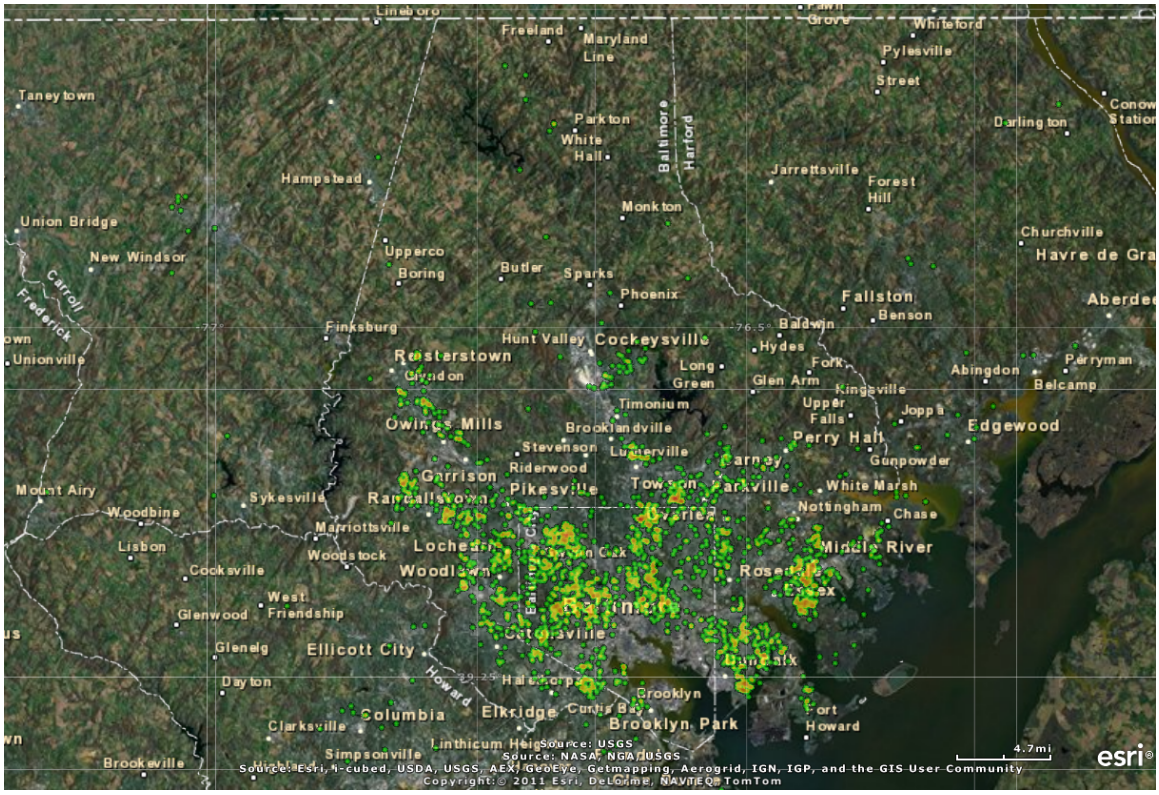


Figure 16. Case Study: Prior Anchor Points.kml (Using JtcTest1.dbf and a 1000 ft. bandwidth)

Examining this graph, we notice first that the distribution of prior anchor points is not solely contained within Baltimore City and Baltimore County. The map shows nonzero concentrations in Harford County, Howard County, and Carroll County. In fact, there are even nonzero concentrations even farther afield in areas that were cropped out from this particular figure. On the other hand, there are significant variations between the population map and the distribution of known offenders. For example, both the center of Baltimore City and the northeast corner of Baltimore City show a significant population, but relatively few offenders. This could be because the map accurately represents the fact that there are differing numbers of potential offenders in different geographic regions. It could also, however, be a limitation of the underlying data set. We do not know if the known locations of offenders is statistically biased in some fashion; we also do not know if the underlying bandwidth of 1000 feet is sufficiently large. This knowledge would need to be provided by the analyst.

Note also that, as was the case for the Census data, regions where the prior distribution are identically zero are assumed not to contain the offender. Thus since the prior distribution is zero in the downtown area of Baltimore City, the tool will not even search for an offender there.

The tool provides two sets of data that describe travel patterns each in two different formats, one as a comma separated .csv file and the second with the same content but as as a plain text tab

delimited file. The first set of data is the prior estimate of average offense distance which uses the historical data to estimate the distribution of the average distance an offender is willing to travel to offend, based solely on the data provided in the Historical Distances Box. The second piece of data is the estimated offender average distance, which is calculated from the prior distribution and the locations of the crime sites for this particular offender; it is an estimate of the average distance this particular offender is willing to travel to offend.

These data need to be interpreted differently. The prior average offense distance represents the probability that an single offender chosen from the population of all offenders will travel that average distance to offend; hence this represents the variation in the population of offenders. The estimated average offense distance however, is a probability density for the single offender under consideration who committed the crimes in the series; here the variation is caused by our lack of knowledge of the precise behavior of the offender. We would like to know not only the offender's average offense distance but also the home location; the best that the mathematics can do however is to make some estimates of the likely values.

Both of these files can be opened in a range of external programs for subsequent analysis. For example, the comma separated value .csv files can be opened in Excel and a graph generated; this is illustrated in Figure 17. The horizontal scale in the graph is spherical distance, measured in

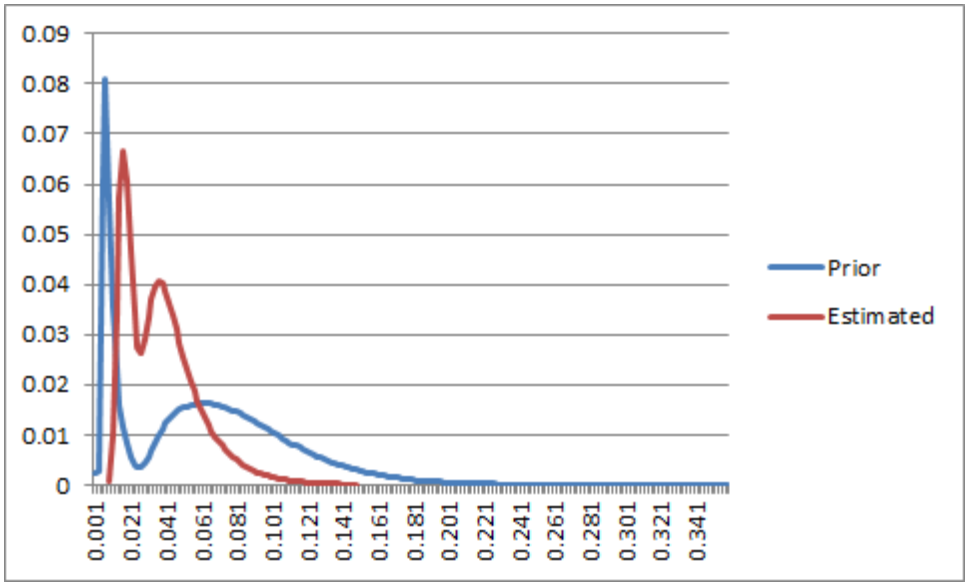


Figure 17. Case Study: Distribution of prior average offense distances and estimated average offense distance for this offender

decimal degrees. This can be converted to a conventional distance in the usual fashion; the mark at a spherical distance of 0.101 corresponds roughly to a distance of 7 miles.

The prior shows that most offenders travel a very small distance to offend. Witness the initial peak in the blue graph at roughly 0.001, which is roughly 0.07 miles; there is also a secondary peak at roughly 0.07, or 4.8 miles. On the other hand, the estimate of the offender's average offense distance is that is lies roughly in the distance range 0.01 - 0.08, equivalent to an average offense distance between 0.7 and 5.5 miles.

Finally, the key results for the analyst are the estimates of the location of the home base of

the offender produced by the prototype. Because the tool takes some time to run, the tool writes the coarse file as soon as it is complete, even as it is working to complete the fine results. Figure 18 shows the coarse results file in .kml format represented in ArcGIS Explorer.

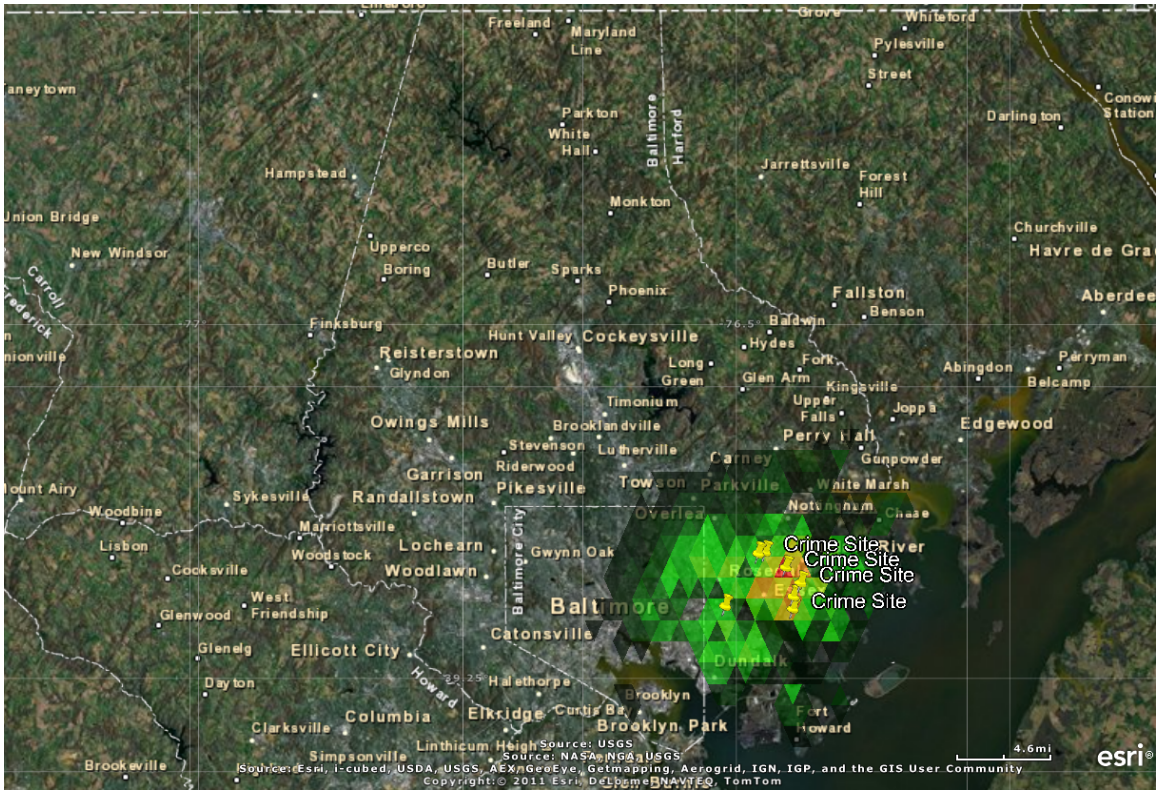


Figure 18. Case Study: CoarseResults.kml

Figures 19 and 20 show the fine results in .kml format represented in ArcGIS Explorer; the latter of these two zoomed in the region around the crime series. In all of these .kml files, regions shaded red are the areas the prototype considers most likely to contain the anchor point; yellow regions are less likely, green regions less likely still, followed by unshaded regions. The .kml files for the results also indicate the locations of the crime scenes.

The shapefiles for the results have the same structure as other shapefiles; the probability density is scaled and stored as a sixteen digit number with eight digit precision; the WGS84 projection is used for the data. However, the actual crime site locations are not included in the shapefile.

The prototype also returns a comma separated values .csv file for the results. The first six elements of each line are the longitude followed by the latitude of the three vertices that form a triangle; the seventh and last entry is the probability that the offender's home is located in that triangle. Note that the sum of all of the probabilities is 0.995 rather than the expected 1.0; this is deliberate behavior. When the prototype refines the coarse approximation to the fine approximation, it does so in order from the coarse triangle most likely to contain the offender towards the least likely triangles, stopping when the fine search area is expected to contain the offender's home base 99.5% of the time. Continuing beyond that point dramatically increases computation time to no significant benefit.

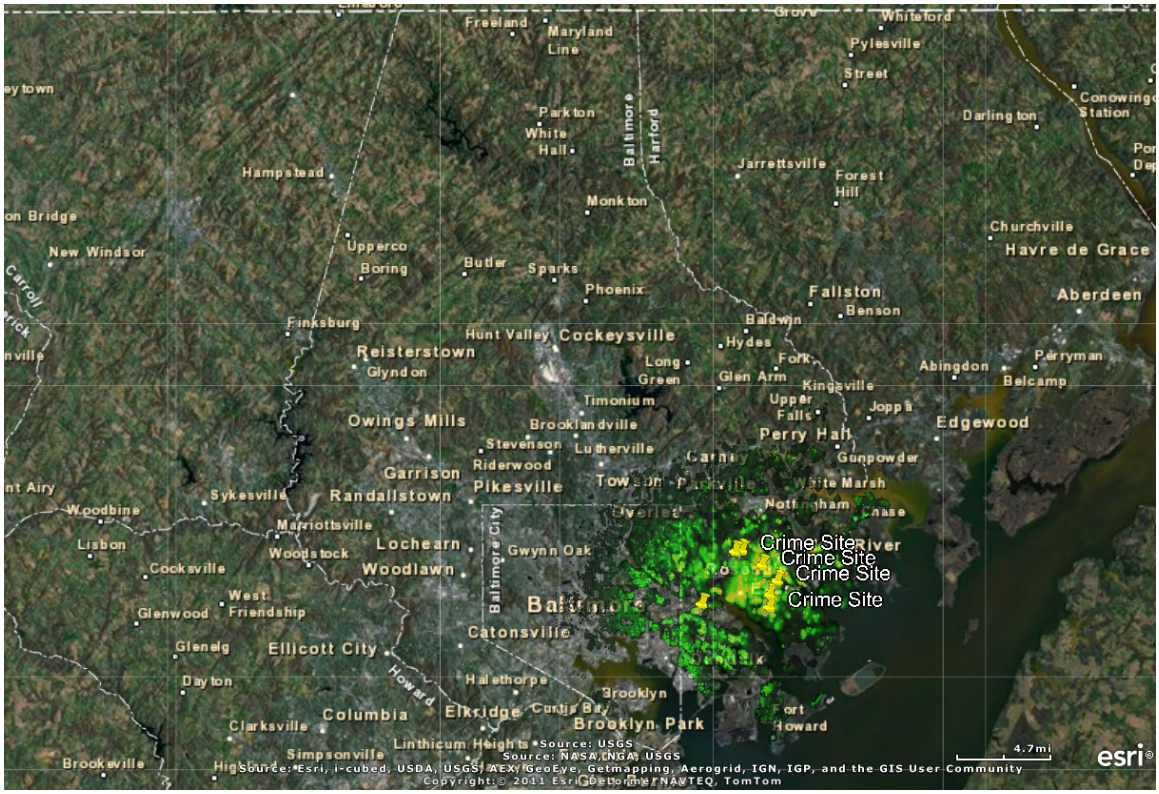


Figure 19. Case Study: FineResults.kml, using Census data for prior anchor point distribution

Looking at the scale on the maps in Figures 19 and 20, we see that some crime pairs are separated by three or four miles; this is part of the reason why the tool estimated that the offender's average offense distance is between 0.7 and 5.5 miles.

The maps show significant structure at the local level; this is caused by both the model of offender behavior and the fine resolution available in the 2010 Census data. The proposed search area avoids commercial and industrial areas near the waterfront; it also avoids the local airport and mall.

To see how the local structure in the search area is driven by the data provided to the prototype, we can re-run the prototype but rather than using 2010 Census data to generate the prior distribution of offender anchor points, suppose that we use the anchor points of the offenders who committed the 2000 robberies in our data set to generate the prior distribution of anchor points. We have already seen the resulting prior distribution in Figure 16. Running the analysis, we obtain the fine results of Figures 21 and 22; again the latter is zoomed in near the crime series itself.

Comparing the two results, it is clear that the prediction is significantly affected by the assumed prior distribution of offenders. The search area when using historical data to generate the prior is much smaller than the search area where the Census data is used to generate the prior. This, of course, is expected as a comparison of the priors themselves has already shown us that there are many areas with significant populations but without significant numbers of offenders. By using the historical data with a 1000 foot bandwidth, we have explicitly assumed that the offender for this series has an anchor point within 1000 feet of one of the other offenders in the data set. If this

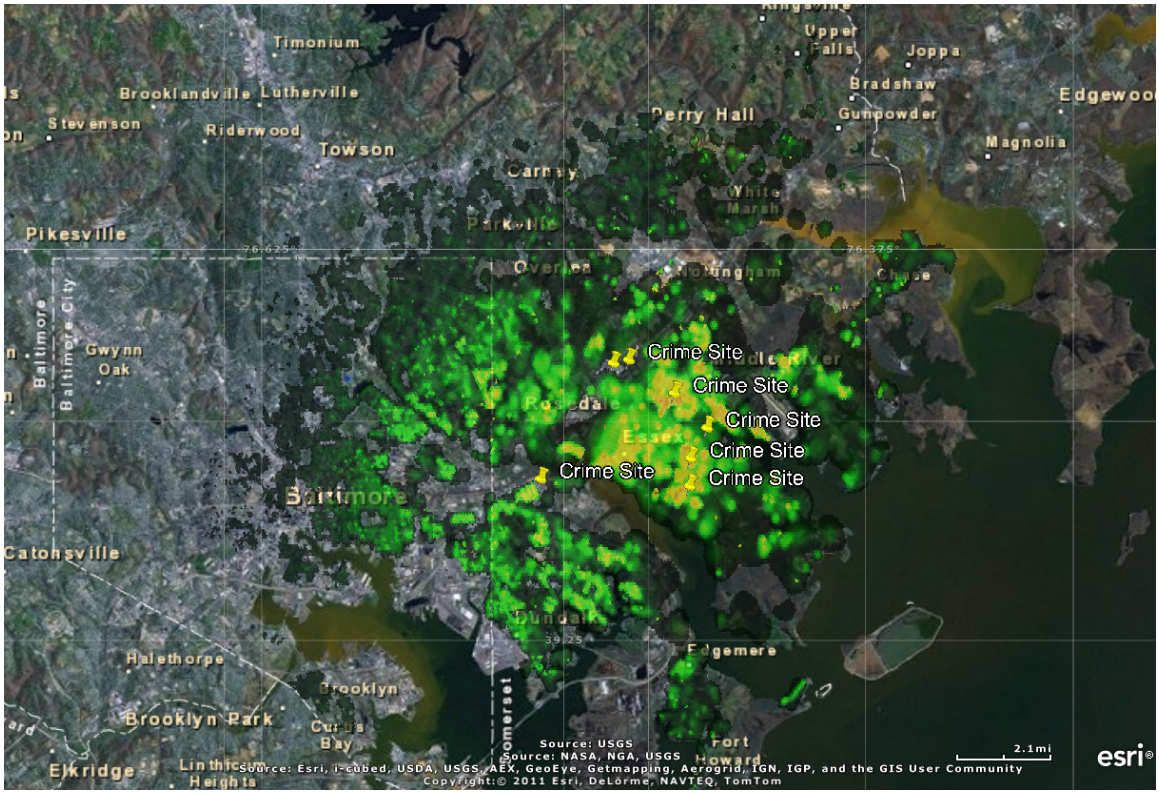


Figure 20. Case Study: FineResults.kml zoomed in near the crime series, using Census data for prior anchor point distribution

assumption is true, then the tool has quickly and correctly significantly pruned the size of the search area. On the other hand, if this assumption is false, then the offender will not be located in the search area returned by the tool and the result is in error.

A careful look at the search are in Figure 22 will show a portion of the search area in the bay. As mentioned earlier, this is not a flaw in the algorithm but rather reflects the fact that the locations in the used historical data set have been moved by up to 1/4 mile from their correct location.

Another impact of the historical prior is that it significantly reduced the computation time for this problem. Because the areas where the historical prior are assumed not to contain the offender, they do not need to be checked. While the prototype took sixteen and a half hours to generate the results using the 2010 Census data, that was cut to under five and a half hours when the historical data was used.

Analysis of the Prototype's Effectiveness

When evaluating the effectiveness of the tool, it is important to note that there are two components to effectiveness- the first is how often the search area correctly contains the offender, and the second is how large the search region actually is. Clearly these are competing factors; one can make the search area contain essentially every offender by choosing the search area sufficiently large. Similarly, an algorithm can produce tiny search areas, but if they only rarely contain the offender, then they are of little value.

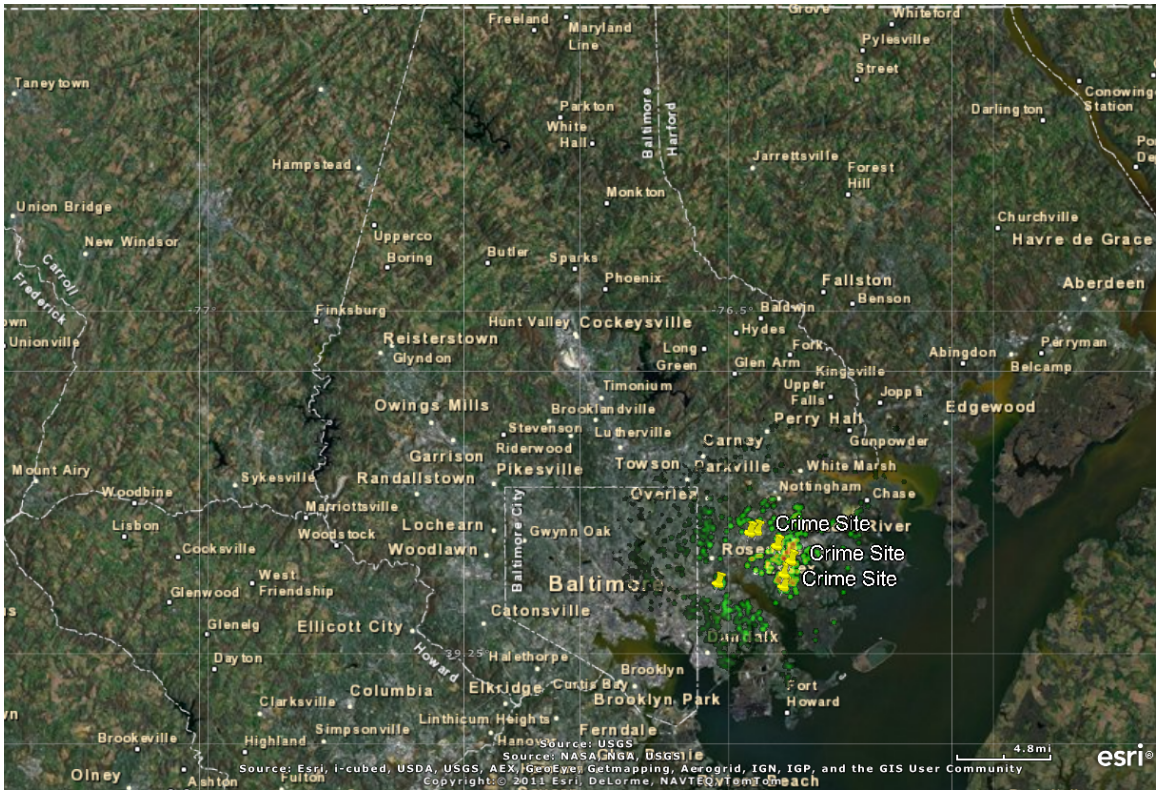


Figure 21. Case Study: FineResults.kml, using historical data for prior anchor point distribution

Using historical data as the basis of the prior distribution of offender anchor points, the prototype produced search areas comparable in size to those produced by circle theory, meaning that they are roughly the size of a circle that contains the two crimes that are farthest apart. However, the tool is much more accurate than circle theory. Tested on both residential burglary and non-residential burglary data from Baltimore County the prototype's search area contained the offender 70% (residential burglary) or 74% (non-residential burglary) of the time.

Paulsen (2006) compared different geographic profiling strategies by examining all 247 solved crime series of three or more crimes in Baltimore County between 1994 and 1997; this data set included residential burglaries as well as commercial robberies, larcenies, auto thefts, street robberies, and arsons. In his analysis, Rigel was considered correct 15% of the time and Dragnet 11% of the time; the accuracy of CrimeStat varied with the distance decay function and ranged from 5-19%. Interestingly, simply measuring a one mile circle around the mean center, median center, or center of minimum distance resulted in accuracy rates of 25-29%. On the other hand, the profile size produced by all of these analyzed methods was much smaller than the profile size produced by the prototype. The median profile size for the prototype is roughly 30 square miles, while in Paulsen's study none of the methods had an average profile area in excess of 16 square miles, and nearly all were much smaller. The centrographic measures (mean center, median center, or center of minimum distance) had a search area of roughly three square miles. See also the later analysis of Leitner and Kent (2009) who also considered some Bayesian methods.

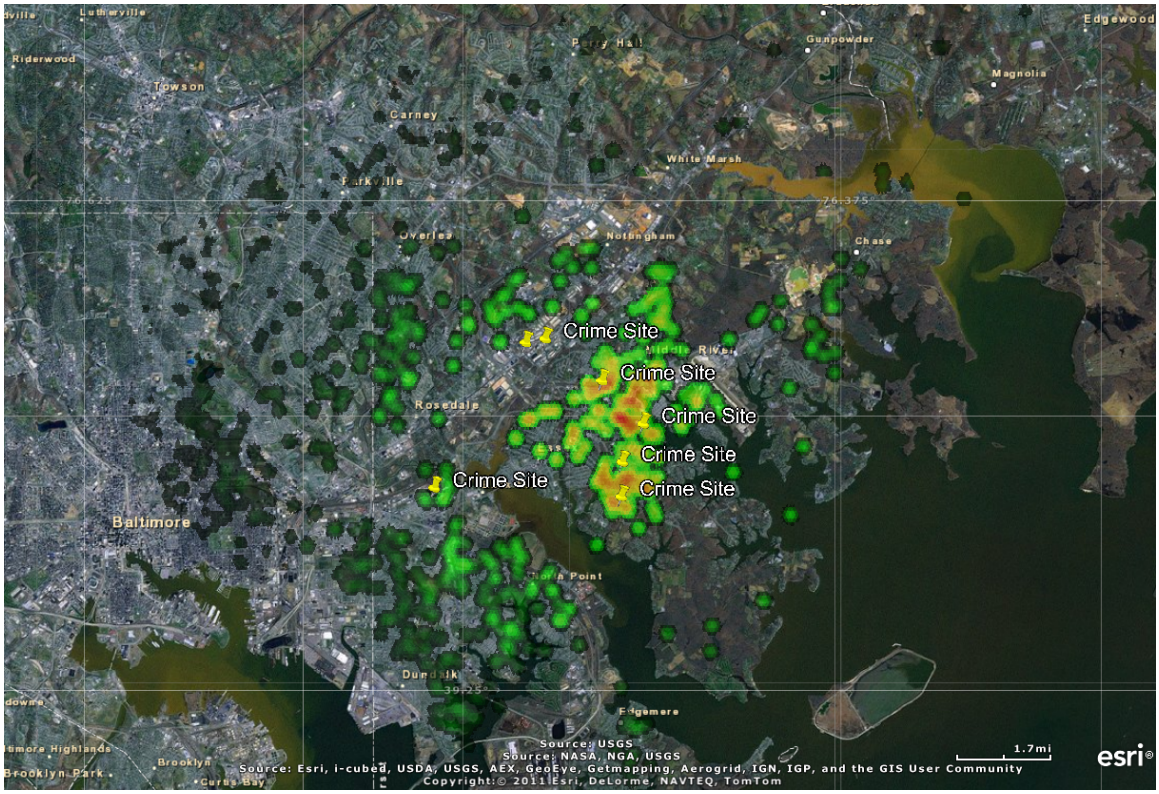


Figure 22. Case Study: FineResults.kml zoomed in near the crime series using historical data for prior anchor point distribution

Residential Burglaries. The initial data set for analysis consists of 6,217 solved residential burglaries committed in Baltimore County between 1986 and 2009. Available data about each crime include the date and location of the offense; data about the offender includes the offender's age, sex, date of birth, and location of their residence.

We identified 237 series of five or more offenses. To construct a series, we assumed that crimes committed by an offender with the same residence, date of birth, race, and sex were committed by the same offender. This method of selecting offenders may not be completely accurate. It is possible that more than one person can share their residence, date of birth, race, and sex. It is also possible that a single offender may have multiple addresses during the course of their series. However, the data contains no instance where the date of birth matched and the location of the residence did not, suggesting that these issues are not common. Note also that the location of the residence in the data set may not be the actual anchor point of the offender at the time of the series. Further, a series is identified solely as a sequence of crimes committed by the same offender. These patterns do not account for the temporal distribution of the offenses, not do they account for potentially confounding factors like the presence of additional offenders.

For each of the 237 series, the prototype was run, and the resulting geoprofile calculated. Historical data from the data set was used to generate the pattern of historical crimes, the average distance offenders are willing to travel, and the prior geographic distribution of offender anchor points. To avoid any potential confounding of results caused by the fact that all of the crimes in the

data set were solved, the locations of only some of the offenders were considered to be known.

To generate the pattern of historical crimes, all 6,217 crime locations were used, as the crime locations are known whether or not the offender has been identified. On the other hand, for each run we examined each offender other than the one who committed the series under consideration. A random choice was then made; 25% of the time the data for that offender was retained and that offender was considered "known". The other 75% of the time, the data for that offender was not included in the development of either the prior distribution of the offender's anchor points or in the distribution of the historical distances. In this fashion, we ensured that the prototype did not have complete data about the distribution of crimes; it also had absolutely no data from the series under study other than the locations of the offenses themselves.

The bandwidth for the pattern of historical crimes was set at 500 feet, while the bandwidth for the prior distribution of offender anchor points was set at one mile. These were selected after preliminary analysis showed that they were sufficiently large to give prior maps that were representative of the full data set.

The 237 series break down into 114 marauders and 123 commuters. This split into roughly half commuter and half marauders is in line with expectations observed for burglary patters; see for example the work of Kocsis and Irwin (1997), Kocsis et al. (2002), Meaney (2004), or Sarangi and Youngs (2006). The offender's home was within the search area provided by the prototype in 165 of the 237 crime series; including 107 of the 114 marauders (94%) and 58 of the 123 commuters (47%). Correctly identifying the great majority of the marauders is not a significant theoretical or practical advance, as that has already been accomplished by Circle Theory. On the other hand, correctly identifying nearly half of the commuters is very significant.

Of course, creating a search area that contains the offender is only half of the issue; the second half is the size of the resulting search area. Figure 23 shows the size of the search area (in square

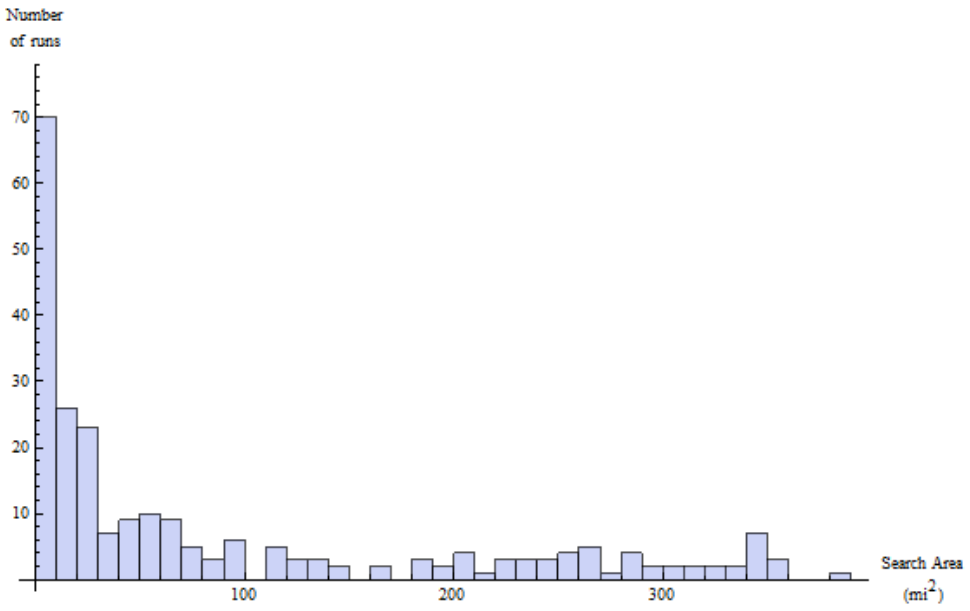


Figure 23. Distribution of Geoprofile Search Area for 237 Serial Burglary Series in Baltimore County

miles) of the geoprofiles returned for all 237 runs. Examining the graph, we see that nearly half of the runs, 119 in all, have a total search area of less than 30 square miles.

To properly understand this graph, let us compare it to the corresponding search area predicted by Circle Theory. In Figure 24, we plot the area of the circle whose diameter is formed by the segment connecting the two crimes that are farthest apart. Not only does this figure give a

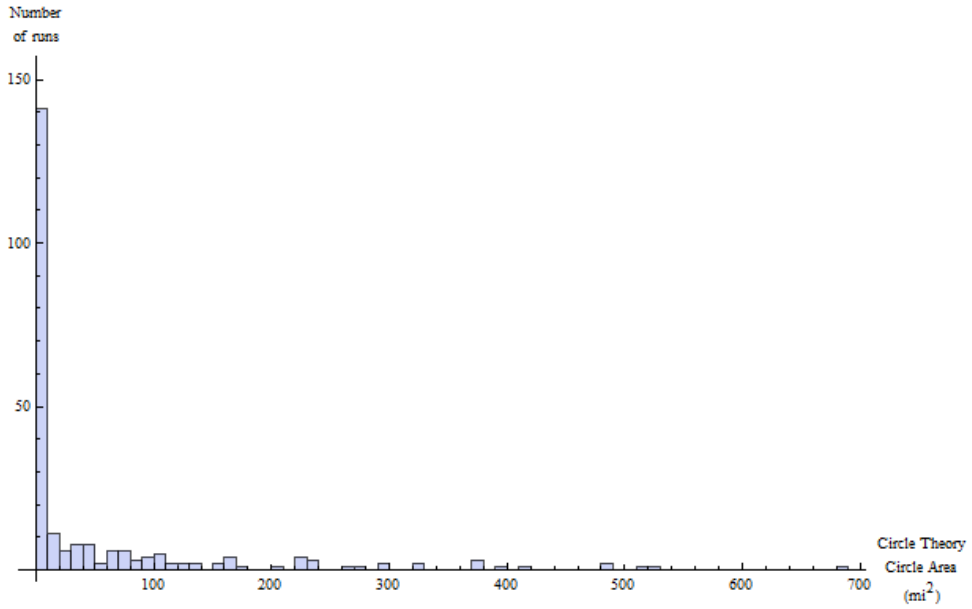


Figure 24. Distribution of Area of the Canter Circle for 237 Serial Burglary Series in Baltimore County

graph of the area of the circle used to determine if the offender is a commuter or a marauder, it also provides an estimate of the overall size of the crime series. Two facts jump out from Figure 23. First is that, at least for many series, the overall size of the Canter circle is quite small; in 141 cases that circle has an area of no more than 10 square miles. On the other hand, the tail is also much longer, and some of the circles are almost twice as large as the largest search area produced by the prototype.

To better compare the search areas, Figure 25 plots the search area for both methods together in the same graph; here the search area is along the vertical axis and the number of series crimes on the horizontal axis. To improve readability, the search area was truncated to 250 square miles or less.

The prototype generally has slightly larger search areas than those produced by Circle theory. However, most of the variation is in the very small search regions; the distribution of larger search regions are comparable. Essentially the prototype is less likely to generate very small search regions, while the distribution of moderate and large search regions is comparable to what is generated by circle theory.

To better compare the individual search areas rather than the aggregates, Figure 26 has been developed. For each of the 237 crime series, a point was plotted; the horizontal axis gives the size of the search area generated by the prototype while the vertical axis the size of the search area produced by circle theory. This figure clearly confirms our observation that when circle theory gives

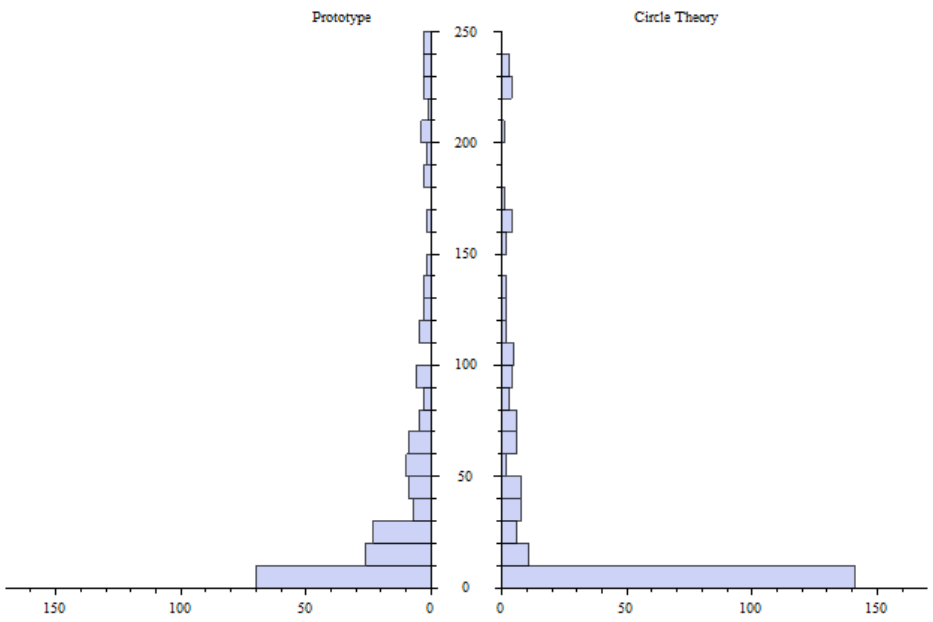


Figure 25. Comparing the Distribution of the Geoprofile Search Area to the Area of the Canter Circle for 237 Serial Burglary Series in Baltimore County

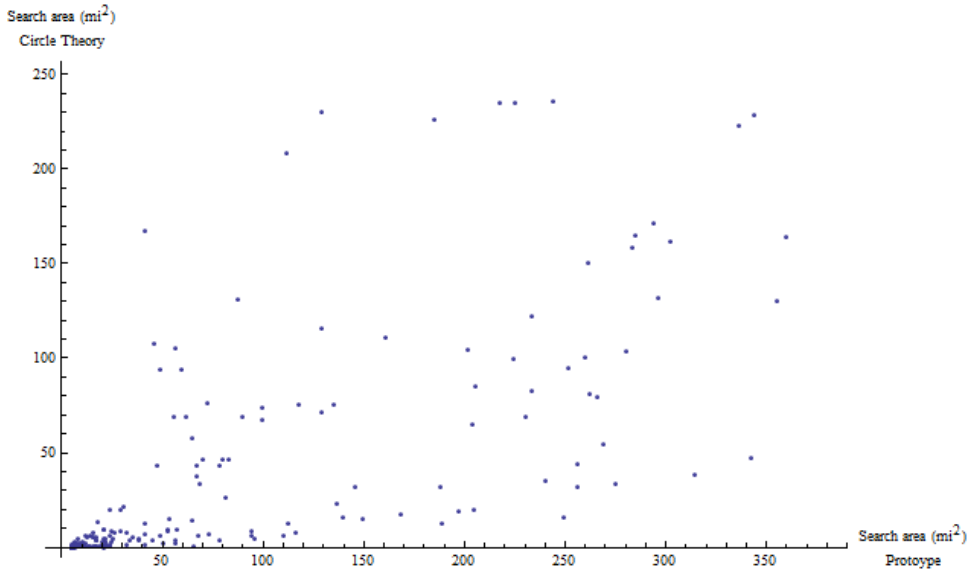


Figure 26. Prototype Geoprofile Search Area Versus Canter Circle Area, for each of 237 Serial Burglary Series in Baltimore County

a very small search area, the prototype gives a somewhat larger search area. On the other hand, there appears to be little correlation between the search areas produced when either are large.

So far we have discussed how often the offender's anchor point lies in the full search area produced by the prototype and compared that area to the area produced by Canter's circle theory. However, unlike circle theory, the prototype produces a prioritized search area; when the geoprofile is used it is expected that the search would proceed from the regions considered most likely to contain the offender to the regions least likely. Rather than looking solely at the total search area, consider the portion of the search area that is ranked at the same or a higher priority than the location that contains the offender. Doing so, we obtain Figure 27 which considers the 163 runs where the prototype's search area contained the offender.

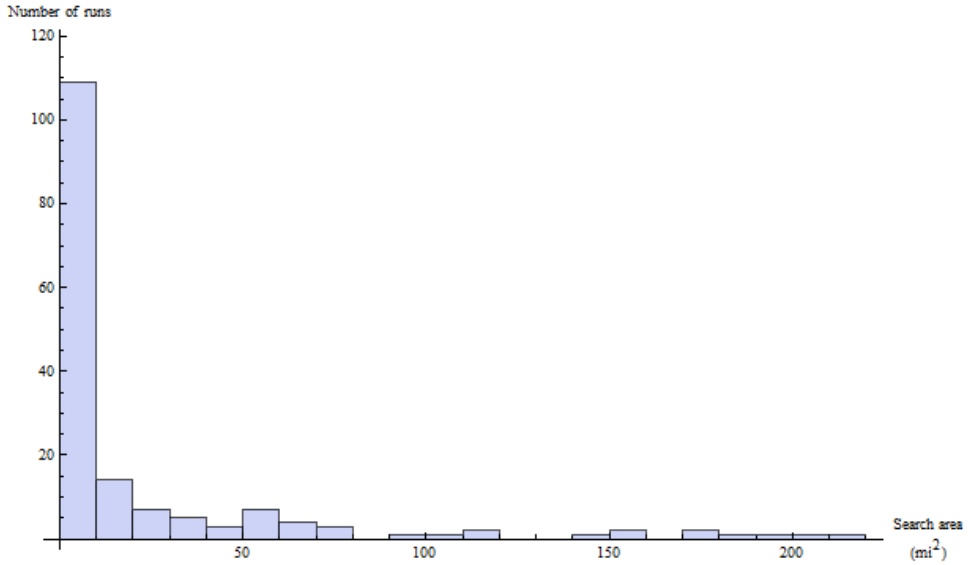


Figure 27. Size of the Portion of the Search Area at Least as Likely as the Actual Offender's Home Location, for 163 Serial Burglary Series in Baltimore County where the Prototype's Search Area Contains the Offender's Home Location

Here we see that in 109 of the 163 successful runs, the prototype located the offender within a search area of 10 square miles or less. In only a small handful of cases was the offender located in the far reaches of the search area.

The prototype does not just return a prioritized search area; mathematically it returns a probability density function. In particular, for any geographical region S, the prototype returns an estimate of the probability that the offender lies within the region S

Given a geographic point h, we can then find the percentile rank of the probability density of that point. To do so, form the set S consisting of all geographic points whose probability density is greater than or equal to the value of the probability density at h. The percentile rank of the probability density at h is the probability that the offender lies within that resulting set S, interpreted as a percentage.

If the model is correct, then we would expect that the percentile rank of the probability density of the actual anchor point is 5% or less roughly 5% of the time; similarly the percentile rank of the probability density of the actual anchor point should be between 5% and 10% another 5% of the

time and so on. To compare this theoretical prediction with our model, we present Figure 28 which is a histogram of the percentile rank of the probability density of the actual anchor point for all of the 163 runs where the search area contained the offender's anchor point.

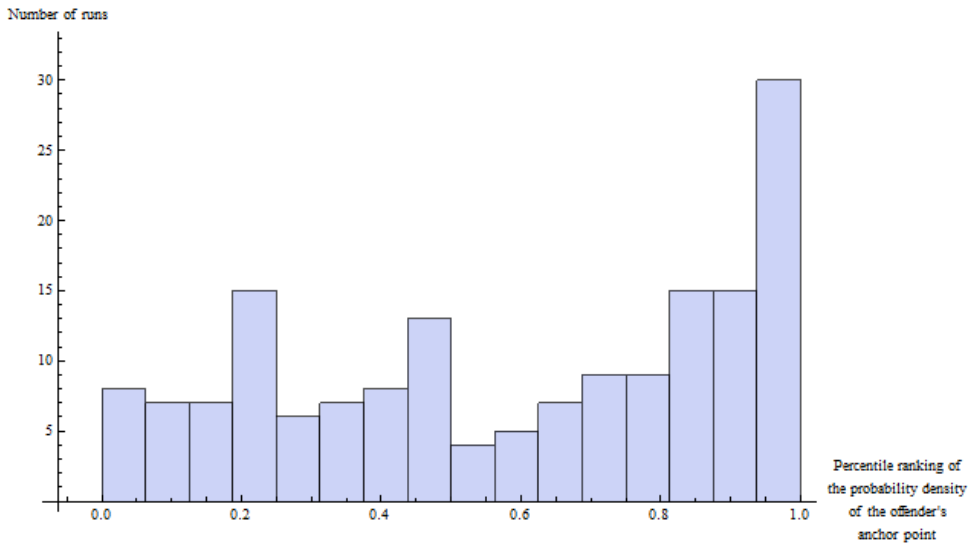


Figure 28. Histogram of the Percentile Rank of the Probability Density of the Actual Offender's Anchor Point for 163 Serial Residential Burglary Series in Baltimore County where the Prototype's Search Area Contains the Offender's Home Location

The observed data well matches our prediction in that it is nearly constant across all of the equally sized bins. The major exception is that there are far more entries in the last bin, where the percentile rank is between 93.75% and 100%.

If this observation were made for all 237 runs, rather than just the 163 runs where the search area contained the offender's anchor point, then we would have compelling evidence that the model was correctly producing the probability density function for the location of the offender's anchor point. The fact that the model missed 30% of the offenders however tells us that there is clear improvement that needs to be made in the model. One possible explanation for the observed behavior is that the model is roughly correct some fraction of the time, but dramatically wrong other times; this would be consistent with the observed behaviors. One possible mechanism for this would be if some fraction of the home addresses recorded in the data set did not actually agree with the anchor point of the offender during the time the crimes were committed.

The total run time of the prototype is long. Figure 29 plots a histogram of the total run time for the program. One fact jumps out immediately; the tail is long with six runs taking longer than 200 hours, and one taking 400 hours to complete. That said, the great majority of the runs are much faster. Zooming in on the histogram in Figure 30, we see that 96 of the 237 runs were completed in less than four hours, 119 in less than eight hours, and 155 in less than 24 hours.

Non Residential Burglaries. To cross validate the results seen for residential burglaries, we performed a similar analysis on non-residential burglaries committed in Baltimore County. The data set consists of 2,650 solved non-residential committed between 1989 and 2008. As in the case of residential burglaries, data consisted of the date and location of the offense as well as the race,

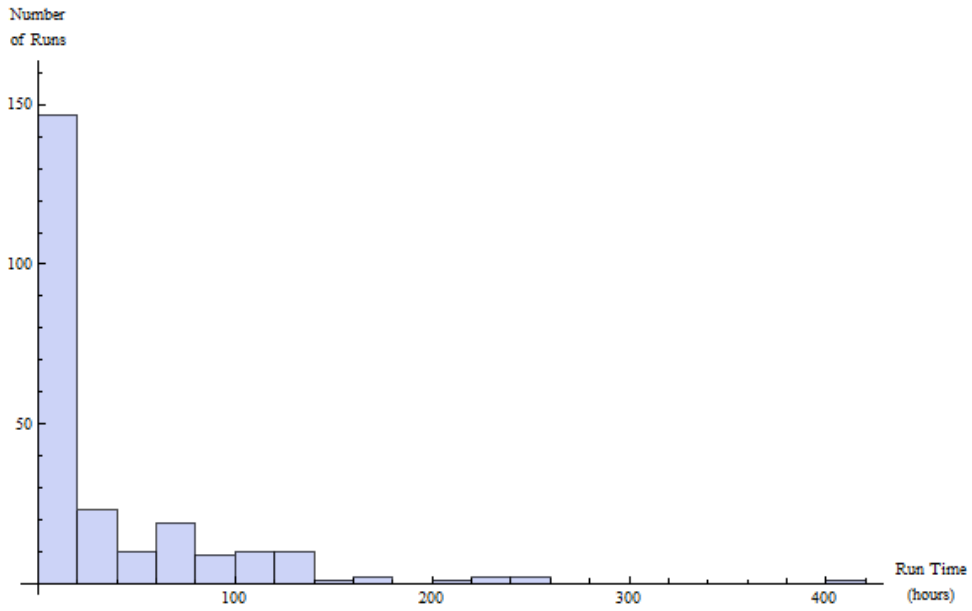


Figure 29. Histogram of the Run Times for the Prototype on 237 Serial Burglary Series in Baltimore County

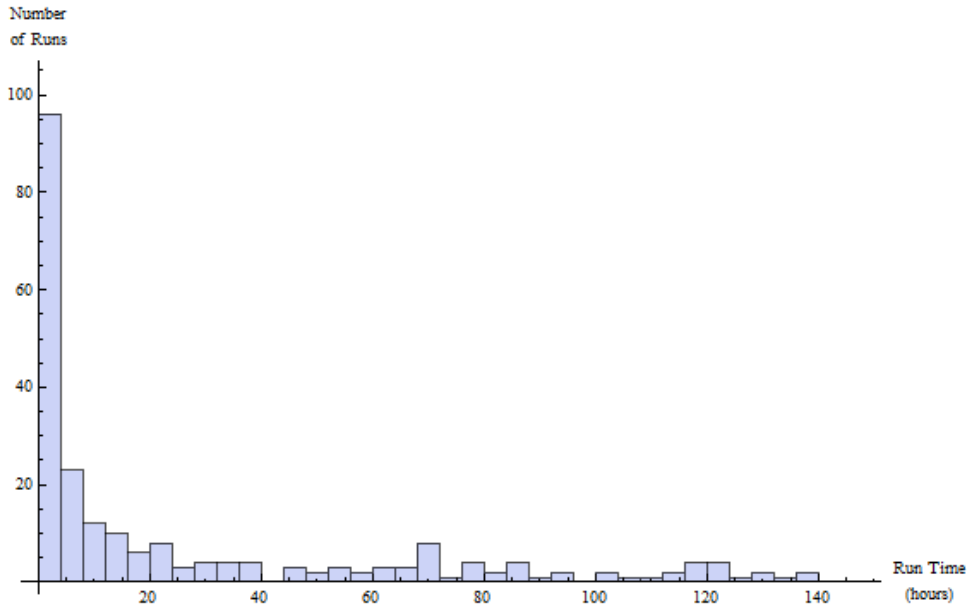


Figure 30. Zoomed in Histogram of the Run Times for the Prototype on 237 Serial Burglary Series in Baltimore County

sex, date of birth, and the location of the offender's home. Series were identified as a collection of crimes committed by someone with the same sex, race, date of birth, and home location.

All 2,650 crimes were used to generate the prior distribution of crimes. For each run, only a sample of offenders were treated as known. There was a 25% chance that an offender would be considered "known" and their home location used to generate the prior distribution of anchor points and the prior distribution of average offender distances. The actual offender for the series under consideration was, of course, excluded.

The bandwidth for the historical crimes remained at 500 feet and the bandwidth for the prior distribution of offender anchor points remained at one mile.

Overall, the data set contained 38 commuters and 36 marauders. The search area for the prototype contained the home of the offender for 66% (25 of 38) of the commuters and 83% (30 of 36) of the marauders, for a combined accuracy rate of 74%, (55 of 74) even better than the 70% observed for residential burglaries.

The size of the search areas produced for non-residential burglaries are comparable to those produced for residential burglaries (Figure 31); they are also comparable to the size of the corresponding Canter circles as seen in Figure .

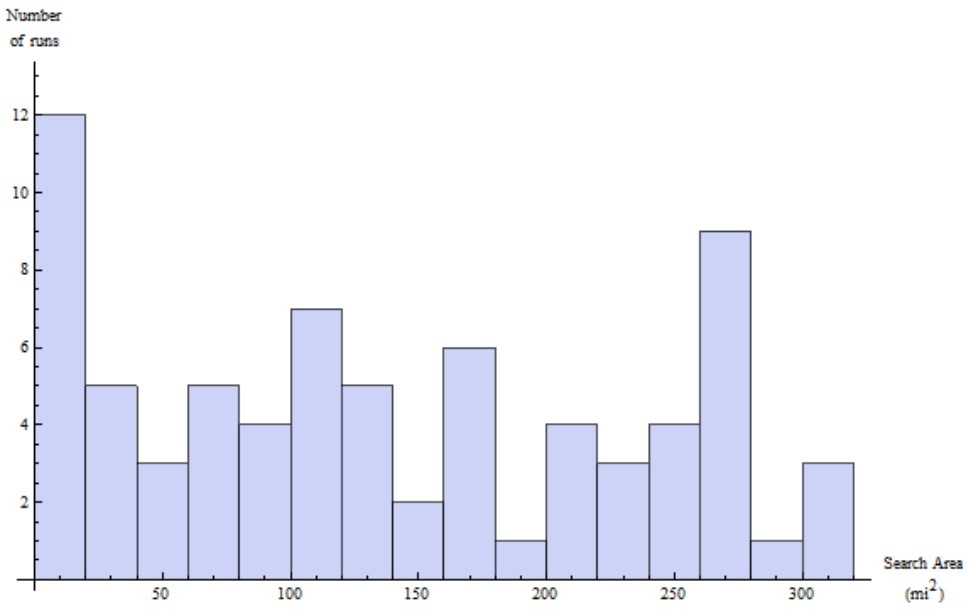


Figure 31. Distribution of the Geoprofile Search Area for 74 Non-Residential Burglary Series in Baltimore County

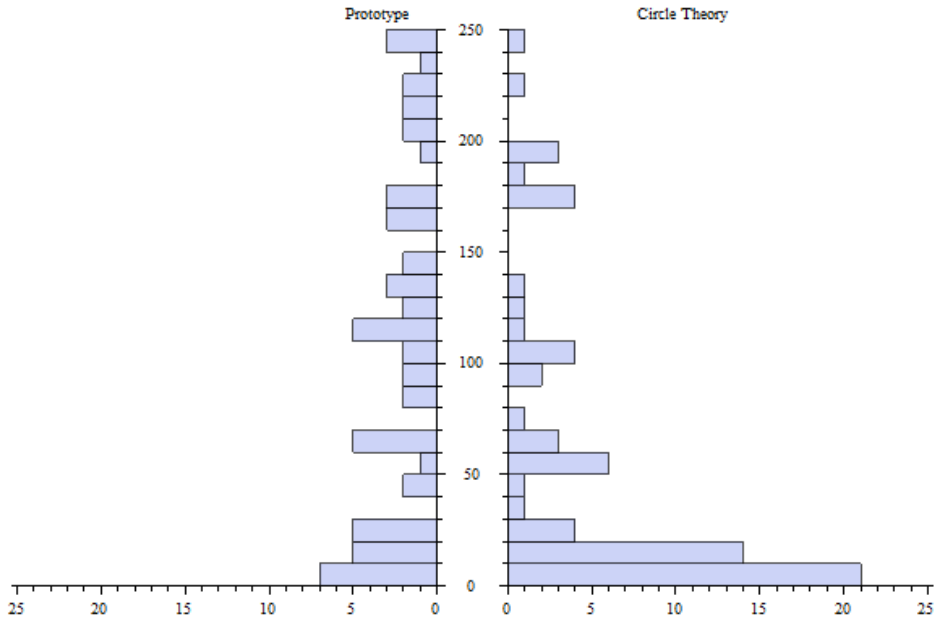


Figure 32. Comparing the Distribution of the Geoprofile Search Area to the Area of the Canter Circle for 74 Non-Residential Burglary Series in Baltimore County

Further, as we saw for the residential burglary data, when we look only at the portion of the prioritized search area that needed to be examined before the offenders home is located, then the search area is even smaller. As seen in Figure 33, 37 of the 55 successful searches find the offender in prioritized search areas smaller than 30 square miles, with 26 smaller than 10 square miles.

Looking at the percentile ranking of the probability density of the offender's anchor point, if our theory is correct, we would expect to see a distribution that is roughly constant. Examining the result (Figure 34), we observe the same behavior seen for residential burglary, namely a histogram that is mostly constant but one that has an outsize peak at the highest percentile range. The explanation of this behavior remains- clearly there are improvements to be made with the model.

The distribution of run times for the non-residential burglaries (Figure 35) is similar to what was observed for residential burglaries, save that the most extreme elements of the tail are no longer present. The longest run time was just over 163 hours; there were eight residential burglary series that took even longer to be analyzed, the longest of which took over 400 hours to complete.

The Prototype's Internal Structure

The software prototype actually consists of two separate programs: a graphical user interface and a separate analysis engine. The analysis engine handles all of the scientific analysis; it uses various files as input, and produces the various output files. The graphical user interface prompts the user for the required data elements; when they are all specified it will then write a parameter file for the analysis engine, then call the analysis engine and wait for the result.

One advantage of this architecture is that is separates the different functions into separate programs that can each be maintained or customized separately. This modular form also allows the analysis engine to be run without the graphical user interface; this feature is particularly valuable

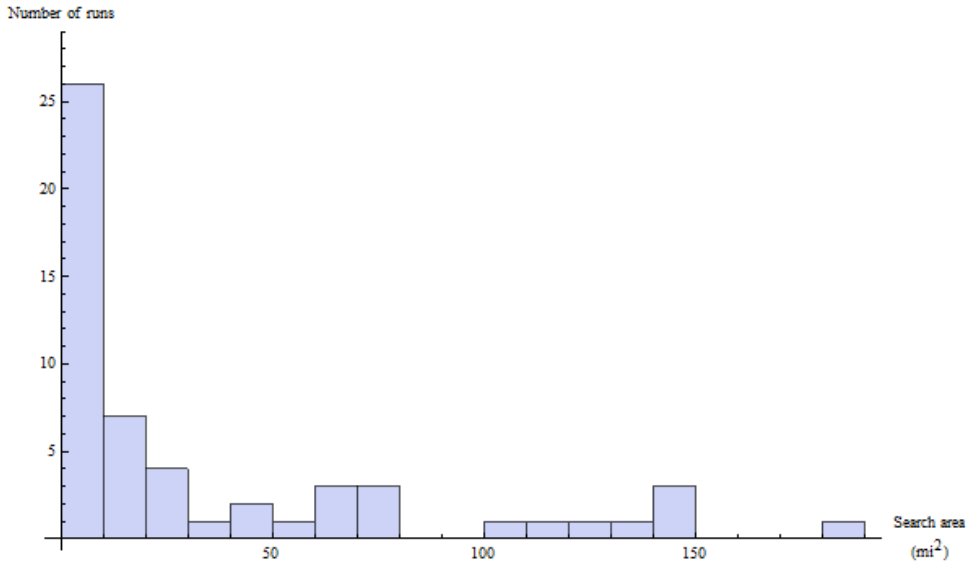


Figure 33. Size of the Portion of the Search Area at Least as Likely as the Actual Offender's Home Location, for 55 Serial Non-Residential Burglary Series in Baltimore County where the Prototype's Search Area Contains the Offender's Home Location

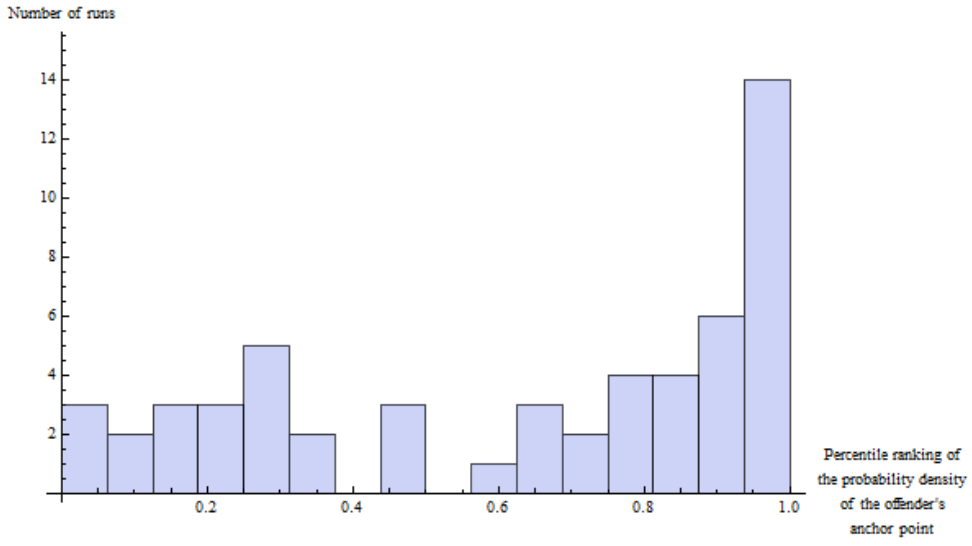


Figure 34. Histogram of the Percentile Rank of the Probability Density of the Actual Offender's Anchor Point for 55 Serial Non-Residential Burglary Series in Baltimore County where the Prototype's Search Area Contains the Offender's Home Location

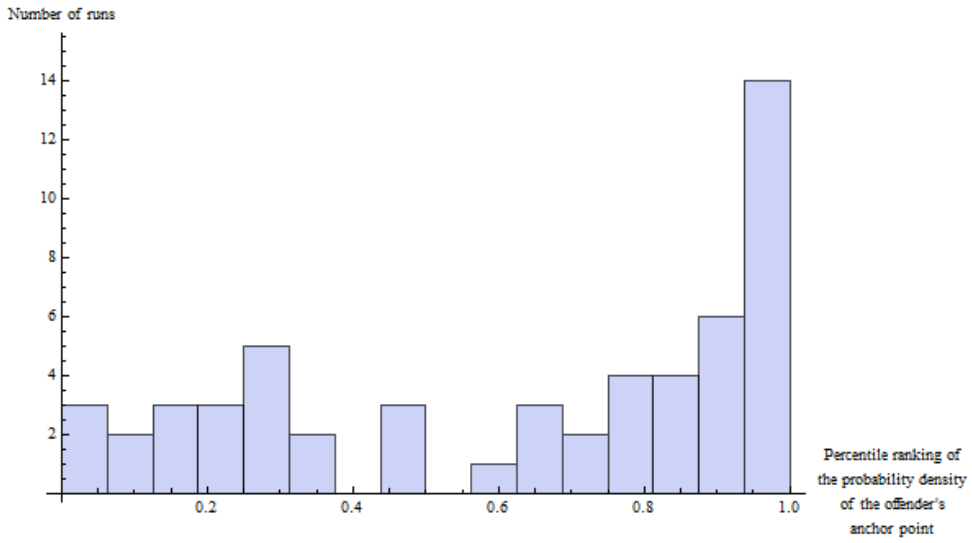


Figure 35. Histogram of the Run Times for the Prototype on 74 non-Residential Burglary Series

when large numbers of similar runs need to be made, as the program can simply be called by another script.

The Command Line Analysis Tool. The command line analysis tool requires as input:

- The locations of the elements of the crime series
- The locations of historical crimes of the same type as the series under consideration. These are used to estimate the attractiveness of a location to an offender using hot spot methods; the bandwidth in the hot spot method can be specified manually and by default it uses the mean nearest neighbor distance.
- The locations of both the crime location and the anchor point of the offender for a series of representative crimes. This is used to calibrate the distance decay behavior used in the model.
- A method to estimate the prior distribution of potential offender anchor points before the crime series is analyzed. There are two choices- either to use data from the 2010 US Census, or to provide the locations of anchor points for historical offenders. In the latter case, kernel density methods are used to develop the distribution over the geographic region using a bandwidth that is either manually specified or equal to the mean nearest neighbor distance between offender anchor points.
- The directory in which the results will be stored.

These data are provided via a plain text parameter file, described below. When launched, the program begins with a number of preliminary analyses, including

- Developing an estimate of the distribution of the average distance (in decimal degrees following a great circle of the earth) that offenders are willing to travel. It is assumed that each offender has a distance decay function that can be well modeled with a Rayleigh distribution, but that different offenders can have different average offense distances. The results of this analysis are provided in the files `Distribution of Average Offense Distance.csv` and `Distribution of Average Offense Distance.txt`.
- Estimating the relative attractiveness of different geographic locations for this crime type. Maps are provided in Shapefile format and in .kml format, as well as providing a .csv file with the

relative values.

• Estimating the distribution of anchor points in the geographic region, before any information about the location of the crime series is used. Maps are provided in Shapefile format and in .kml format, the relative values are also provided in a .csv file.

The program then performs a coarse search over the entire geographic region to find the location(s) most likely to contain the anchor point of the offender. The results of this coarse search are then provided in Shapefile format, in .kml format, as well as in a .csv file of the probabilities.

Once the coarse search completes, those areas that are most likely to contain the offender's anchor point are searched again, but now using a much finer geographic resolution. The results of this fine search are also provided in Shapefile format and in .kml format, as well as in a .csv file of the probabilities.

The program then terminates.

The Parameter File. The program uses a two stage process to obtain its input. First it reads a parameter file; this contains the names of files that contain the required data elements as well as the directory to place the program results.

The name of the parameter file can be passed as the argument of the program; if no parameter is passed, then the program looks for the parameter file with name `./Parameters/Parameters.txt`.

The directives in the parameter file must be given in the order specified below.

Triangle Circumradius. The geographic region under study is subdivided into a mesh of equilateral triangles (the coarse mesh), and each of these triangles is then subdivided into 256 subtriangles, resulting in the fine mesh. The size of the triangles in the coarse mesh is specified by this directive. This is measured in decimal degrees along a great circle of the earth, and is usually kept at the default value of 0.01.

```
Triangle Circumradius = 0.01
```

Crime Series Data File Name. The next required directive(s) describe the file that contains the locations of the elements of the crime series. Like all data, locations are specified using longitude and latitude in decimal degrees, using the WGS-84 reference. The program can take this data in three different formats: a plain text file, a .csv file, or as a shapefile. If a plain text file is used, then the parameter file needs to contain a line like

```
Crime Series Data File Name = C:\Users\Analyst\Desktop\CrimeData\Series.txt
```

If the data is contained in a .csv file, the line should read

```
Crime Series Data File Name = C:\Users\Analyst\Desktop\CrimeData\Series.csv
```

The more complex situation is when the data is contained in a shapefile. In this case the parameter file needs three lines, one with the name of the corresponding .dbf file that contains the data and two lines to specify the field names.

```
Crime Series Data File Name = C:\Users\Analyst\Desktop\CrimeData\Series.dbf
Crime Series Longitude Field Name = INCIDX
Crime Series Latitude Field Name = INCIDY
```

Historical Crime Data File Name. The next required element in the parameter file is information about the file that contains the longitude and latitude of historical crimes of the same type as the series under study. As always, locations are specified using longitude and latitude in decimal degrees, using the WGS-84 reference. The program can take this data in three different formats: a plain text file, a .csv file, or as a shapefile. If a plain text file is used, then the parameter file needs to contain a line like

```
Historical Data File Name = C:\Users\Analyst\Desktop\CrimeData\HistoricalCrimes.txt
```

If the data is contained in a .csv file, the line should read

```
Historical Data File Name = C:\Users\Analyst\Desktop\CrimeData\HistoricalCrimes.csv
```

The more complex situation is when the data is contained in a shapefile. In this case the parameter file needs three lines, one with the name of the corresponding .dbf file that contains the data and two lines to specify the field names.

```
Historical Data File Name = C:\Users\Analyst\Desktop\CrimeData\HistoricalCrimes.dbf
Historical Crimes Longitude Field Name = INCIDX
Historical Crimes Latitude Field Name = INCIDY
```

Target Density Bandwidth. The historical crime data is used to generate a map of target attractiveness using a kernel density parameter estimation technique. The bandwidth can either be specified automatically as the mean nearest neighbor distance of the crime data, or it can be specified manually. In the first case, the parameter file must next contain the directive

```
Target Density Manual Bandwidth = no
```

To manually specify the bandwidth, the parameter file must instead contain directives in the general form

```
Target Density Manual Bandwidth = yes
Target Density Bandwidth = 500 ft.
```

The bandwidth can be specified in feet (ft.), miles (mi.), meters (m), or kilometers (km).

Historical Distances File Name. The next required directive(s) describe the files that contain the locations of both the crime site and the anchor point for a collection of solved crimes; these data are used to determine the distribution of the average offense distance for offenders in the jurisdiction. The data can be provided as a plain text file, as a .csv file, or as a shapefile.

```
Historical Distances File Name = C:\Users\Analyst\Desktop\CrimeData\HistoricalDistances.txt
```

If the data is contained in a .csv file, the line should read

```
Historical Distances File Name = C:\Users\Analyst\Desktop\CrimeData\HistoricalDistances.csv
```

The more complex situation is when the data is contained in a shapefile. In this case the parameter file needs five lines, one with the name of the corresponding .dbf file that contains the data and four lines to specify the field names.

```
Historical Distances File Name = C:\Users\Analyst\Desktop\CrimeData\HistoricalDistances.dbf
Historical Distances Point 1 Longitude Field Name = INCIDX
Historical Distances Point 1 Latitude Field Name = INCIDY
Historical Distances Point 2 Longitude Field Name = HOMEX
Historical Distances Point 2 Latitude Field Name = HOMEY
```

Prior Anchor Point Data. The next required directives specify how the program is to estimate the prior distribution of offender anchor points. The two possible ways are either to use data from the 2010 US Census, or to use kernel density parameter estimation on a set of locations of known offender anchor points.

A typical set of directives in the parameter file that use the US Census data is

```
Anchor Point Prior Distribution Data Set = census
Number of regions = 2
State = MD
County Code = 510
State = MD
County Code = 005
Race / Ethnic group = Unknown
Sex = Unknown
Minimum Age = 0
Maximum Age = Maximum
```

There is no limit on the number of regions that can be specified. The graphical user interface limits users to four, but that is a limitation solely of the graphical user interface; the command line analysis program will happily run as many as are specified.

For each region, there needs to be a corresponding state and county code directive. The value for the state field is the usual two letter abbreviation for the state. The county code is the three digit U.S. Census Bureau code (see `http://www.census.gov/geo/www/ansi/download.html`); this does not include the corresponding two digit state code.

For the Race/Ethnic group directive, allowable values are:

- Unknown
- White
- Black or African American
- American Indian or Alaska Native
- Asian
- Native Hawaiian or Pacific Islander
- Two or more races
- Hispanic or Latino
- White, not Hispanic or Latino

For the Sex directive, allowable values are:

- Male
- Female
- Unknown

For the Minimum Age directive, only the following ages are valid:

- 0,5,10,15,18,20,21,22,25,30,35,40,45,50,55,60,62,65,67,70,75,80,85

while for the Maximum Age directive only the following ages are valid:

- 4,9,14,17,19,20,21,24,29,34,39,44,49,54,59,61,64,66,69,74,79,84,maximum

The reason for these values is the fact that block level Census data with information about the age, sex, and race/ethnic group of the population are only available in certain age ranges.

The second option is to use historical data to generate the prior distribution of the offender's anchor point. This is accomplished in a similar fashion as other historical data sets. The data themselves can be in plan text, .csv, or shapefile formats. The prior is then generated via kernel density parameter and the bandwidth can be specified manually or automatically.

To indicate that historical data is used to generate the prior, start with the directive

```
Anchor Point Prior Distribution Data Set = manual
```

For a plain text file or a .csv file, then just specify the name

```
Anchor Point Prior Distribution File = C:\Users\Analyst\Desktop\CrimeData\offender_prior.txt
```

If the data are located in a shapefile, then specify both the file name and the field names

```
Anchor Point Prior Distribution File = C:\Users\Analyst\Desktop\CrimeData\offender_prior.dbf
Prior Anchor Point Longitude Field Name = HOMEX
Prior Anchor Point Latitude Field Name = HOMEY
```

After specifying the file type, we also need to specify the bandwidth. To ask the prototype to use the calculated mean nearest neighbor distance, use the directive

```
Anchor Point Density Manual Bandwidth = no
```

To manually choose the bandwidth, use the directives

```
Anchor Point Manual Bandwidth = yes
Anchor Point Bandwidth = 500 ft.
```

Allowable units for the bandwidth are feet (ft), miles (mi), meters (m), and kilometers (km).

Results Directory. The next required directive specifies the directory that will be used to store the results; it has the general form

```
Results Directory =  C:\Users\Analyst\Desktop\CrimeData\Results
```

The directory must already exist; the program will not create it.

Base Point. The final directive is optional. One approach to testing the code is to run it against solved crime series. If a base point is specified in the parameter file, then the code will compare the base point to the search area and say that the base point either was or was not inside the generated search area.

In the case that the base point is inside the search area, it will order the coarse triangles from most likely to least likely, and report which triangle contained the base point; it will perform the same analysis for the collection of fine triangles as well. It is specified with a directive of the form

```
Base Point = -76.76806,39.30946
```

Using The Program. To run the analysis program, simply run it from a command prompt, specifying the name of the parameter file.

Scripting The Program. One of the advantages of the command-line tool is that it is well suited to being used in automated scripts. For example, here is a short Python script that was used to run the tool through the 237 residential burglary series discussed earlier. The data for each individual run was contained in its own directory, named “001” through “237”.

```
# Script RunProfiler.py
import multiprocessing
import os
import subprocess

def f(i):
   output_path = "C:\Users\moleary\Desktop\Testing\Results"
   program = "C:\Users\moleary\Desktop\Testing\Profiler\Profiler.exe"

   parameter = output_path + "/{:03}/data/Parameters.txt".format(i)
   output_file_name = output_path + "/{:03}/data/output.txt".format(i)
   output_file = open(output_file_name,'w')

   p = subprocess.Popen([program,parameter],stdout=output_file)
   print "Case {} started, subprocess PID is {}".format(i,p.pid)
   p.wait()
   output_file.close()
   print "Case {} finished.".format(i)

# Windows parallel processing protection....
if __name__ == '__main__':

   # We want to use all cores but 2, so that the system will remain responsive...
   cpu_count = multiprocessing.cpu_count()
   if(cpu_count > 2):
      cores = cpu_count - 2
   else:
      cores = 1

   print "Using {} cores from the {} total".format(cores,cpu_count)
   # Build the pool
   pool = multiprocessing.Pool(cores)
```

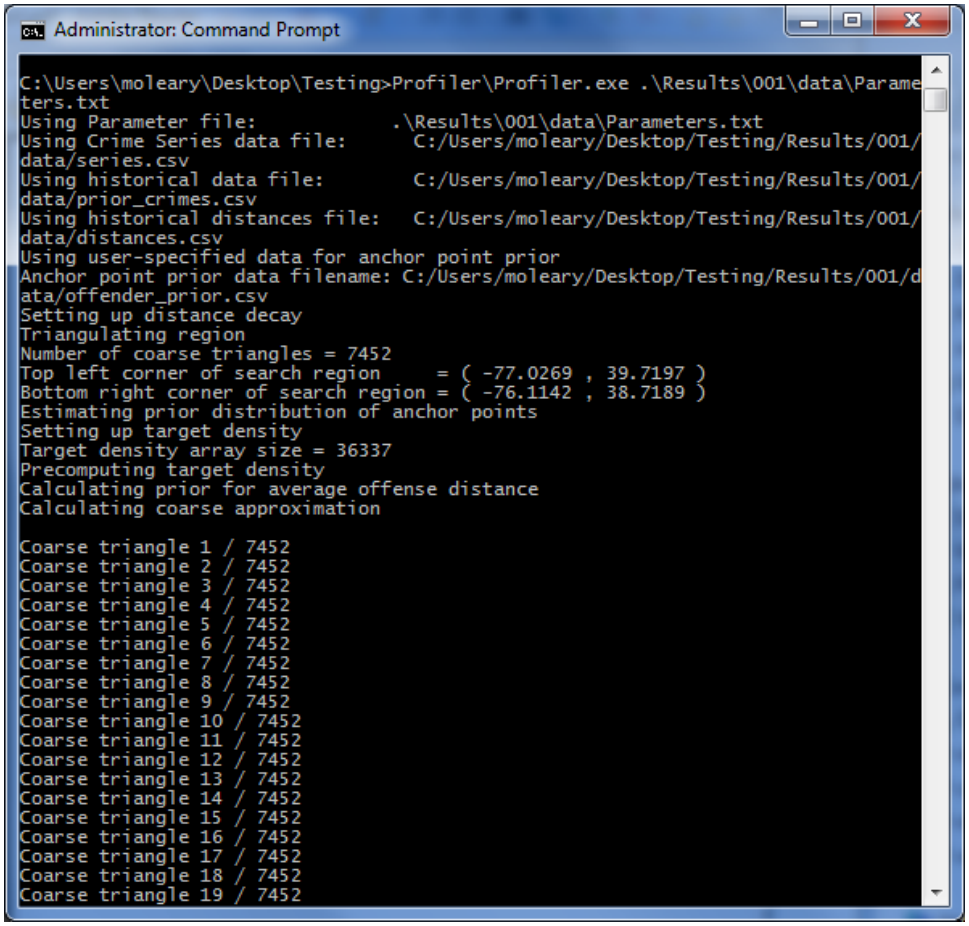


Figure 36. Running the command line analysis tool

```
for i in range(1,238):
   result = pool.apply_async(f,(i,))

pool.close()
pool.join()
```

Not only does this script manage starting and stopping the Profiler program, it uses multi-threading to run multiple instances in parallel, and stores the output from the tool in separate files to aid in subsequent analysis. The only output sent to the screen from the script are the notifications when the script starts a new Profiler session or an existing session completes.

Compiling the Analysis Program. The current version of the prototype was written in C++ and compiled on Microsoft Windows using the MinGW suite and the gcc compiler (version 4.5.2). It should compile and run on a Linux system that contains the required libraries, but this has not been tested.

The program requires the Lapack++ libraries (v. 2.5+). It is important to use a recent version (available from `http://lapackpp.sourceforge.net/`). There is a much older version of Lapack++ available at the NIST web site; it is not suitable. These libraries are licensed under the LGPL.

The program also requires the use of the Shapefile C Library, available from `http://shapelib.maptools.org/`. Use of this library is also available under a LGPL license.

The Graphical User Interface. The graphical user interface is a completely separate program from the analysis program. When run, the graphical user interface prompts the user to provide the necessary data for the analysis engine to run. It generates a parameter file that contains that contains this information, then calls the analysis engine in a separate process. It processes the output from the analysis engine to return the status back to the user through the graphical user interface.

The tool itself is written in C++ and Qt 4.7.0, using Qt Creator 2.0.1. Both the library and the development tool are available from `http://qt.nokia.com/downloads`. The Qt libraries are licensed under the LGPL.

The graphical user interface also requires the Shapefile C Library.

Mathematical Theory

Foundation

Mathematical and criminological theory form the foundation on which the prototype is based. To understand the theory, we begin by establishing some common mathematical notation that will be used throughout. A geographic point $\mathbf{x}$ will have two components $\mathbf{x} = \langle x^{(1)}, x^{(2)} \rangle$; these are simply the distances of the point from a pair of perpendicular reference axes. We assume that the offender has a single well defined anchor point during the crime series denoted by $\mathbf{z}$, and we assume that the series under study has n linked crimes at the locations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$.

The fundamental assumption of the method is that there is a function $P(\mathbf{x})$ that gives the probability density that an offender will commit a crime at the location $\mathbf{x}$. Though we are modeling the behavior of the offender with a probability density, this is not meant to imply that the offender necessarily has a random component in the selection of crime site locations. Rather, this randomness reflects the lack of knowledge that we have about the behavior of the offender. With the probability density known, then the probability that the offender commits a crime in the geographic region R is calculated by adding up the values of P within R, giving us the probability $\iint_R P(\mathbf{x})\, dx^{(1)}\, dx^{(2)}$.

On what variables should the probability density $P(\mathbf{x})$ depend? Clearly it should depend on the offender's anchor point $\mathbf{z}$. We also assume that it depends on the average distance the offender is willing to travel, which will be denoted by α. Not all offenders have the same propensity or ability to travel, and we expect that both the travel patterns and crime patterns of a fifteen year old in an urban neighborhood are likely to be different than a forty year old truck driver living in a rural area. In some of our later analysis we will investigate models where the location of the next crime in a series depends also on the locations of the previous elements in the crime series. For the moment though, we will assume that the probability density P depends only on the anchor point $\mathbf{z}$ and the average offense distance α giving us the function $P(\mathbf{x} \,|\, \mathbf{z}, \alpha)$.

What functional form should the probability density $P(\mathbf{x} \,|\, \mathbf{z}, \alpha)$ take? We assume that it depends on two different factors. First is a distance decay component, and second are geographic characteristics of the target site itself.

The distance decay function will be called D, and it is a function of the distance $d = d(\mathbf{x}, \mathbf{z})$ from the anchor point $\mathbf{z}$ to the crime site $\mathbf{x}$. The distance decay function D will also depend on the average distance the offender travels to offend α, so that

$$P(\mathbf{x} \,|\, \mathbf{z}, \alpha) \propto D(d(\mathbf{x}, \mathbf{z}), \alpha).$$

There are multiple reasonable models for both the distance decay function itself and for the underlying distance metric. The prototype uses the straight line distance as the distance metric[1] and a bivariate normal distribution for the offender's distance decay function. This then has the simple form

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

An extensive discussion of distance decay functions in general and the motivation for this particular choice in the prototype follow in subsequent sections.

Though distance is important, it clearly is not the only factor involved in the selection of a crime site. In the most obvious example, if an analyst is examining a series of liquor store robberies, then those robberies all took place at the location of a liquor store. Any reasonable model of offender behavior needs to account for the fact that there are some locations that cannot possibly be the location of an offense site, as they do not contain a liquor store. The situation can often be more nuanced than just the absence of a particular type of target. Consider street robberies as an example. Unlike liquor store robberies which can happen only in certain well defined locations, street robberies can occur on any street. On the other hand, some streets are far more likely to be the location of a street robbery than others. This can be caused by a dearth of targets, but it can also be due to other local factors, like the lighting and the local police presence.

With this in mind, we also assume that there is a function $G(\mathbf{x})$ that describes the (relative) likelihood that the location $\mathbf{x}$ will be chosen as a target and that

$$P(\mathbf{x}\,|\,\mathbf{z},\alpha) \propto G(\mathbf{x}).$$

Locations with large values of $G(\mathbf{x})$ are considered to be more likely offense targets that regions with low values; locations where $G(\mathbf{x}) = 0$ are considered to be places where the crime cannot occur.

To estimate this function, the prototype uses the locations of past crimes of the same general type as the series under consideration. The distribution is then constructed via kernel density parameter estimation. No attempt is made to explain why some locations are more likely to be an offense site than others; instead these differences are simply acknowledged and measured. The approach to this function is the same as it was in the previous version of the prototype. It uses the quartic kernel $K(\mathbf{y}\,|\,\lambda)$ with bandwidth λ in the form

$$K(\mathbf{y}\,|\,\lambda) = \begin{cases} \frac{3}{\pi\lambda^6}(|\mathbf{y}|^2-\lambda^2)^2 & \text{if } |\mathbf{y}| \le \lambda, \\ 0 & \text{if } |\mathbf{y}| \ge \lambda. \end{cases}$$

Suppose that the historical crimes have taken place at the locations $\{\mathbf{c}_1,\mathbf{c}_2,\ldots,\mathbf{c}_N\}$; then $G(\mathbf{x})$ is given by

$$G(\mathbf{x}) = \sum_{i=1}^{N} K(\mathbf{x}-\mathbf{c}_i\,|\,\lambda)$$

for an appropriate bandwidth λ. As already noted in the discussion of the prototype, the bandwidth λ can be specified by the analyst or can be calculated as the mean nearest neighbor distance between offense locations.

[1] Internally, the prototype stores points as longitude-latitude pairs, and calculates the distance between points in decimal degrees using a spherical approximation for the surface of the earth; these can be converted back to straight line distances in the usual fashion.

Though we have specified that $P(\mathbf{x} \mid \mathbf{z}, \alpha) \propto D(d(\mathbf{x}, \mathbf{z}), \alpha)$ and $P(\mathbf{x} \mid \mathbf{z}, \alpha) \propto G(\mathbf{x})$, we cannot simply then state that P is equal to the product of D and G. Indeed, because P is a probability density function (for each choice of $\mathbf{z}$ and α), we know that we must have

$$\iint P(\mathbf{x} \mid \mathbf{z}, \alpha)\, dx^{(1)}\, dx^{(2)} = 1$$

for each value of $\mathbf{z}$ and α. To ensure this, we must scale the density, while being aware that this scaling depends on both $\mathbf{z}$ and α. Define the normalization function

$$N(\mathbf{z}, \alpha) = \left[\iint P(\mathbf{x} \mid \mathbf{z}, \alpha)\, dx^{(1)}\, dx^{(2)}\right]^{-1};$$

then we can set

$$P(\mathbf{x} \mid \mathbf{z}, \alpha) = D(d(\mathbf{x}, \mathbf{z}), \alpha) G(\mathbf{x}) N(\mathbf{z}, \alpha) = \frac{D(d(\mathbf{x}, \mathbf{z}), \alpha) G(\mathbf{x})}{\iint P(\mathbf{y} \mid \mathbf{z}, \alpha)\, dy^{(1)}\, dy^{(2)}}.$$

This completes the specification of the model for offender behavior.

Suppose for the moment that the offender has committed a single crime at the location $\mathbf{x}$. Bayes' theorem then implies that the joint distribution of the offender's anchor point $\mathbf{z}$ and average offense distance α satisfies

$$P(\mathbf{z}, \alpha \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid \mathbf{z}, \alpha)\pi(\mathbf{z}, \alpha)}{M(\mathbf{x})}.$$

Here $P(\mathbf{x} \mid \mathbf{z}, \alpha)$ is the model of offender behavior, $\pi(\mathbf{z}, \alpha)$ is the prior distribution of offender anchor point and average offense distance before accounting for the location of the crime, and $M(\mathbf{x})$ is the marginal density of the crime site locations. More precisely, the marginal density is given by

$$M(\mathbf{x}) = \iiint P(\mathbf{x} \mid \mathbf{z}, \alpha)\pi(\mathbf{z}, \alpha)\, dz^{(1)}\, dz^{(2)}\, d\alpha.$$

Because M does not depend on either $\mathbf{z}$ or α, it can be considered constant; then to avoid the difficulty in calculating this constant, we simply can simply replace the equality with a proportionality:

$$P(\mathbf{z}, \alpha \mid \mathbf{x}) \propto P(\mathbf{x} \mid \mathbf{z}, \alpha)\pi(\mathbf{z}, \alpha). \tag{1}$$

The prior $\pi(\mathbf{z}, \alpha)$ represents our knowledge of the distribution of both the offender's anchor point $\mathbf{z}$ and the offender's average offense distance α before we account for the location of the crime. To model this, we assume that the two components can be modeled separately and independently, so that

$$\pi(\mathbf{z}, \alpha) = H(\mathbf{z})\pi(\alpha)$$

where $H(\mathbf{z})$ is the prior distribution of offender anchor points, and $\pi(\alpha)$ is the prior distribution for the average offender distance.

The prototype provides two ways to estimate the prior distribution of anchor points $H(\mathbf{z})$. If the anchor point locations $\{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_M\}$ of past offenders are known, then the prior can be estimated using kernel density parameter estimation as we did before, so that

$$H(\mathbf{z}) = \sum_{i=1}^{M} K(\mathbf{z} - \mathbf{h}_i \mid \lambda)$$

for an appropriate bandwidth λ, which in the prototype is either manually specified by the analyst or selected as the mean nearest neighbor distance between offenders.

If instead, the analyst wishes to use U.S. Census data to generate the prior, then the prototype queries the data for the location, population count (modified by any provided demographic data) and land area of each block; it then calculates

$$H(\mathbf{z}) = \sum_{i=1}^{N_{\text{blocks}}} = p_i K(\mathbf{z} - \mathbf{q}_i \,|\, \sqrt{A_i})$$

where each block has population p_i, location $\mathbf{q}_i$ and for each block we have chosen a different bandwidth equal to the side length of a square with the same area A_i as the block.

Estimating the prior distribution $\pi(\alpha)$ of average offense distances is much more complex, but is unchanged from the original prototype; we refer the interested reader to the original NIJ report (O'Leary, 2009b, pp. 28–41).

With the various priors known, (1) then becomes

$$P(\mathbf{z}, \alpha \,|\, \mathbf{x}) \propto P(\mathbf{x} \,|\, \mathbf{z}, \alpha) H(\mathbf{z}) \pi(\alpha).$$

To determine the probability density for just the anchor point, rather than the joint distribution of the anchor point and the average offense distance, we marginalize to find

$$P(\mathbf{z} \,|\, \mathbf{x}) \propto \int P(\mathbf{x} \,|\, \mathbf{z}, \alpha) H(\mathbf{z}) \pi(\alpha) \, d\alpha. \tag{2}$$

The case of multiple crimes requires a more complex model of offender behavior; in particular we need to specify the probability density $P(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \,|\, \mathbf{z}, \alpha)$ given that the offender commits crimes in all of the locations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. The simplest solution is to assume that the crime site locations are all selected independently; in this case

$$P(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \,|\, \mathbf{z}, \alpha) = \prod_{i=1}^{n} P(\mathbf{x}_i \,|\, \mathbf{z}, \alpha).$$

The argument that led to (2) can then be repeated, yielding

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

and so

$$P(\mathbf{z} \,|\, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) \propto \int \left[\prod_{i=1}^{n} P(\mathbf{x}_i \,|\, \mathbf{z}, \alpha) \right] H(\mathbf{z}) \pi(\alpha) \, d\alpha. \tag{3}$$

This completes the foundational mathematics of the model, and represents the starting point for our subsequent analysis; it is also the approach taken in the originally released prototype. The biggest weakness in the approach outlined so far is the underlying assumption that the locations of the crime sites are independent. Significant time was spent during the project examining that assumption, and we have evidence that this independence assumption is flawed, both in terms of its theoretical justification as well as how its effectiveness in the results produced by the prototype. In subsequent sections, we will describe our theoretical analysis of the independence problem, and show some ways in which it could be improved. In another section, we will show one one the consequences of the independence assumption on the behavior of the original prototype, and will explain how the density (3) has been modified in the current version of the prototype to obtain the improved results already described.

Distance Decay

Distance decay is a fundamental notion in criminology. In a paper written and published during the grant period (O'Leary, 2011), I examined the relationship between the two dimensional offense distribution that describes how offenders select targets and the corresponding one-dimensional distance decay function. The following material is largely excerpted from that already published paper.

Dimensionality and Distance Decay. Criminal distance decay is the notion that there is a relationship between the distance from an offender's anchor point to a potential target location, and the likelihood that the offender chooses to offend in that location. To make this idea mathematically precise, define the distance decay function to be the probability density that the offender chooses a target at a specified distance from their anchor point. Call this function $D(r)$; then to find the probability P that the distance to the offense is between the numbers a and b, we simply sum the density and calculate $P = \int_a^b D(r)\, dr$. Note that because $D(r)$ is a probability density and because the distance r must be nonnegative, we know that $D(r) \geq 0$ for all r and $\int_0^\infty D(r)\, dr = 1$.

Throughout what follows, we will continue to use the phrase "distance decay" because it historically has been the term used to describe this phenomenon. However, we will not assume that the distance decay distribution is monotone decreasing, but instead explicitly allow for the possibility that increasing the distance from the anchor point can increase or decrease the probability density that an offense takes place at that distance.

Because offenders select targets rather than distances, it is clear that the truly fundamental quantity that describes offender target selection is the two-dimensional probability distribution that describes how offenders select targets.

For simplicity in what follows, we will always assume that the offender's anchor point is located at the origin in this coordinate system. We define the offenders offense distribution $T(\mathbf{x}) = T(x^{(1)}, x^{(2)})$ to be the probability density that the location $\mathbf{x}$ is selected by the offender as the location of an offense. Then, for any geographic region Ω, we can find the probability P that an offense occurs within Ω by calculating $P = \iint_\Omega T(x^{(1)}, x^{(2)}) dx^{(1)} dx^{(2)}$. Because T is a probability density, we know that $T(x^{(1)}, x^{(2)}) \geq 0$ and $P = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} T(x^{(1)}, x^{(2)})\, dx^{(1)}\, dx^{(2)} = 1$. We remark that Rossmo (2000, p. 197) calls a three-dimensional map of the offense distribution a jeopardy surface.

As we have already noted, the behavior of the offender, and thus the offense distribution T can depend on factors unique to the individual committing the offenses, the crime type and crime characteristics, and on the local geographic, demographic and other characteristics of the region.

The geometry of space imposes a fundamental relationship between the two-dimensional offense distribution $T(x^{(1)}, x^{(2)})$ and the corresponding one-dimensional distance decay distribution $D(r)$. To see this, suppose that we are using Euclidean distance, and we want to determine the probability P that an offense occurs at a distance between r and $r + \Delta r$. To calculate this, we sum the values of T on the annulus Ω with inner radius r and outer radius $r + \Delta r$, giving

$$P = \iint_{r \leq |x| \leq r + \Delta r} T(x^{(1)}, x^{(2)})\, dx^{(1)}\, dx^{(2)}.$$

If we make the simplifying assumption that the offense distribution T depends only on the distance to the anchor point, then T is roughly constant on that annulus; say $T(\mathbf{x}) \approx T(r)$ for $r \leq |x| \leq r + \Delta r$. The area A of the annulus can be calculated simply by taking the difference of the area

enclosed by the outer circle from the area enclosed by the inner circle, so $A = \pi(r+\Delta r)^2 - \pi r^2 = 2\pi r\Delta r + \pi(\Delta r)^2$; thus the probability that the offense takes place in our annulus is the product of the probability density and the area of the region, and so satisfies

$$P \approx (2\pi r\Delta r + \pi(\Delta r)^2)T(r).$$

On the other hand, our definition of the distance decay distribution $D(r)$ tells us that the probability that the offense lies at a distance between r and $r+\Delta r$ is $P = \int_r^{r+\Delta r} D(s)\, ds$. If we make the same simplifying assumption and assume that D is roughly constant on the interval $[r, r+\Delta r]$ with the value $D(r)$, we also have the approximation

$$P \approx [(r+\Delta r) - r]D(r) = (\Delta r)D(r)$$

formed by taking the product of the length of the interval with the value of the function on that interval. Combining these two different expressions of the same quantity P, and canceling the factor Δr from both sides, we obtain the relationship

$$D(r) \approx (2\pi r + \pi(\Delta r))T(r)$$

and for small Δr, we see that

$$D(r) \approx (2\pi r)T(r).$$

Although this derivation proceeded via a number of mathematical approximations, these approximations are not germane and it is simple to replace this argument with a formal proof. The fundamental relationship between the offense distribution and the distance decay function

$$D(r) = 2\pi r T(r)$$

holds whenever the underlying two-dimensional offense distribution T depends only on distance where Euclidean distance is being used.

To illustrate this fundamental result, suppose we place 300 points uniformly randomly throughout the square shown in Figure 37, effectively choosing the constant value of $T = 1/4$ on the square $[-1,1] \times [-1,1]$. Four subregions are shown that figure, a disk of radius $1/20$ and annuli of width $1/20$ with inner radii at $1/4$, $1/2$, and $3/4$. A simple count shows us that there are no points within the inner disk while the annuli contain 6, 11, and 21 points- quite near the expected results, which would be 0.6, 6.5, 12.3 and 18.3. Though the hypothesized underlying two-dimensional offense distribution is constant, our fundamental relationship tells us that the associated distance decay function should be linear, increasing as r gets large and tending to zero as r tends to zero- which is exactly what we observe.

The notion that one must pay attention to the difference between the one-dimensional distance decay and the two-dimensional offense distribution has already been noted by Rengert, Piquero, and Jones (1999) in their critique of the work of van Koppen and de Keijser (1997).

To understand the consequences of the fundamental relationship, let us consider what happens when we apply it to some common distributions. Probably the most natural two-dimensional offense distribution is the bivariate normal centered at the offender's anchor point, which we have placed at the origin of our coordinate system so

$$T(\mathbf{x}) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{|\mathbf{x}|^2}{2\sigma^2}\right).$$

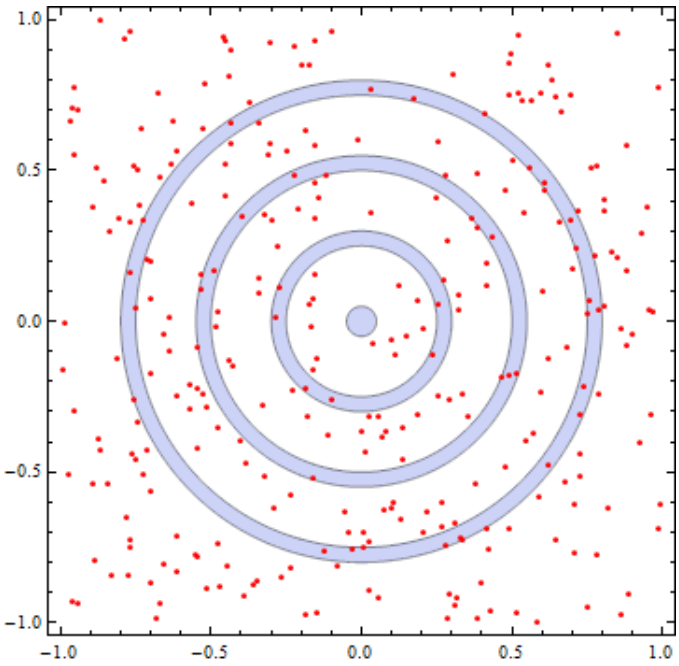


Figure 37. 300 points placed uniformly randomly throughout a square.

In this case, the corresponding distance decay function $D(r)$ is

$$D(r) = \frac{r}{\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right)$$

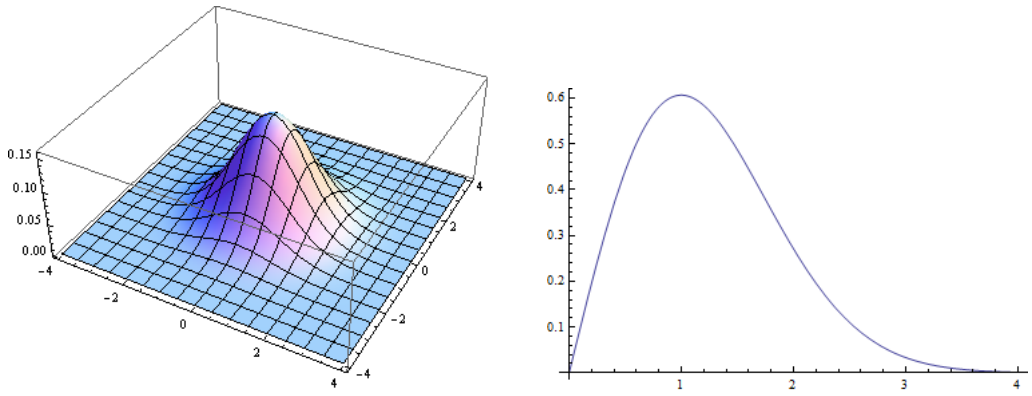
which can be recognized as a Rayleigh distribution. If we graph both of these with $\sigma = 1$, we obtain Figure 38.

Despite the shape of the distance decay function in Figure 38b, it is clear from the offender distribution in Figure 38a that this offender's behavior does not exhibit a buffer zone. Indeed, the offense distribution shows that the offender is more likely to offend at locations closer to their anchor point rather than less likely. The distance decay curve $D(r)$ vanishes as $r \to 0$ not because of the existence of a buffer zone, but rather because of the underlying two-dimensional nature of space. As $r \to 0$, the area available to offend decreases, and thus the distance decay function decreases, precisely in line with our fundamental result.

If the distance decay function is known, it is possible to use the fundamental result to construct a two-dimensional offense distribution that matches that distance decay. This two-dimensional offense distribution is not unique, and given any distance decay function, there are an infinite number of two-dimensional offense distributions with the same distance decay function. However, there is only one two-dimensional distribution that both matches a given distance decay curve and depends solely on the distance from the anchor point to the potential offense site; this satisfies

$$T(\mathbf{x}) = \frac{1}{2\pi|\mathbf{x}|} D(|\mathbf{x}|)$$

if we continue to assume the use of a Euclidean distance.



(a) Offense distribution (b) Distance decay function

Figure 38. Two views of the same process- the offense distribution is bivariate normal and the distance decay function is Rayleigh

The CrimeStat Manual (Levine, 2010, Chp. 10) provides five built-in choices for a distance decay curve; they are

- Linear

$$D(r) = \begin{cases} A + Br & \text{if } A + Br \geq 0 \\ 0 & \text{if } A + Br < 0 \end{cases}$$

- Negative exponential

$$D(r) = Ae^{-Br}$$

- Normal

$$D(r) = \frac{A}{S\sqrt{2\pi}} \exp\left(-\frac{(r-\bar{r})^2}{2S^2}\right)$$

- Lognormal

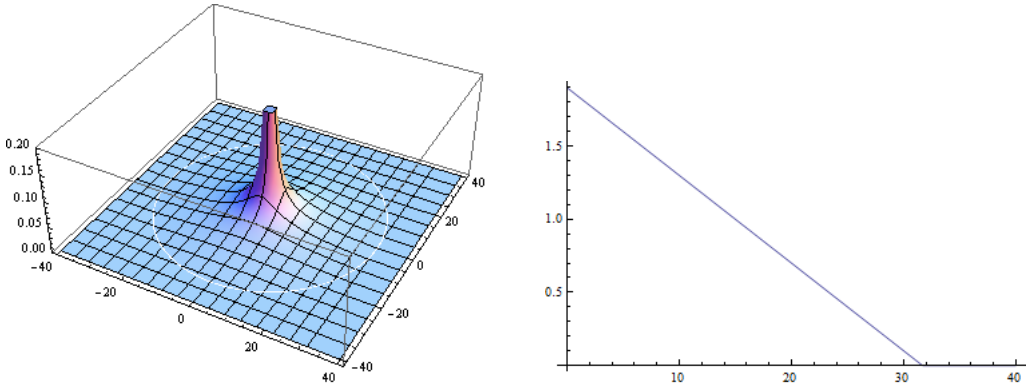
$$D(r) = \frac{A}{r^2 S\sqrt{2\pi}} \exp\left(-\frac{[\ln(r^2)-\bar{r}]^2}{2S^2}\right)$$

- Truncated negative exponential

$$D(r) = \begin{cases} Br & \text{if } r \leq r_p \\ Ae^{-Cr} & \text{if } r > r_p \end{cases}$$

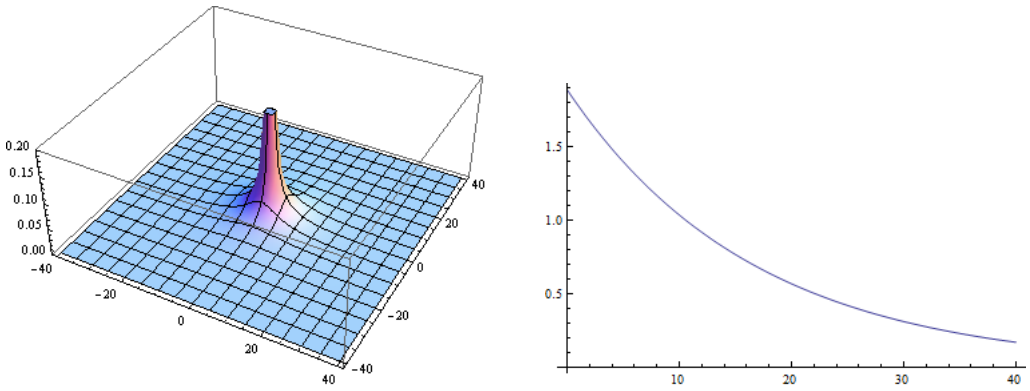
We can then graph each of these distance decay functions together with the corresponding two-dimensional offense distribution; these are shown in Figures 39–43.

To create each of these figures, we used the default parameters provided in the CrimeStat Manual (Levine, 2010, Chp. 10); thus the horizontal axes in each of should be considered to be distances measured in miles. Because we are using the parameters from that manual, the distance decay curves do not satisfy the normalization requirement $\int_0^\infty D(r)\, dr = 1$. This does not impact CrimeStat because it uses these functions to generate hit scores where comparisons are made between regions where these scores are large and areas where these scores are small. Each of these distance decay curves can be made into a properly scaled probability distribution by simply scaling each function by an appropriate multiplicative constant.



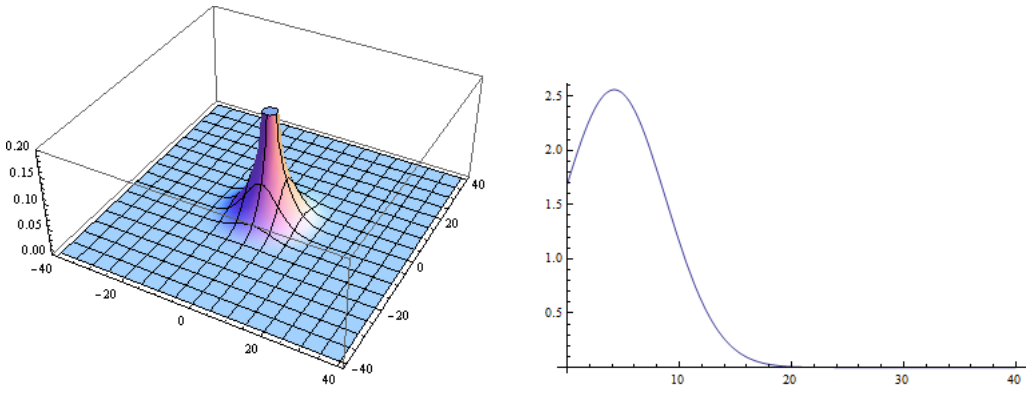
(a) Offense distribution (b) Distance decay function

Figure 39. Linear distance decay of Levine; $A = 1.9$, $B = -0.06$.



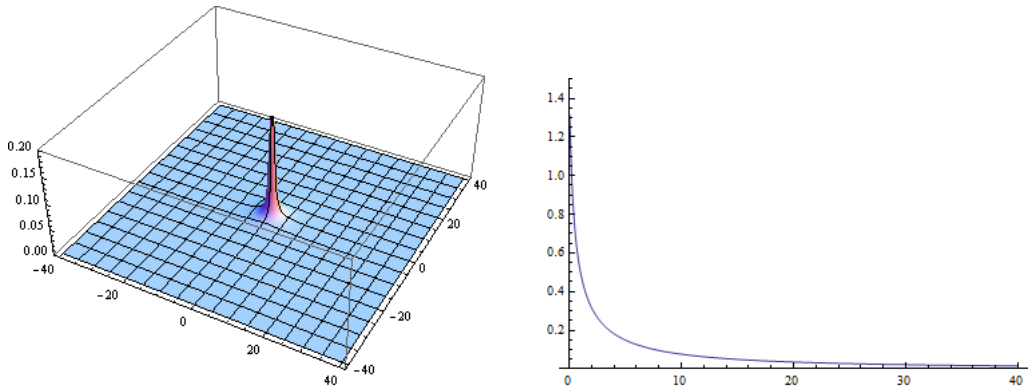
(a) Offense distribution (b) Distance decay function

Figure 40. Negative exponential distance decay of Levine; $A = 1.89$, $B = -0.06$.



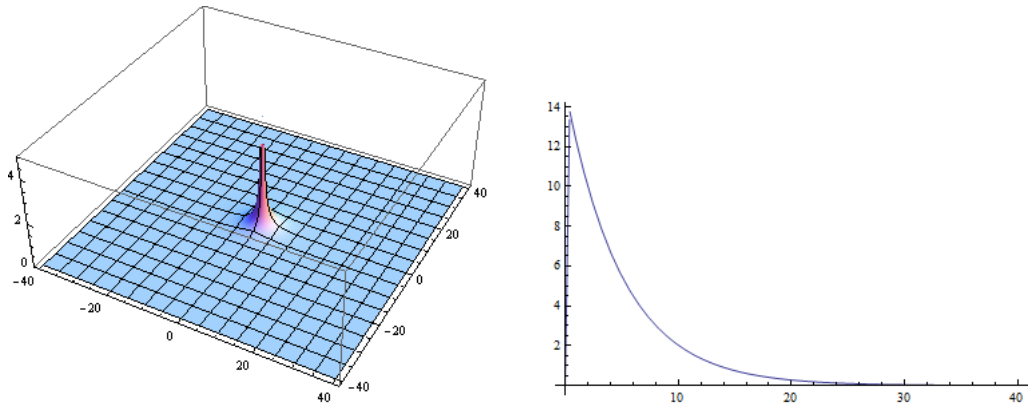
(a) Offense distribution (b) Distance decay function

Figure 41. Normal distance decay of Levine; $A = 29.5$, $S = 4.6$, $\bar{r} = 4.2$



(a) Offense distribution (b) Distance decay function

Figure 42. Lognormal distance decay of Levine; $A = 8.8$, $S = 4.6$, $\bar{r} = 4.2$



(a) Offense distribution (b) Distance decay function

Figure 43. Truncated exponential distance decay of Levine; $A = 14.95$, $B = 34.5$, $C = 0.2$, $r_p = 0.4$

Examining the graphs of the offense distribution for each of these, we see that none of these distributions exhibit a buffer zone. In fact, in all of these graphs, the two-dimensional offense distribution is strongly concentrated at the origin, and in four of the five graphs, the offense distribution becomes infinite as we approach the origin. The only exception is the truncated exponential in Figure 43, though that may not be clear from Figure 43a. A closer look near the origin of that distribution provided in Figure 44 shows that the two-dimensional offense distribution actually exhibits a flat plateau in the region $r < r_p$.

Rossmo (2000) takes a different approach; he uses a Manhattan distance metric together with a piecewise rational function for the distance decay function. However, there is an analogue of our fundamental result for the Manhattan distance.

Suppose that we want to find the probability that an offense occurs where the Manhattan distance is between m and $m + \Delta m$. To calculate this, we need to sum the values of T over the annular region Ω outside the square $|x^{(1)}| + |x^{(2)}| \leq m$ and inside the square $|x^{(1)}| + |x^{(2)}| \leq m + \Delta m$, giving us

$$P = \iint_{m<|x^{(1)}|+|x^{(2)}|\leq m+\Delta m} T(x^{(1)}, x^{(2)})\, dx^{(1)}\, dx^{(2)}$$

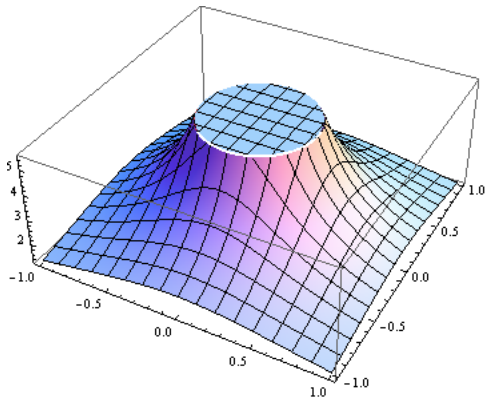


Figure 44. The center of the offense distribution generated for the truncated negative exponential distance decay of Levine

We again assume that T is roughly constant on this region, with a value that depends only on the Manhattan distance m so $T(x) \approx T(m)$. To find the area of this region, we note that the region $|x^{(1)}| + |x^{(2)}| \leq m$ is a square of side length $m\sqrt{2}$; thus to find the area A of the region $m < |x^{(1)}| + |x^{(2)}| \leq m + \Delta m$, we take the difference of the areas of the concentric squares and find $A = ((m + \Delta m)\sqrt{2})^2 - (m\sqrt{2})^2 = 4m\Delta m + 2(\Delta m)^2$. Thus

$$P \approx (4m\Delta m + 2(\Delta m)^2)T(m).$$

On the other hand, our definition of the distance decay distribution tells us that the probability that the offense lies at a Manhattan distance between m and $m + \Delta m$ is

$$P = \int_m^{m+\Delta m} D(s)\, ds \approx [(m + \Delta m) - m]D(m) = (\Delta m)D(m)$$

where again we are assuming that D is roughly constant on the interval $[m, m + \Delta m]$ with value $D(m)$. Combining these two expressions for P and considering Δm small, we see that for Manhattan distance, the fundamental relationship is

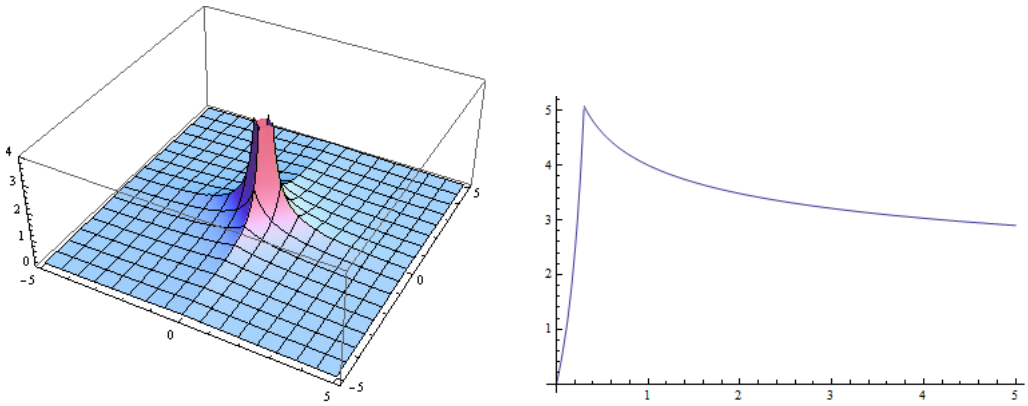
$$D(m) = 4mT(m).$$

Again, though our argument proceeded by approximation, this can be proven rigorously.

Rossmo uses the offense distribution rather than the distance decay; his form is

$$T(x^{(1)}, x^{(2)}) = \begin{cases} \dfrac{k}{(|x^{(1)}| + |x^{(2)}|)^f} & \text{if } |x^{(1)}| + |x^{(2)}| \geq b, \\ \dfrac{kb^{g-f}}{(2b - |x^{(1)}| - |x^{(2)}|)^g} & \text{if } |x^{(1)}| + |x^{(2)}| \leq b. \end{cases}$$

For criminal profiling, Rossmo recommends the choice $f = g = 1.2$, (Rossmo, 1995, p. 341) (see also Le Comber, Nicholls, Rossmo, and Racey (2006, §3.2) and Raine, Rossmo, and Le Comber (2009) for applications outside criminal profiling). The buffer zone parameter b is set depending on the characteristics of the crime series, while k is simply a scaling parameter (Rossmo, 1995, p. 341).



(a) Offense distribution (b) Distance decay function

Figure 45. Rossmo model, $f = g = 1.2$, $b = 0.3$, $k = 1$

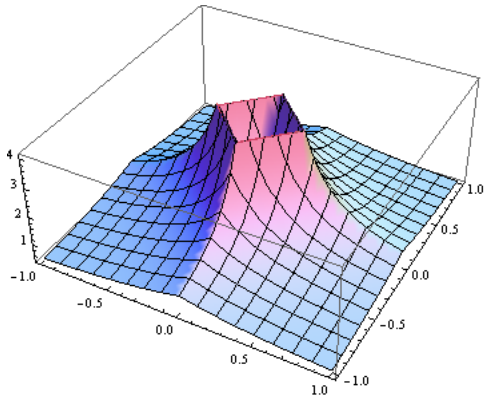


Figure 46. The center of the offense distribution of Rossmo

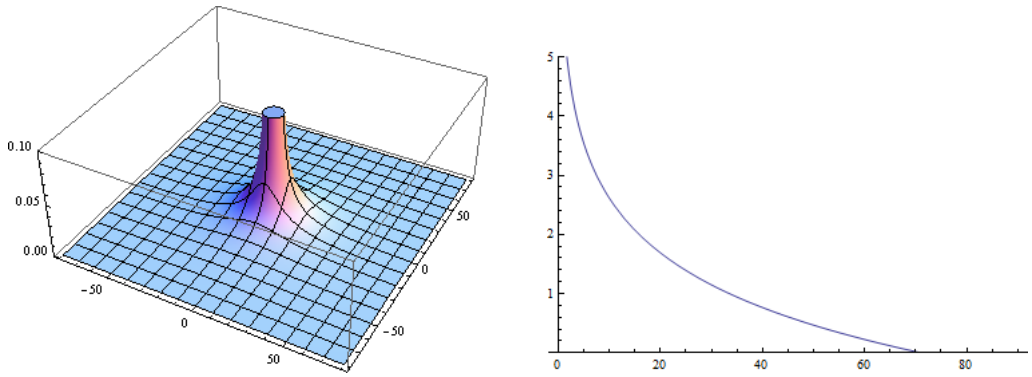
We graph the offense distribution and the distance decay function in Figure 45. We see immediately the effect caused by the use of the Manhattan distance rather than the Euclidean distance in that the offense distribution is no longer radially symmetric; this is more clearly seen in Figure 46 which shows the center of the offense distribution. We also notice that the distance decay function decays rapidly as $m \to 0$, while it decays very slowly for large values of m.

We remark that, at least for the values $f = g = 1.2$ the model of Rossmo cannot represent an offense distribution, as the tails are too large to allow for the required normalization $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} T(x^{(1)}, x^{(2)})\, dx^{(1)}\, dx^{(2)} = 1$ regardless of the choice of the parameters b and k as we always have $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} T(x^{(1)}, x^{(2)})\, dx^{(1)}\, dx^{(2)} = \infty$. One solution to this issue would be to truncate the values of the offense distribution outside some finite but large region. One concern with this approach is that this necessarily imply that the values of the other parameters, especially the parameter k, would vary significantly with the size of the region used in for the truncation.

Canter and Hammond (2006) examine four different forms for distance decay:

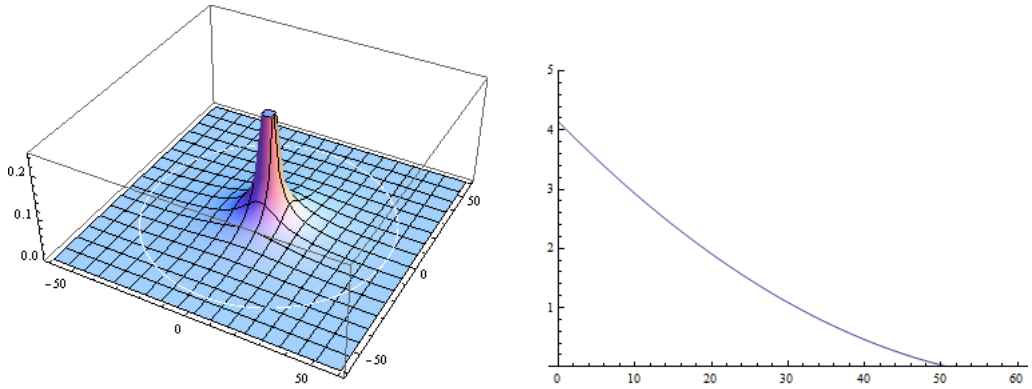
- Logarithmic

$$D(r) = \begin{cases} A + B\ln r & \text{if } A + B\ln r \geq 0 \\ 0 & \text{if } A + B\ln r < 0 \end{cases}$$



(a) Offense distribution (b) Distance decay function

Figure 47. Logarithmic distance decay of Canter and Hammond; $A = 5.6735$, $B = -1.3307$



(a) Offense distribution (b) Distance decay function

Figure 48. Quadratic distance decay of Canter and Hammond; $A = 0.000981$, $B = -0.13129$, $C = 4.14137$

- Negative exponential

$$D(r) = Ae^{-Br}$$

- Quadratic

$$D(r) = \begin{cases} Ar^2 + Br + C & \text{if } Ar^2 + Br + C \geq 0 \\ 0 & \text{if } Ar^2 + Br + C < 0 \end{cases}$$

- Linear

$$D(r) = \begin{cases} A + Br & \text{if } A + Br \geq 0 \\ 0 & \text{if } A + Br < 0 \end{cases}$$

Two of these distributions have already been analyzed, so we provide the corresponding graphs for just the logarithmic and the quadratic distributions using the parameters selected by Canter and Hammond (2006). In their work, distances were measured in kilometers, so this represents the horizontal axes in these graphs.

Estimating the distance decay function. Knowledge of the structure and form of the offense distribution of an individual offender would clearly be valuable, not only from a theoretical perspective but also for its potential value in developing improved geographic profiling algorithms.

Given the difficulty and complexity of the full problem, most attention has been paid to the problem of simply determining the structure and form of the offender's distance decay function.

In the CrimeStat manual, Levine (2010, pp. 10.24 ff.) recommends using aggregated crime data for a jurisdiction to select the parameters that appear in the provided analytically defined distance decay functions; this process is illustrated with data from Baltimore County. In particular, he recommends choosing the parameters in the decay curve so as to best fit the aggregate data. He also shows how CrimeStat can be used to generate a experimentally determined distance decay function that is an even better fit for the aggregate data.

In comparison, Rossmo (1995, p. 341) states that "The value of the distance exponent ($f = 1.2$) was selected from a gravity model formulation developed to describe interprovincial migration of criminal fugitives (Rossmo, 1987, p. 136)." The value of the buffer zone parameter b is selected heuristically and is dependent on the mean nearest neighbor distance, while the multiplicative scaling constant k is primarily used as a computational aide.

In other applications of this model, the values of these parameters are selected to optimize the effectiveness of the geoprofile. Indeed, Le Comber et al. (2006) apply this model to construct geographic profiles in animal foraging. In their analysis, they hold the constants k and B fixed, with the latter dependent on the trip distance of the species under study. Then a collection of values for the exponents f and g were selected; for each exponent pair, the effectiveness of the geoprofile was evaluated to determine the best choice for the exponents. A similar approach was taken by Raine et al. (2009).

This technique of choosing distance decay parameters so as to optimize the effectiveness of the geoprofile was taken earlier by Canter et al. (2000). They examined a collection of negative exponential models for distance decay, and then considered variants that contained buffer zones and/or plateaus; in all 285 different functions were evaluated for their effectiveness as geographic profiling tools.

Canter and Hammond (2006) take a hybrid approach; they selected four general models for offender distance decay- logarithmic, negative exponential, quadratic, and linear. For each model, the parameters that describe the model were obtained by fitting the decay curve to a collection of aggregate data. Each function generated a slightly different geographic profiling algorithm; these algorithms were compared for effectiveness.

Although all of these approaches to the geographic profiling problem select a distance decay function, their focus naturally is on determining the effectiveness of geographic profiling algorithms, rather than on our problem of estimating the offense distribution or the distance decay function of an individual offender.

The primary issue in any approach to the determination of the distance decay behavior of an individual offender is one of data. Though researchers have access to data sets from a number of jurisdictions covering a wide range of crime types, the data available about any one offender is necessarily limited. In general the series size for an individual offender is small, and though there are series where the number of elements is large, it is possible that offenders who successfully complete a large series are special in some way. As a consequence, the individual series contain insufficient amounts of data to reliably estimate much more than some elementary parameters of the series, for example the mean and the standard deviation of the distance from the offender's anchor point to the crime sites.

If all offenders behaved in the same fashion, then is would be a simple matter to aggregate data collected across a suitably large set of offenders to obtain a reliable estimate of their distance

decay patterns. However, the literature has well established significant variation in offense patterns depending on characteristics of the offender, the criminal event, and the local geography.

It was noted by van Koppen and de Keijser (1997) that it may not be appropriate to try to draw inference about the distance decay patterns of an individual offender based on the distance decay behavior observed when aggregated across offenders. Though Rengert et al. (1999) disagreed with many of the conclusions in that work, they unequivocally stated that "We do not dissent from Van Koppen and De Keijser's assertion that researchers cannot and should not make inferences about individual behavior with data collected at the aggregate level."

Recently, Smith, Bond, and Townsley (2009) analyzed residential burglary in Northhamptonshire during 2002-04. They examined 32 offenders who committed series of at least 10 crimes; together they committed 590 burglaries. They compared the aggregate distance decay curve for all offenders with the distance decay curves found by aggregating only offenders who were alike in age and found significant qualitative differences between these graphs, providing direct evidence that offenders do not in general possess the same distance decay patterns. In fact, they found that roughly two-thirds of the variation in offense distances can be ascribed to variations between offenders, rather than to variation in an individual offender.

Townsley and Sidebottom (2010) then examined a larger data set of residential and non-residential burglaries involving more than 1,300 offenders and 16,000 offenses. They found that nearly half of the variation in offense distances can be explained by variation between offenders.

Given these facts, is is possible to estimate the quantitative behavior of an individual from the available aggregate data?

Coefficient of Variation. To proceed, we are going to analyze a data set provided by the Baltimore County Police Department. Our data set for analysis consists of solved residential burglaries committed in Baltimore County between 1986 and 2008. For each incident, we have identified the location of the offense and the location of the home of the offender; we also have basic demographic information about the offender, including the age, sex, race, and date of birth of the offender. The data set contains 5859 offender / offense pairs, with 2890 individual offenders committing 4542 separate burglaries. The data set contains 322 series of four or more burglaries committed by the same offender.

Individual offenders were identified by matching the provided demographic information and home location. We acknowledge that it is possible that two or more different individuals may have the same demographic information and the same home address; it is also possible that a single individual offender could possess different home addresses at different times. However, the data set contains no instances where the date of birth matched and the home address did not, suggesting that these concerns are likely to be of little practical impact. We also note that the home address recorded in the data set may not be the actual anchor point for the offender at the time of the offenses. A series is identified solely as a collection of burglaries committed by the same offender; as a consequence we do not account for confounding factors like the presence of multiple offenders.

Of the 322 identified series, the mean length of a series is 8.1 crimes, with a median of 6 and a maximum of 54. A histogram of the number of series as a function of the series length is provided in Figure 49.

Suppose that the distance decay behavior of an individual offender is governed by a negative exponential distribution in the usual form

$$D(r|\beta) = \frac{1}{\beta} e^{-r/\beta},$$

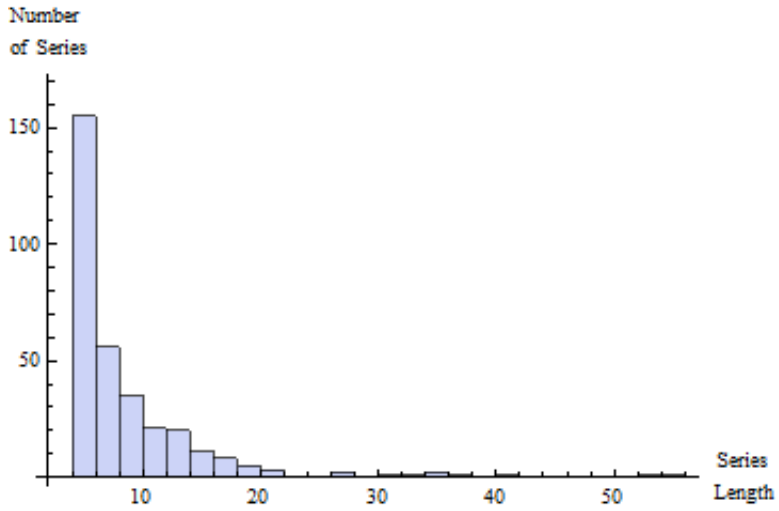


Figure 49. Baltimore County Residential Burglaries, 1986-2008. Number of series as a function of the length of the series

but that the parameter β may be different for each offender. Then we do not expect that the distribution of offense distances aggregated across all offenders satisfies a negative exponential distribution. To illustrate this, suppose that offenders fall into two categories, with the fraction p_1 with parameter β_1 and the fraction p_2 with parameter β_2. Then it is clear to see that, sampling across offenders, we obtain the aggregate distribution

$$A(r) = p_1 D(r|\beta_1) + p_2 D(r|\beta_2)$$

which is not a negative exponential. More generally, if the parameter β is distributed across offenders according to a probability density $\mu(\beta)$ then the distribution of offense distances $A(r)$ aggregated across offenders would then satisfy

$$A(r) = \int_0^\infty D(r|\beta)\mu(\beta)\, d\beta = \int_0^\infty \frac{1}{\beta} e^{-r/\beta}\mu(\beta)\, d\beta. \tag{4}$$

Although each individual offense series is far too small to reliably estimate the offense distribution or even the distance decay function of an individual offender, we can estimate both the mean and the standard deviation of the distance from the home location to the offense site for each series. Doing so lets us estimate the coefficient of variation of each series, which is defined to be the ratio of the standard deviation to the mean of a distribution.

We can also calculate the mean and standard deviation of the negative exponential distribution; both the mean and the standard deviation are β, and hence the coefficient of variation is identically 1. The significance of this result is that coefficient of variation does not depend on the parameter β. Thus, if all individual offenders select targets according to a negative exponential, then we expect that the sample coefficient of variation should be roughly 1, regardless of how the parameter β is distributed across offenders in the population.

To test this hypotheses, we plotted the mean and the standard deviation of the distance from the crime site to the offender's home in Figure 50. If individual offenders all have a negative

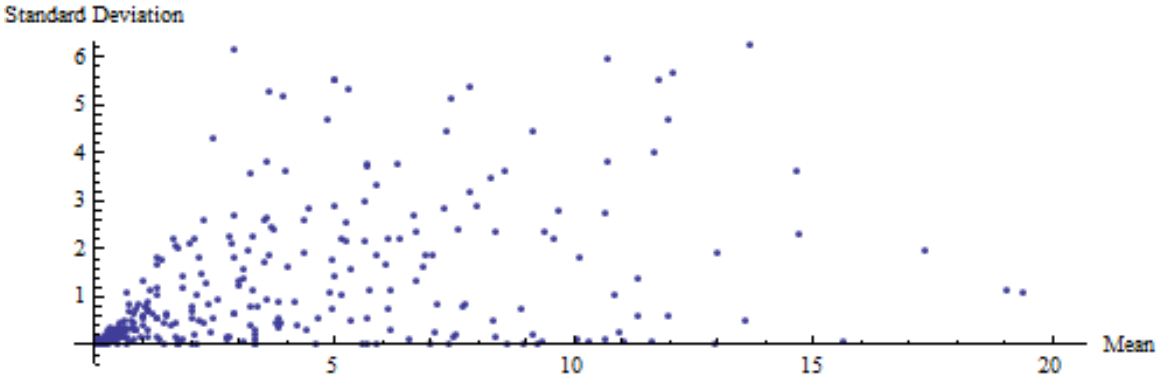


Figure 50. Mean and standard deviation in miles of the distance from crime site to anchor point for 322 residential burglary series in Baltimore County, 1986–2008

Name	Distribution	Mean	Standard Deviation	Coefficient of Variation
Negative Exponential	$\frac{1}{\beta}e^{-r/\beta}$	β	β	1
Logarithmic	$A + B\ln r$	$\frac{1}{4}e^{-A/B}$	$\frac{\sqrt{7}}{12}e^{-A/B}$	$\frac{\sqrt{7}}{3} \approx 0.882$
Normal ($\bar{r} = 0$)	$\frac{2}{\sigma\sqrt{2\pi}}\exp\left(-\frac{r^2}{2\sigma^2}\right)$	$\sqrt{\frac{2}{\pi}}\sigma$	$\sqrt{\frac{\pi-2}{\pi}}\sigma$	$\sqrt{\frac{1}{2}(\pi-2)} \approx 0.756$
Linear	$A + Br$	$\frac{2}{3A}$	$\frac{\sqrt{2}}{3A}$	$\frac{1}{\sqrt{2}} \approx 0.707$
Rayleigh	$\frac{r}{\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right)$	$\sqrt{\frac{\pi}{2}}\sigma$	$\sqrt{\frac{4-\pi}{2}}\sigma$	$\sqrt{\frac{4}{\pi}-1} \approx 0.522$

Table 1
Coefficient of variation of commonly selected distance decay curves

exponential distance decay function, then we would expect these points to lie close to the line with slope 1. Examining the graph, it is clear that the typical ratio of standard deviation to mean is much less than unity. To investigate this further, we plot a histogram of the coefficients of variation for the distance decay function of our serial offenders and obtain Figure 51 which confirms our observation that the coefficient of variation is less than unity for nearly all of our observations and suggests that the negative exponential model for distance decay is not a good fit for the behavior of individual offenders.

We can also calculate the coefficient of variation for the remaining commonly chosen models for offender distance decay. In many cases, the coefficient of variation is independent of all of the parameters in the distribution, letting us repeat this analysis. Doing so, we find that the coefficient of variation is constant for the logarithmic model (0.882), the normal model with $\bar{r} = 0$, (0.756), the linear model (0.707), and the Rayleigh model (0.522). These results are summarized in Table 1. Comparing these analytically computed values to our data, we find that none of these models appear to provide a compelling explanation of the observed offender behavior. Indeed, the calculated

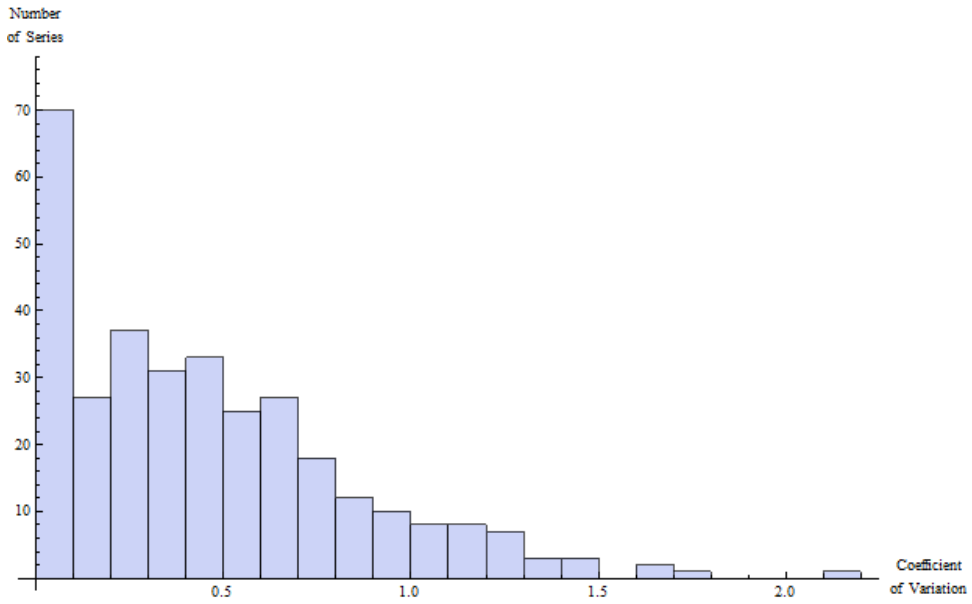


Figure 51. Histogram of the coefficient of variation for distance from offense site to offender's home for 322 residential burglary series in Baltimore County, 1986–2008

coefficients of variation are all at least 0.522, while the mean of the coefficients of variation for the 322 series is 0.455 and the median is 0.390.

When we calculate the coefficient of variation for other distributions, we find that the result depends on one or more of the parameters in the distribution. For example, for the lognormal distribution

$$D(r|\sigma,\mu) = \frac{1}{r\sigma\sqrt{2\pi}} \exp\left[\frac{-(\ln r - \mu)^2}{2\sigma^2}\right]$$

we find that the coefficient of variation is $\sqrt{e^{\sigma^2} - 1}$. Since we do not know the distribution of σ in the offender population, we can draw no conclusions about the appropriateness of the lognormal model. On one hand, we can construct a distribution of σ that will allow us to match the observed data. On the other hand, we can repeat this process with the other models, like the normal (with $\bar{r} \neq 0$), the truncated exponential, or the quadratic model. Though the expression of the coefficient of variation for those models is algebraically unpleasant, we can still find distributions of the parameters for their models that also allow us to match their coefficient of variation to the observed data. Thus, we do not possess evidence either for or against any of these models.

Some statistical work has been done to determine the distribution of the coefficient of variation from a known underlying distribution. In the case where the underlying distribution is normal, McKay (1932) approximated the distribution of the coefficient of variation by showing that an appropriate transform can be well approximated by a chi-squared distribution. Much later Vangel (1996) gave an improved approximation. Iglewicz and Myers (1970) compare different numerical approximations and the exact values of the percentage points for the sample coefficient of variation when the underlying distribution is normal using a variety of approximations including that of McKay (1932) and Hald (1952); see also Hendricks and Robey (1936) who approached the problem differently. Koopmans, Owen, and Rosenblatt (1964) developed confidence intervals for the coef-

(a) Pattern #1 (b) Pattern #2

Figure 52. If the crime sites are blue circles and the offender's anchor point a red square, then which pattern should be called a commuter and which should be called a marauder?

ficient of variation when the underlying distribution is normal or lognormal, while Linhart (1965) developed approximate limits for the coefficient of variation when the underlying distribution is gamma.

In particular, though we have qualitative evidence to suggest that none of the models in table 1 is sufficient to describe the distance decay behavior of offenders, we are unable to present a quantitative estimate of the potential significance of this statement. We also note that we are not the first researchers to apply the coefficient of variation to problems of distance decay; see Smith et al. (2009) who calculated the coefficients of variation for the different aggregated samples provided by Snook (2004).

Improving Distance Decay Models

Commuters and Marauders. Looking carefully at Figure 50, it is clear that there are a number of series where the mean distance is quite large while the corresponding standard deviation is quite small; indeed for many of the points in Figure 50 the standard deviation is essentially zero. This clearly suggests a commuter pattern to the offender's series as the offender appears to travel essentially the same distance to offend for each element of the series. The presence of these commuters in the data set may explain in part why the observed mean coefficient of variation is smaller than any of the proposed models.

One weakness of the original classification scheme of Canter and Larkin (1993) of offenders as commuters or marauders is that it is a binary classification- either an offender is a commuter or the offender is a marauder, depending on whether or not the offender's anchor point lies within the circle whose diameter is formed by the two crimes that are farthest apart. However, what should be said of an offender whose anchor point lies near or even on the boundary of the circle? A very small change in the location of the offender's anchor point, even just across the street, could result in a change in the offender's classification. In reality such an offender is exhibiting a mixture of commuter and marauder behavior- see Figure 52 for an example.

We have created a different way to differentiate commuters and marauders that explicitly

allows for this type of mixture behavior. Suppose that the offender's crimes are located at the sites $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ and that their anchor point is located at $\mathbf{z}$. Then for any number $1 \leq p < \infty$ define the number

$$\mu_p = \left[\frac{\min_{\mathbf{y}} \sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{y})^p}{\sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{z})^p} \right]^{1/p}$$

where the distance between points is measured by the distance metric d. Note that the number μ_p is independent of the precise units used to measure distances; because it is a ratio of distances it gives the same values whether d measures distances in miles, kilometers, meters, or any other unit. Note also that $0 \leq \mu_p \leq 1$ regardless of the series and the choice of p.

To see the relationship between the numbers μ_p and the distinction between commuters and marauders, consider for a moment the special case where $p = 1$. Then it is clear that

$$\min_{\mathbf{y}} \sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{y}) = d(\mathbf{x}_i, \mathbf{y}_{\text{cmd}})$$

where $\mathbf{y}_{\text{cmd}}$ is the center of minimum distance of the crime series; indeed this is the definition of the location of the center of minimum distance. Then

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

From this we see that if the anchor point $\mathbf{z}$ is close to the center of minimum distance $\mathbf{y}_{\text{cmd}}$, then the numerator and denominator are roughly the same and so the ratio is roughly one; at the same time if the anchor point is close to the center of minimum distance then the offender is exhibiting a classic marauder pattern. On the other hand, suppose that the offender is a commuter. Then the crimes cluster far away from the anchor point. In this case the distance from the anchor point to each crime is going to be much larger than the distance from the center of minimum distance of the crime series to each crime; in this case the ratio is roughly zero.

To see the same behavior for $p = 2$, suppose that the distance metric is Euclidean distance. Then it is known that the point $\mathbf{y}$ that minimizes the function

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

is precisely the centroid

$$\mathbf{y}_{\text{centroid}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i.$$

Thus, we have the definition

$$\mu_2 = \sqrt{\frac{\sum_{i=1}^{n} |\mathbf{x}_i - \mathbf{y}_{\text{centroid}}|^2}{\sum_{i=1}^{n} |\mathbf{x}_i - \mathbf{z}_{\text{anchor}}|^2}}.$$

Again, if the offender is exhibiting a marauder pattern, then the anchor point and the centroid of the series will be similar and the ratio roughly 1; if the offender is exhibiting a commuter pattern then the crimes will be close to the centroid of the crime series but far away from the anchor point and the ratio is roughly zero.

Returning to Figure 52, pattern #1 is a Canter & Larkin commuter, while pattern #2 is that of a Canter & Larkin marauder. The values of μ_2, however, are roughly the same; pattern #1 has $\mu_2 = 0.56$ while pattern #2 has $\mu_2 = 0.58$.

Relationship of μ_p to Canter & Larkin Marauders & Commuters. There is direct but subtle relationship between the values of μ_p and the Canter & Larkin notion of marauder. For any crime series, define

$$\mu_\infty = \lim_{p\to\infty} \mu_p = \lim_{p\to\infty} \left[\frac{\min_{\mathbf{y}} \sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{y})^p}{\sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{z})^p} \right]^{1/p}.$$

If all of the crimes are contained in the circle whose diameter is formed by the segment connecting the two crimes that are farthest apart and Euclidean distance is used, then every Canter & Larkin marauder necessarily satisfies $\mu_\infty \geq 1/2$.

To prove this assertion, suppose without loss of generality that two of the crimes farthest apart are $\mathbf{x}_1$ and $\mathbf{x}_2$, so that

$$|\mathbf{x}_i - \mathbf{x}_j| \leq |\mathbf{x}_1 - \mathbf{x}_2|$$

for any other pair i, j. Let $2r = |\mathbf{x}_1 - \mathbf{x}_2|$ and $\mathbf{c} = \frac{1}{2}(\mathbf{x}_1 + \mathbf{x}_2)$ so that $\mathbf{c}$ is the midpoint of the segment from $\mathbf{x}_1$ to $\mathbf{x}_2$ and r is the distance from that midpoint to either $\mathbf{x}_1$ or $\mathbf{x}_2$.

For any choice of $\mathbf{y}$, we know that

$$|\mathbf{x}_1 - \mathbf{y}| + |\mathbf{x}_2 - \mathbf{y}| \geq |\mathbf{x}_1 - \mathbf{x}_2| = 2r$$

so that either $|\mathbf{x}_1 - \mathbf{y}| \geq r$ or $|\mathbf{x}_2 - \mathbf{y}| \geq r$. Thus

$$\left(\sum_{i=1}^{n} |\mathbf{x}_i - \mathbf{y}|^p \right)^{1/p} = r \left(\sum_{i=1}^{n} \frac{|\mathbf{x}_i - \mathbf{y}|^p}{r^p} \right)^{1/p} \geq r.$$

We can then conclude that

$$\min_{\mathbf{y}} \left(\sum_{i=1}^{n} |\mathbf{x}_i - \mathbf{y}|^p \right)^{1/p} \geq r.$$

On the other hand, we also know that

$$\min_{\mathbf{y}} \left(\sum_{i=1}^{n} |\mathbf{x}_i - \mathbf{y}|^p \right)^{1/p} \leq \left(\sum_{i=1}^{n} |\mathbf{x}_i - \mathbf{c}|^p \right)^{1/p}.$$

Using the assumed fact that all of the crimes $\mathbf{x}_i$ are contained in the circle, we know that $|\mathbf{x}_i - \mathbf{c}| \leq r$ for all i, and hence

$$\min_{\mathbf{y}} \left(\sum_{i=1}^{n} |\mathbf{x}_i - \mathbf{y}|^p \right)^{1/p} \leq \left(\sum_{i=1}^{n} r^p \right)^{1/p} \leq n^{1/p} r.$$

As a consequence

$$r \leq \min_{\mathbf{y}} \left(\sum_{i=1}^{n} |\mathbf{x}_i - \mathbf{y}|^p \right)^{1/p} \leq n^{1/p} r$$

and so

$$\lim_{p\to\infty} \min_{\mathbf{y}} \left(\sum_{i=1}^{n} |\mathbf{x}_i - \mathbf{y}|^p \right)^{1/p} = r.$$

A similar argument shows that for any sequence of nonnegative numbers $\{a_i\}_{i=1}^n$ we have

$$\lim_{p\to\infty} \left(\sum_{i=1}^{n} a_i^p \right)^{1/p} = \max_{1\leq i\leq n} a_i.$$

Indeed, let $M = \max_{1\leq i\leq n} a_i$. Then because the ratio a_i/M is equal to one at least once

$$\left(\sum_{i=1}^{n} a_i^p \right)^{1/p} = M \left(\sum_{i=1}^{n} \frac{a_i^p}{M^p} \right)^{1/p} \geq M$$

while because all of the ratios a_i/M are no more than one

$$\left(\sum_{i=1}^{n} a_i^p \right)^{1/p} = M \left(\sum_{i=1}^{n} \frac{a_i^p}{M^p} \right)^{1/p} \leq M \left(\sum_{i=1}^{n} 1 \right)^{1/p} = M n^{1/p}.$$

Thus

$$M \leq \lim_{p\to\infty} \left(\sum_{i=1}^{n} a_i^p \right)^{1/p} \leq M \lim_{p\to\infty} n^{1/p}$$

and the claim follows.

Returning to the discussion of μ_∞, we then see that if $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n$ are all in the circle, then

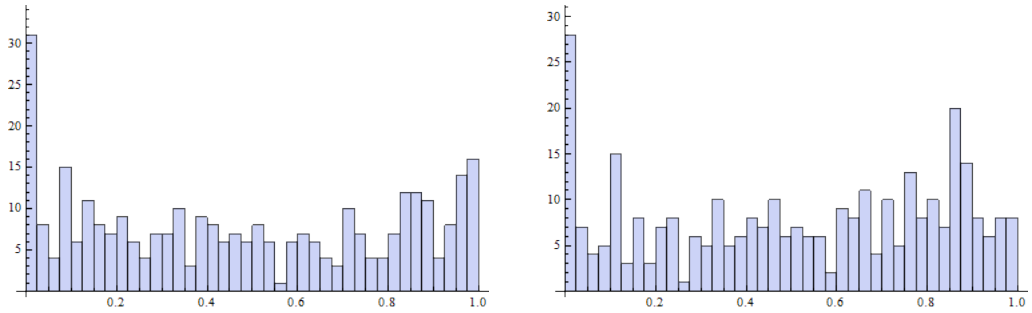
$$\mu_\infty = \lim_{p\to\infty} \mu_p = \frac{\lim\limits_{p\to\infty} \min\limits_{\mathbf{y}} \left(\sum\limits_{i=1}^{n} |\mathbf{x}_i - \mathbf{y}|^p \right)^{1/p}}{\lim\limits_{p\to\infty} \left(\sum\limits_{i=1}^{n} |\mathbf{x}_i - \mathbf{z}|^p \right)^{1/p}} = \frac{r}{\max\limits_{1\leq i\leq n} |\mathbf{x}_i - \mathbf{z}|}.$$

So, if the anchor point $\mathbf{z}$ lies in the circle, then

$$|\mathbf{x}_i - \mathbf{z}| \leq |\mathbf{x}_i - \mathbf{c}| + |\mathbf{c} - \mathbf{z}| \leq 2r$$

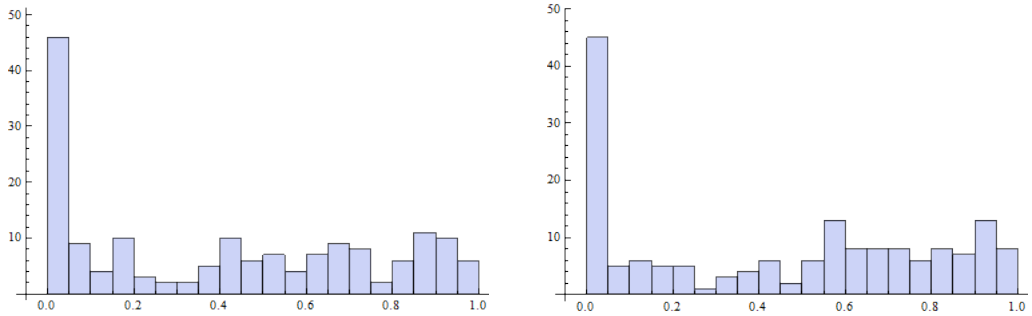
and so

$$\mu_\infty \geq \frac{r}{2r} \geq \frac{1}{2}.$$



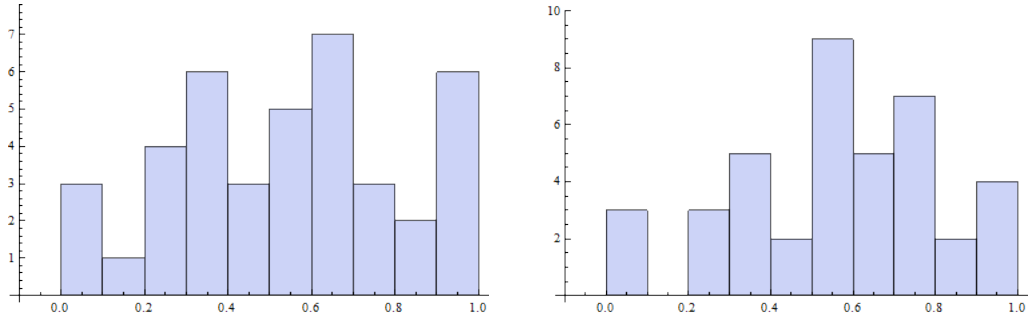
(a) μ_1 (b) μ_2

Figure 53. Histogram of the distribution of μ_1 and μ_2 over 322 solved residential burglary series with at least four crimes in Baltimore County, 1986-2008.



(a) μ_1 (b) μ_2

Figure 54. Histogram of the distribution of μ_1 and μ_2 over 167 solved non-residential burglary series with at least four crimes in Baltimore County, 1989-2008.



(a) μ_1 (b) μ_2

Figure 55. Histogram of the distribution of μ_1 and μ_2 over 70 solved bank robbery series with at least four crimes in Baltimore County, 1993-2009.

Applications of μ_p to Baltimore County Data. To see the implication of these definitions on real data, we have analyzed our data from Baltimore County. We have 5,863 solved residential burglaries from 1986-2008, with 322 identified crime series of at least four crimes. We also have 2,643 solved non-residential burglaries from 1989-2008, and identified 167 series with at least three crimes. Finally, we have 602 solved bank robberies from 1993-2009 and identified 70 series with at least three crimes.

Examining the histogram of the frequency of μ_1 and μ_2 across the residential burglary series (Figure 53), we see little difference between the histograms; we also see that there does not appear to be a sharp distinction between commuters and marauders, but rather that some offenders behave like one or the other, while many show a mixture of these behaviors. Non residential burglary (Figure 54) shows a decided preference for commuters, while bank robbery (Figure 55) shows a slight preference towards marauders, especially in the histogram of μ_2.

Scaled Distances & the Rayleigh Distribution. Many analysts when faced with the task of estimating the distance decay behavior of an offender start by looking at the distance decay curve aggregated across all offenders, despite the danger of the ecological fallacy. For example, if we were to examine the aggregate distance decay function across 5,863 studied residential burglaries in Baltimore County, we obtain Figure 56 which appears consistent with the usual negative exponential

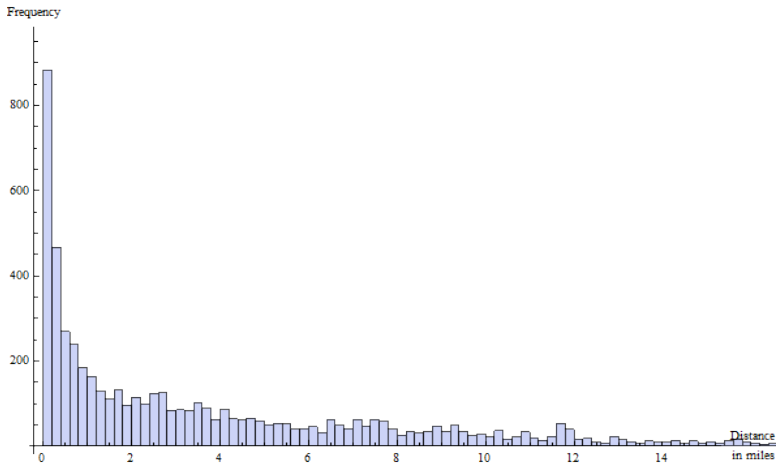


Figure 56. Baltimore County Residential Burglaries, 1986-2008. Distance from offender's home to crime site, aggregated across all offenders.

models for offender behavior. However, if we assume both that offenders behave like a negative exponential but that different offenders have different parameters, then we would not expect to see a negative exponential in the aggregate; this was already noted in (4).

To move forward while avoiding the ecological fallacy, we need to know that all of the individuals exhibit the same behavior prior to aggregation; in this fashion we will know that the aggregate also reflects that same behavior. If the only quantity that varies between offenders is the average distance that the offender is willing to travel, then the scaled distance would exhibit the same behavior. Thus if an offender with anchor point $\mathbf{z}$ and average offense distance α commits crimes at the locations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, then we can consider not the distances $d(\mathbf{x}_i, \mathbf{z})$, but rather the scaled distances $d(\mathbf{x}_i, \mathbf{z})/\alpha$. Of course, for a real offender the average offense distance α is not

known, but it can be estimated; let

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

We then consider the scaled distances

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

where $\hat{\alpha}$ is different for different serial offenders. If we then plot the histogram of the scaled distances rather than the distances, we obtain the very different histogram, Figure 57. Unlike the ag-

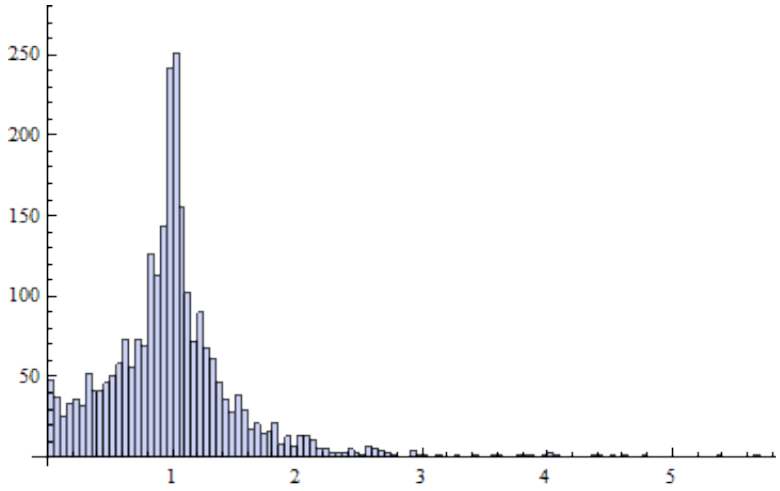


Figure 57. Histogram of scaled distances ρ for 322 residential burglary series in Baltimore County with at least four elements from 1986-2008

gregate distribution, which appeared well modeled by a negative exponential, the scaled distances have a very different qualitative form, with a peak near $\rho = 1$ and decay both as ρ gets large as well as when $\rho \to 0$.

If we want to draw inferences about individual behavior from the histogram of the raw (unscaled) aggregated data (Figure 56) then to avoid the ecological fallacy we would need to know that different offenders behaved in roughly the same fashion. On the other hand, because the scaled distances show such a different structure, we are led to the conclusion that there is significant variation between offenders, and moreover that conclusions about the behavior of individuals should not be drawn from aggregate distance decay graphs like Figure 56.

If our hypotheses that the only variation between offenders comes from differing offender average offense distances is correct, then the histogram of scaled distances (Figure 57) should accurately represent each individual's scaled behavior. Unfortunately, the histogram is not immediately recognizable as a well-known distribution. It does share some of the qualitative characteristics of a Rayleigh distribution, and we have already seen from our discussion of the coefficient of variation that the Rayleigh was the best of a number of poor fits for the observed behavior.

If each individual offender is a Rayleigh with average offense distance α, then the scaled distance should match a Rayleigh distribution with average offense distance precisely one. We plot both the histogram and that Rayleigh together in Figure 58, but see quite clearly that they do not match. The problem is that the observed data has many more offenders with a scaled distance of

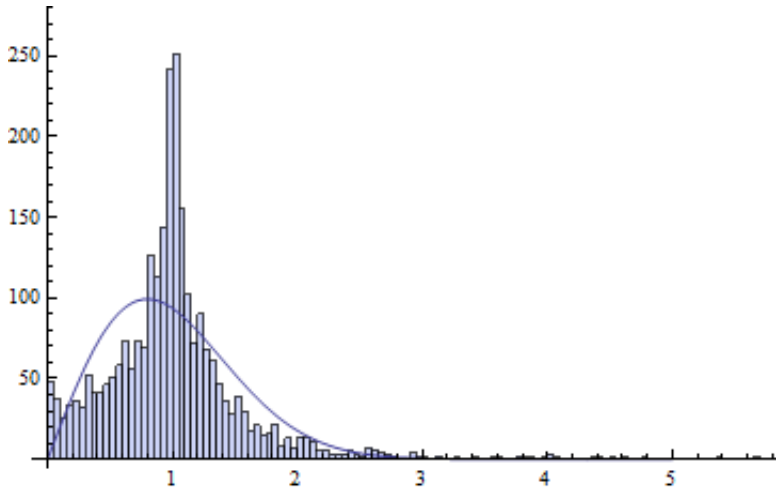


Figure 58. Histogram of scaled distances ρ for 322 residential burglary series in Baltimore County with at least four elements from 1986-2008 compared with a Rayleigh distribution with mean one

roughly one than would be predicted by a Rayleigh distribution.

There is a potential explanation for this behavior though. An offender exhibiting a commuter model of behavior would travel roughly the same distance from the home to the crime site across all of their crimes. Perhaps the oversize peak is due to the influence of commuters in the data? With this in mind, consider Figure 59 that plots the scaled distance ρ against the commuter marauder parameter μ_2. Here we see a clear difference in behavior between commuters and marauders; com-

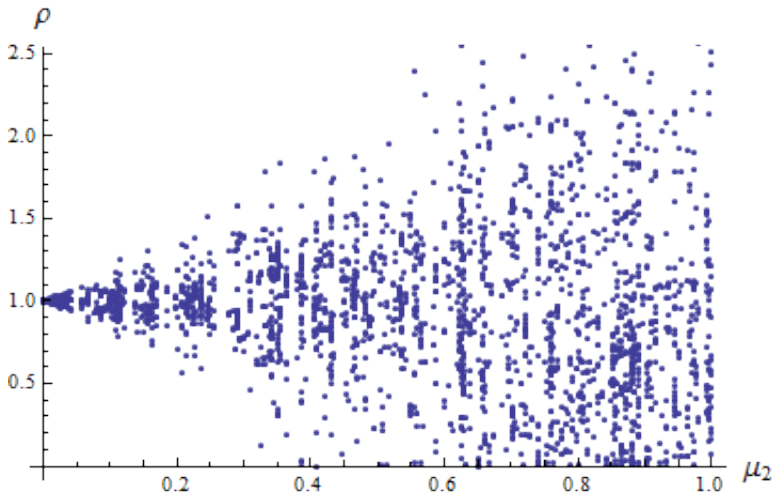


Figure 59. Dependence of scaled distance ρ versus commuter/marauder parameter μ_2 for 322 residential burglary series in Baltimore County with at least four elements

muters, for which $\mu_2 \approx 0$ all have scaled distances ρ of roughly one, while the marauders for which $\mu_2 \approx 1$ show a richer variety of potential behaviors.

Let us focus our attention solely on those offenders who are not strong commuters. In par-

ticular, Figure 60 shows the histogram of scaled distances for those residential burglary series with $\mu_2 \geq 0.25$ and compare the result with a Rayleigh distribution with average one. The degree of

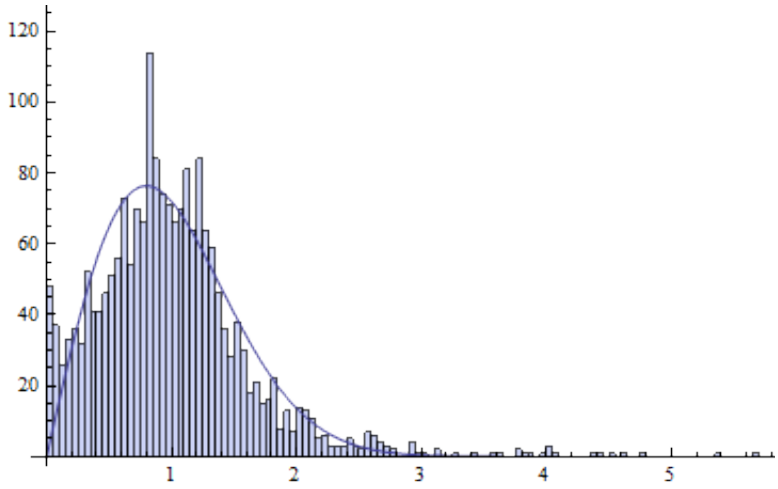


Figure 60. Histogram of scaled distances ρ for residential burglary series in Baltimore County with at least four elements and with $\mu_2 \geq 0.25$ from 1986-2008 compared with a Rayleigh distribution with mean one.

agreement between the theoretical prediction and the observed data is striking, especially when you consider that the theoretical prediction is not a fit to the data, but rather a Rayleigh distribution with mean precisely one.

This match of theory and observation does not appear to be particular to residential burglary. Figure 61 shows agreement between the scaled distances for non-residential burglaries in Baltimore County with $\mu \geq 0.25$ and the Rayleigh distribution with average one, while Figure 62 shows the agreement between scaled distances for bank robberies in Baltimore County with $\mu \geq 0.25$ and the Rayleigh distribution with average one.

It may be possible that these observations are due to something particular to the unusual geography of Baltimore County, especially the way it sits around three sides of Baltimore City. However, we are not the first researchers to consider scaled distances. Warren et al. (1998) examined 108 serial rapists who committed 565 offenses. In Figure 2 of that paper, they plotted the ratio of the distance from residence to the rape locations with the offender's average offense distance for those offenders with five or more crimes. Replotted here as Figure 63, we again see agreement between observation and the theoretical model. In this case, we also note that we did not need to exclude commuters from the data set as we did for residential burglaries, non-residential burglaries, and bank robberies in Baltimore County. On the other hand, circle theory has been shown to be much more effective on rape series than burglary series; indeed in the original paper of Canter and Larkin (1993), they found that 87% of the serial rapists they studied followed a marauder pattern.

Although we have demonstrated the agreement between data and a theoretical prediction for non-commuters, we have no theoretical justification for the exclusion of commuters from the analysis, nor do we have a theoretical justification for the particular limiting value $\mu_2 = 0.25$. It is possible that there is a real distinction that needs to be drawn between offenders that behave strongly as commuters to other offenders. This may be caused either by different patterns of behavior, but

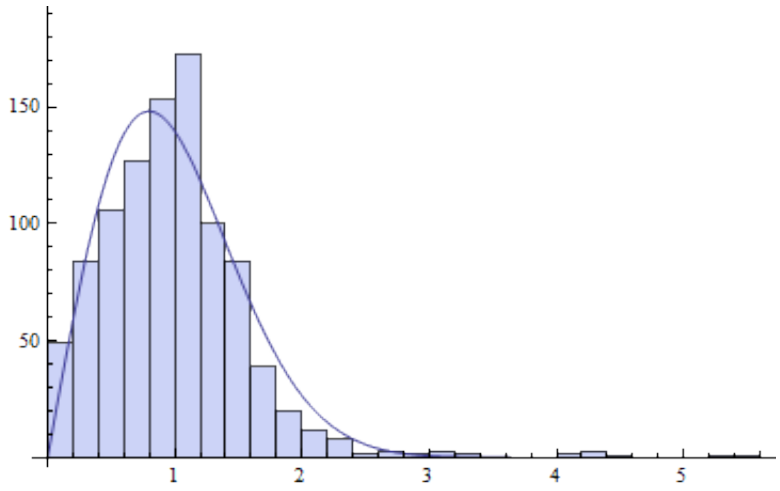


Figure 61. Histogram of scaled distances ρ for non-residential burglary series in Baltimore County with at least four elements and with $\mu_2 \geq 0.25$ from 1989-2008 compared with a Rayleigh distribution with mean one.

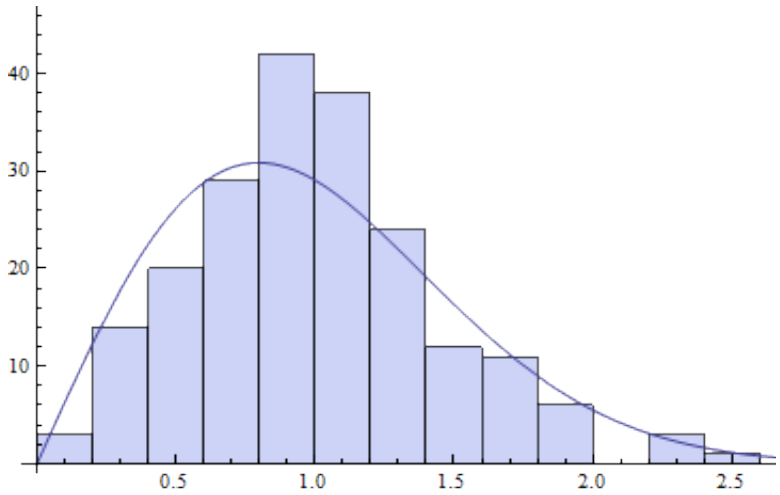


Figure 62. Histogram of scaled distances ρ for bank robbery series in Baltimore County with at least four elements and with $\mu_2 \geq 0.25$ from 1993-2009 compared with a Rayleigh distribution with mean one.

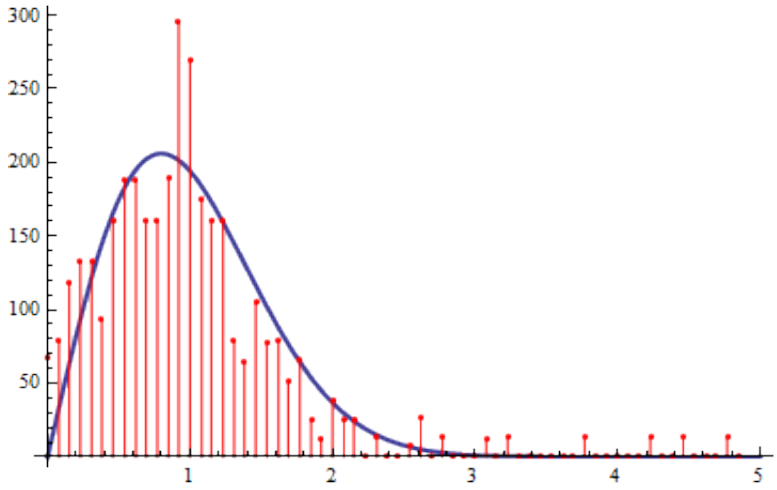


Figure 63. Histogram of scaled distances for serial rape taken from Warren et. al. (1998) compared with a Rayleigh distribution with mean one.

also might be caused by limitations in the data. For example, if an offender behaves as a marauder behaves, but if the offender's home base is mis-identified, then the data would (mistakenly) classify that offender as a commuter. A more interesting possibility is that the underlying model that assumes offenders select offense locations via a bivariate normal distribution is flawed. If is is the case that the error in using the bivariate normal model is larger for commuters than marauders, then we would expect that excluding commuters from the analysis would improve the fit as we have seen.

Coefficient of Variation. Given the agreement between the theoretical Rayleigh distribution with mean one and the observed data when commuters are excluded, it is interesting to return to the coefficient of variation, and repeat that analysis but now with the commuters, defined as those offenders with $\mu_2 < 0.25$, excluded. Plotting both the mean and standard deviation in miles of the distance from the crime site to the anchor point for residential burglaries we now obtain Figure 64; this should be compared with Figure 50.

We can also compare the histogram of the coefficient of variation for the non-commuters. Figure 65 shows the coefficient of variation for just the offenders with $\mu_2 \geq 0.25$; this should be compared to Figure 51 which showed all offenders. We clearly see that a number of offenders whose coefficient of variation is essentially zero are no longer present in these data sets.

We know that the theoretical value for the coefficient of variation for a Rayleigh distribution is roughly 0.522; the mean coefficient of variation for the series with $\mu_2 \geq 0.25$ is 0.552, with a median value of 0.483. We again note that the observed value is close to that for our Rayleigh distribution, but far from the other choices; the coefficient of variation for a negative exponential distribution is one.

Together, these provide compelling evidence to suggest that the distance decay behavior of individual offenders is well-modeled by a Rayleigh distribution, at least for offenders that do not exhibit a marked commuter pattern.

Bivariate Models. The motivation behind the selection of the Rayleigh distribution as a model of offender distance decay was that it was a consequence of the assumption that the offender's two-dimensional distance decay behavior followed a bivariate normal distribution. Though we have

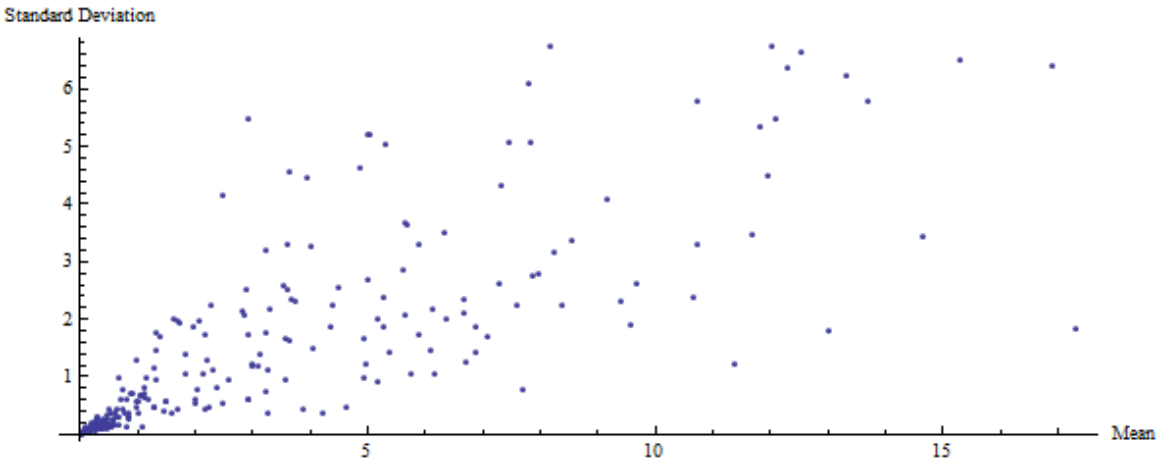


Figure 64. Mean and standard deviation in miles of the distance from crime site to anchor point for residential burglary series in Baltimore County, 1986–2008 with $\mu_2 \geq 0.25$.

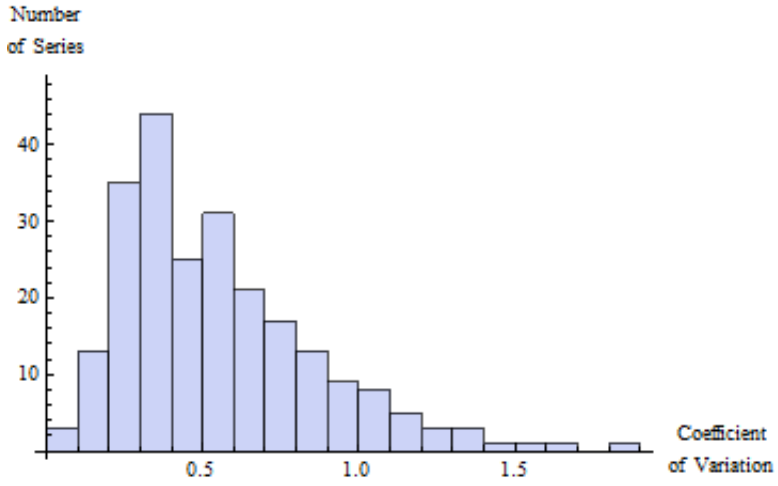


Figure 65. Histogram of the coefficient of variation for distance from offense site to offender's home for residential burglary series in Baltimore County, 1986–2008 with $\mu_2 \geq 0.25$.

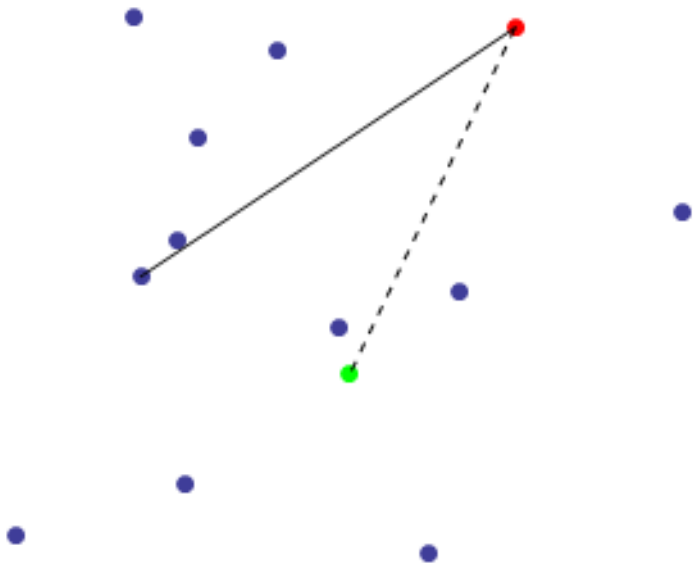


Figure 66. Measuring angles. Blue dots represent crime locations, the red square the offender's anchor point, and the green triangle the centroid of the crime series. For each crime site, measure the angle between the ray from the anchor point to the crime with the referent ray from the anchor point to the centroid of the series.

seen strong evidence that suggests that the Rayleigh distribution is a reasonable model for individual offender one-dimensional distance decay, that does not necessarily imply that the bivariate normal distribution is a good choice to model the two-dimensional distribution of offense sites, as there are an infinite number of bivariate distributions whose distribution of distances is Rayleigh.

To analyze the underlying two-dimensional distribution of offense sites, let us start by considering the angular dependence, if any, in the results. If the distribution of crime sites follows a bivariate normal distribution, then there should be no angular dependence in the results. To measure angles, a referent is needed. In this analysis, we will use the ray from the offender's anchor point to the centroid of the crime series as the referent, and will measure the corresponding angles. Figure 66 illustrates this process; for each crime site, measure the angle between the ray from the anchor point to the crime with the referent ray from the anchor point to the centroid of the series.

If the bivariate normal distribution is a good model for the two-dimensional distribution of offense sites, then we expect that the distribution of these angles should be roughly uniform. This is not, however what is observed. In fact, nearly all of the crime sites are located along the same direction as the path from the anchor point to the centroid. Figure 67 shows the angles for the 322 residential burglary series in Baltimore County, and in nearly every case the direction to the crime site is close to the direction of the centroid.

This disagreement between the expected theoretical distribution and the observed is not due to the presence of commuters as it was in the case of the one-dimensional distribution. Working in the same fashion as the one-dimensional distance decay curves, we plot the joint distribution of angles versus the commuter / marauder parameter μ_2; this is done in Figure 68. Even for relatively

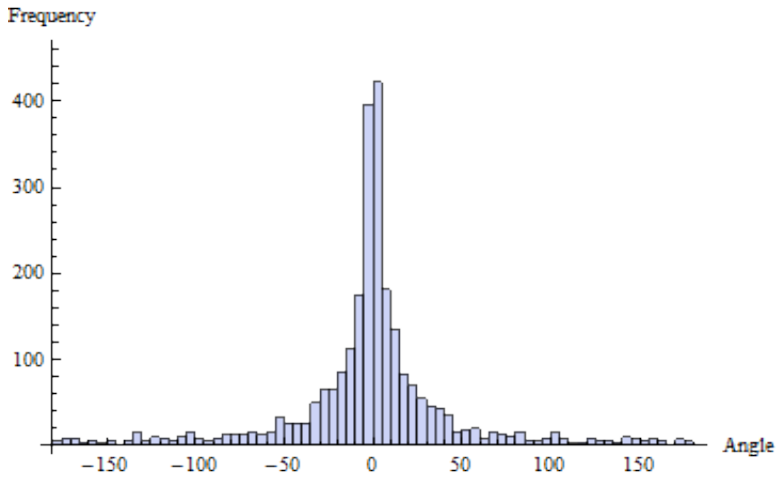


Figure 67. Histogram of angles for 322 solved residential burglary series with at least four crimes in Baltimore County from 1986-2008.

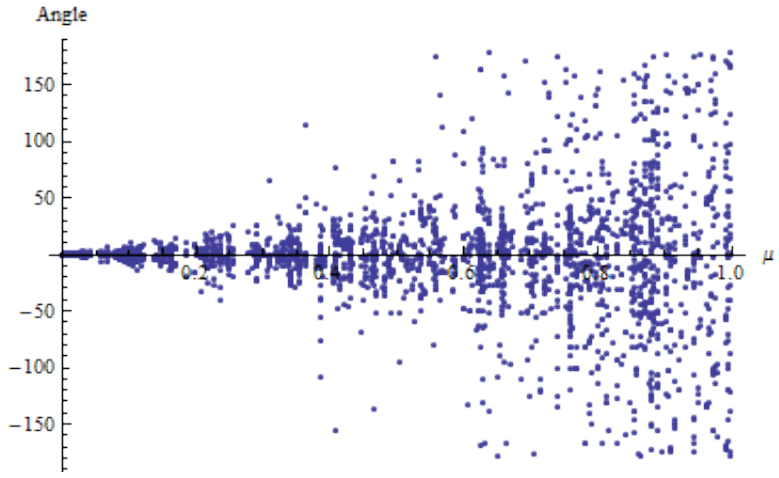


Figure 68. Distribution of angles versus μ_2 for 322 solved residential burglary series with at least four crimes in Baltimore County from 1986-2008.

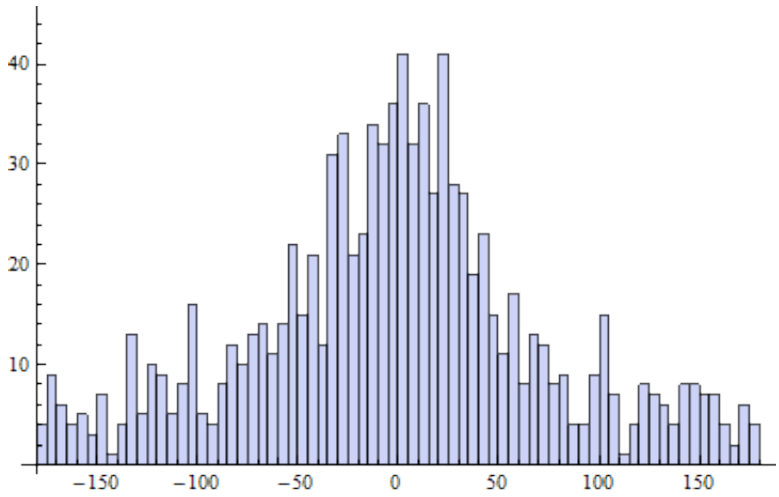


Figure 69. Histogram of angles for solved residential burglary series with at least four crimes in Baltimore County from 1986-2008 with $\mu_2 > 0.7$

large values of μ_2 the observed angles are clustered near the zero angle. This strong tendency to move toward the centroid of the crime series remains no matter how stringently we define commuter. If we remove all of the series save those with $\mu_2 > 0.7$, then despite removing 220 of the 322 series, we still end up with a strong central peak as seen in Figure 69.

To better understand what is being observed, it makes sense to go back to the full bivariate distribution rather than just the angular dependence. To do so, we scale the data set now in three ways. The anchor point for the crime series is set by default at the origin of our coordinate system. Distances from the anchor point to the individual crime sites are scaled again by dividing by the average distance $\hat{\alpha}$ that the offender is willing to travel. Angles are scaled so that the direction from the anchor point to the centroid of the crime site is along the x-axis. Together these have the effect of scaling the centroid of the crime series so that it is located at the point $(1, 0)$. If the underlying distribution of offense sites is bivariate normal with different average offender distances α, then the result of the scaled distribution should be a bivariate normal distribution centered at the origin.

Figure 70 shows what is observed. Clearly it is not a bivariate normal centered at the origin as it is bimodal with a second peak near $(1, 0)$, the centroid of the crime series. The bimodal nature of the observation is made more clear in Figure 71 where the histogram has been smoothed and plotted as a colored contour plot.

Together, we are left with the conclusion that although the distribution of distances from crime site to offender anchor point seems to follow a Rayleigh distribution, at least for marauders, it is just as clear that the bivariate distribution of offense sites is not bivariate normal. Indeed, it is clear from Figures 70 and 71 that there are significant correlations between the locations of the different crime site locations.

One possible explanation for the observed behavior is that offender's go through a two stage process to select targets where the first select a region to offend, then select targets within that region. To examine this hypotheses we can work to understand the distribution of the crime sites not around the offender's anchor point, but rather around the centroid of the individual crime series.

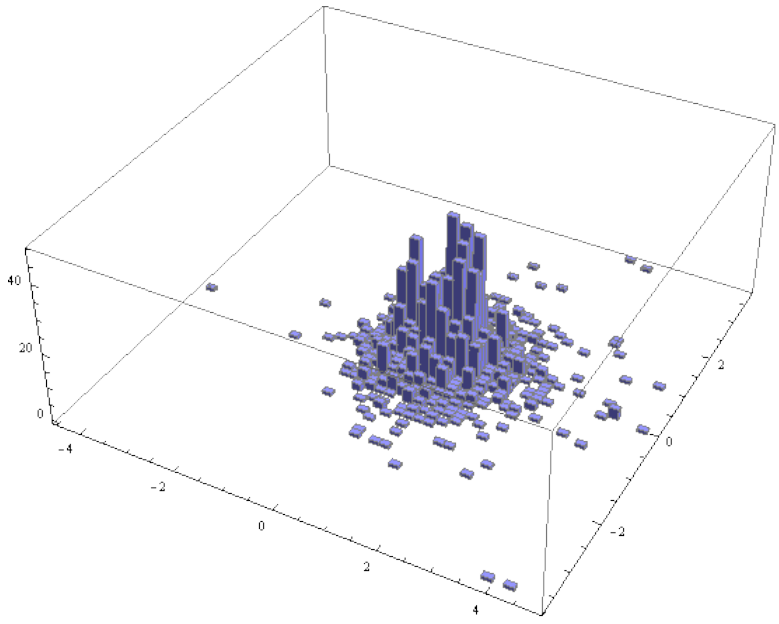


Figure 70. Histogram of scaled crime site locations for 322 residential burglary series with at least four crimes in Baltimore County from 1986-2008. The scaling places the offender's anchor point at the origin and the centroid of the crime series at $(1, 0)$.

In particular, given a crime series with offense locations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, let

$$\mathbf{c} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

be the centroid of the crime series and let

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

be a length scale. Note that $\hat{\alpha}_c$ may be different than, and even unrelated to the offender's average offense distance. We can then consider the set of scaled distances from the crime site to the centroid of the series in the form

$$\rho = \frac{d(\mathbf{x}, \mathbf{c})}{\hat{\alpha}_c}.$$

If each individual offender selects a target from a bivariate normal distribution centered at the centroid of the crime series but with different average distance $\hat{\alpha}_c$, we again expect that the scaled distances should behave like a Rayleigh distribution with mean one, regardless of the offender. Aggregating this presumed identical behavior across all offenders, we should still obtain a Rayleigh distribution with mean one. The observed data for the 322 residential burglary series with at least four crimes in Baltimore County is shown in Figure 72 where it is also compared with the Rayleigh distribution with mean one. We again obtain a solid match between the observed and expected data. More interestingly, this match occurred without the necessity of screening out the commuters; in fact Figure 72 shows a match across all offenders.

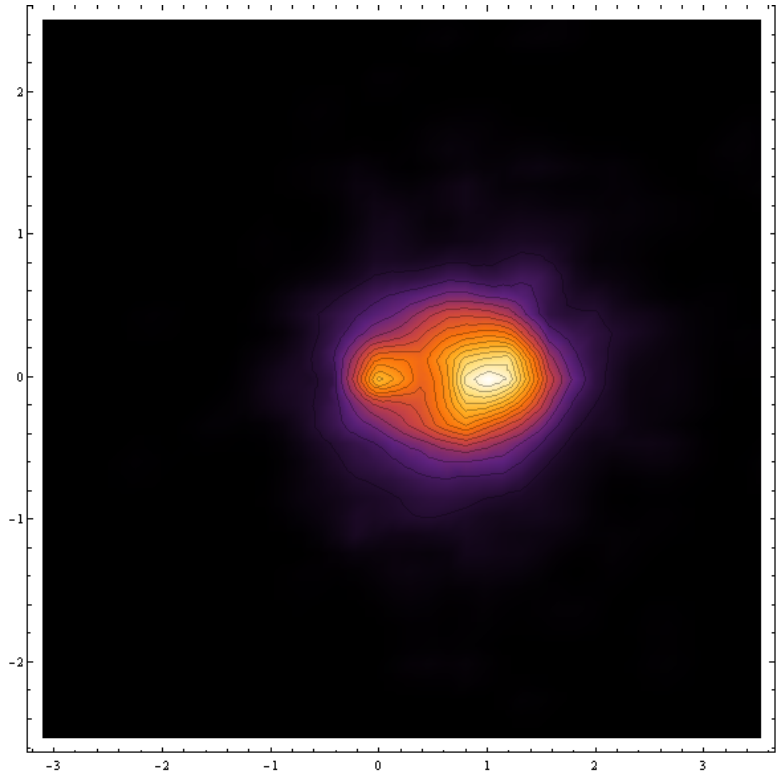


Figure 71. Smoothed histogram of scaled crime site locations for 322 residential burglary series with at least four crimes in Baltimore County from 1986-2008. The scaling places the offender's anchor point at the origin and the centroid of the crime series at $(1, 0)$.

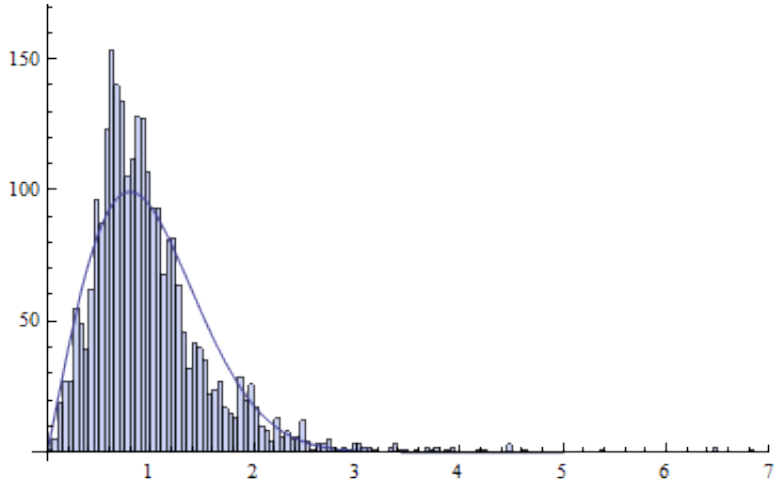


Figure 72. Histogram of scaled distances from the crime site to the centroid of the series for residential burglary series in Baltimore County with at least four elements from 1989-2008 compared with a Rayleigh distribution with mean one.

As we did previously, we can look at the underlying bivariate distribution by first considering the angular dependence. Now the reference ray starts at the centroid of the offender's series and through the offender's home base; the corresponding angle is then between that ray and the ray from the centroid of the crime series through each individual crime site; see Figure 73

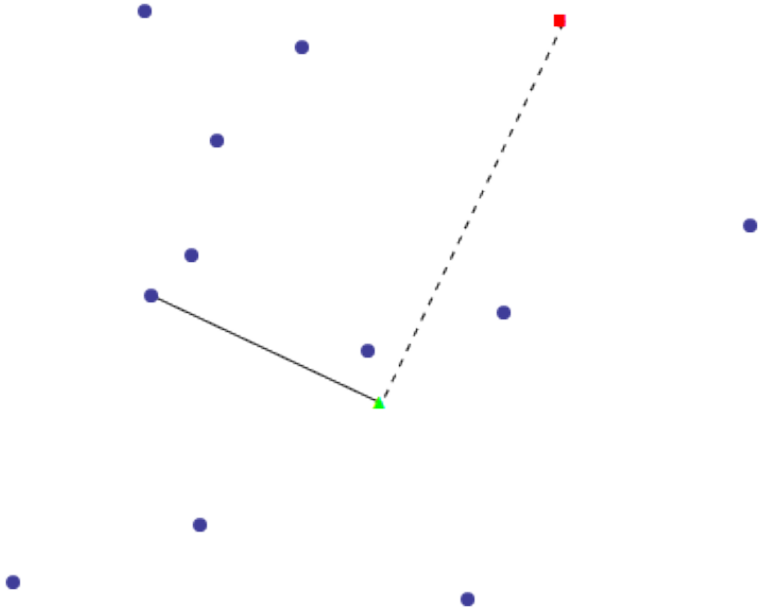


Figure 73. Measuring angles. Blue dots represent crime locations, the red square the offender's anchor point, and the green triangle the centroid of the crime series. For each crime site, measure the angle between the ray from the centroid to the crime with the referent ray from the centroid to the offender's anchor point.

In this case, we find that the uniform distribution is a much better distribution for the observed angular dependence, though some anisotropy remains. Looking at Figure 74 we observe more crimes than expected either directly toward or away from the offender's anchor point, while we observe fewer crimes at right angles to that direction.

The full scaled bivariate distribution is shown in Figure 75 where it is directly compared with a bivariate normal distribution At first glance, it appears to be a solid match for a bivariate normal distribution, but a more careful analysis of the residuals shows that the observed data is larger than predicted directly towards and away from the offender's anchor point, while it is below the prediction in the orthogonal directions. Smoothing and replotting the histogram as is done in Figure 76 shows that the observed data is bimodal, with peaks both towards and away from the offender's actual anchor point, and the larger peak in the direction of the anchor point and the smaller shadow peak directly away from the anchor point.

Together these let us conclude that, at least for residential burglary series in Baltimore County from 1986-2008 with at least four crimes, that there is evidence that the distribution of crime sites is roughly bivariate normal and centered around the centroid of the crime series. There are, however, some noticeable deviations from normality, as directions in line with the offender's home appear preferred to perpendicular directions. We also note that there is a preference for crime sites closer

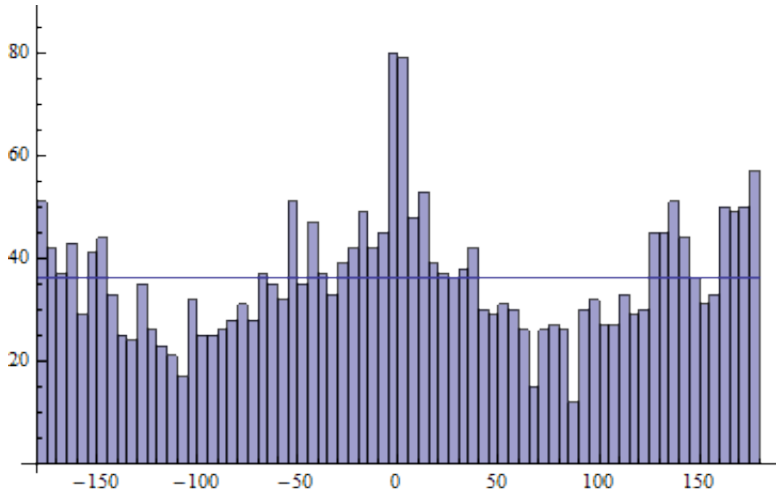


Figure 74. Histogram of angles to the centroid of the crime series for 322 solved residential burglary series with at least four crimes in Baltimore County from 1986-2008.

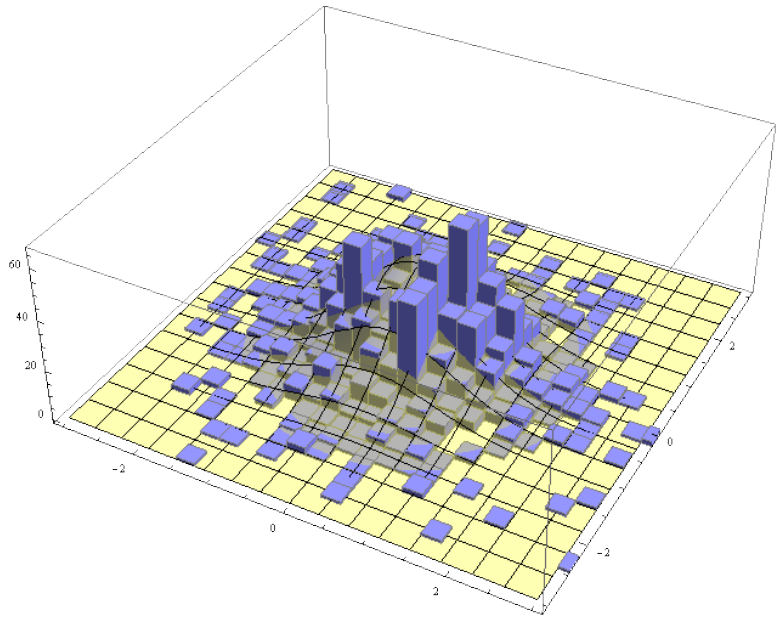


Figure 75. Histogram of scaled crime site locations for 322 residential burglary series with at least four crimes in Baltimore County from 1986-2008. The scaling places the centroid of the crime series at the origin and the anchor point on the positive x-axis.

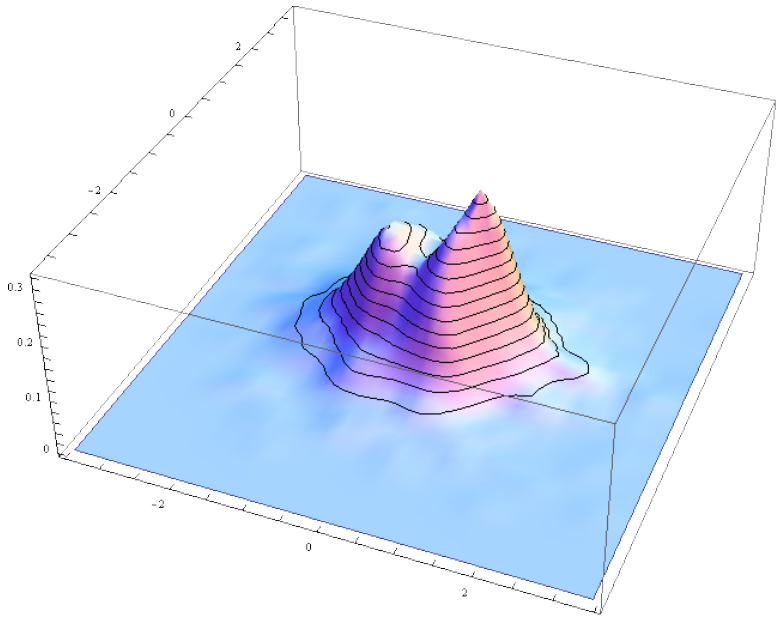


Figure 76. Smoothed histogram of scaled crime site locations for 322 residential burglary series with at least four crimes in Baltimore County from 1986-2008. The scaling places the centroid of the crime series at the origin and the anchor point on the positive x-axis.

to the offender's home than locations farther away.

Our experience with the match between the Rayleigh distribution and the scaled distance to offense for non-commuters however tells us that we need to be careful not to draw overbroad conclusions from this fit. The conclusion that the bivariate distribution of offenses is well modeled by a bivariate normal distribution centered at the centroid of the crime series is of limited explanatory value. Indeed, this characterizes only the observations of the crime sites, but does not explain why the offender would choose those particular sites. In contrast, the original posited model for offender behavior, namely a bivariate normal centered at their anchor point, is much more explanatory, as it predicts the selection of the crime sites from characteristics of the offender, rather that from characteristics of previous offense sites. However, we have already seen that there are significant problems with that bivariate model. Thus, though we have been able to show some fits between some theoretical models and data, it is just as clear that the much work still needs to be done to improve these models.

Models for Offender Behavior with Explicit Dependency Structures

This evidence that the distribution of crime sites can be modeled with a bivariate normal distribution centered around the centroid of the crime series does at least suggest that there may be some validity to a two stage model for offender behavior. To begin to understand this question, in 2012 I began work with a Master's student, Jeremiah Tucker, to develop models for offender behavior that explicitly allow for a dependency structure in the location of the crime sites, meaning that the offender chooses crime site locations not only by their relationship to the offender's anchor point but also based on the location(s) of previously successful offenses. So far this work has been

fruitful, but it is, as yet, not complete. This section will describe the results obtained to date.

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 three models of offender behavior. The first approach, we shall call the normal model, and it is given by

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

This is the same model that we have already described and discussed extensively; it will be used as a baseline for comparison. Notice that subsequent crime sites are chosen independently of the locations of prior crime sites.

The second model we shall call the near repeat model. In this model we assume that the first offense site of the offender is chosen according to the normal model. However, for the second and subsequent offenses, sometimes the offender's target is a repeat or a near repeat of a previous element of the series while other times the offender selects a new target using a bivariate normal distribution centered at the offender's anchor point. Mathematically the probability density function for the second and subsequent crimes has the form

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

where ϵ_{rep} is fixed and small- say 0.005 mi = 26 ft. The distribution of repeat offenses around previous offenses has been modeled with a bivariate normal distribution with average distance ϵ_{rep}. Clearly other distributions could have been chosen; the advantages of the bivariate normal include consistency with the previous results that suggest that the behavior of offenders is well modeled by a bivariate normal centered at the centroid of the crime site. Because $\epsilon_{\text{rep}} > 0$, we allow for both repeat and near-repeat behaviors, and also avoid some of the technical difficulties that would arise if a Dirac delta distribution had been used instead.

The mixture parameter γ represents the probability that the offender chooses to offend at or near a previous target rather than selecting a new target site. Note that all previous target locations are treated equally; there is no preference for earlier or for later elements in the series. Porter and Reich (2012) have begun examining temporal roles in offender crime site selection.

When $\gamma = 0$, the near repeat model reduces to the normal model, however when $\gamma > 0$, the model is fundamentally different as the location of subsequent crime sites will depend directly on the locations of the previous crime sites.

The final model we call the general model. Like the near repeat model, the first crime is chosen according to the normal model. Subsequent offense sites are then chosen according to

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

The only difference between the general model and the near repeat model is that in the general model the parameter ϵ is now unknown and can vary between offenders.

When interpreting the mathematical statements of these models, one must be precise. For example, one interpretation of the general model is that the offender first determines if they plan to offend at or near a previous crime site according to the parameter γ, then with that chosen they then determine the location to offend. Though this is consistent with the mathematical model, it is important to realize that this is not the only way the offender could behave to select targets. Indeed, the offender's decision making process my not involve a random component at all, but instead the randomness in the model may reflect only our lack of knowledge of the offender's behavior and motivations. In particular there are many potential behaviors that have the same mathematical model, and we are not claiming that the offender behaves in any particular way, only that their target selection process can be well-modeled by the given probability density function.

To investigate which of these three models is most useful, we will employ the Akaike Information Criterion with the small sample correction, called the AICc. In this we follow the usual mathematical techniques which are well explained by Burnham and Anderson (2002, Chapter 2). Suppose that the elements of the crime series take place at the locations $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. Then the likelihood function for the normal model depends only on the offender's average offense distance α and the anchor point $\mathbf{z}$ and has the form

$$L(\mathbf{z}, \alpha) = \prod_{i=1}^{n} P(\mathbf{x}_i \,|\, \mathbf{z}, \alpha) = \prod_{i=1}^{n} \left[\frac{1}{4\alpha^2} \exp\left(-\frac{\pi}{4\alpha^2} |\mathbf{x}_i - \mathbf{z}|^2 \right) \right].$$

Computationally, it is usually easier to use the log-likelihood function $\lambda(\mathbf{z}, \alpha) = \ln L(\mathbf{z}, \alpha)$ which has the form

$$\lambda(\mathbf{z}, \alpha) = -n \ln(4\alpha^2) - \frac{\pi}{4\alpha^2} \sum_{i=1}^{n} |\mathbf{x}_i - \mathbf{z}|^2.$$

The likelihood and log-likelihood functions for the explicitly dependent models need to account for that dependence, so for the near-repeat model we have

$$\begin{aligned} \lambda(\mathbf{z}, \alpha, \gamma) = &- \ln(4\alpha^2) - \frac{\pi}{4\alpha^2} |\mathbf{x}_1 - \mathbf{z}|^2 \\ &+ \sum_{i=2}^{n} \ln \left[\frac{\gamma}{i-1} \sum_{j=1}^{i-1} \frac{1}{4\epsilon_{\text{rep}}^2} \exp\left(-\frac{\pi}{4\epsilon_{\text{rep}}^2} |\mathbf{x}_i - \mathbf{x}_j|^2 \right) + \frac{1-\gamma}{4\alpha^2} \exp\left(-\frac{\pi}{4\alpha^2} |\mathbf{x}_i - \mathbf{z}|^2 \right) \right] \end{aligned}$$

and for the general model we have

$$\begin{aligned} \lambda(\mathbf{z}, \alpha, \gamma, \epsilon) = &- \ln(4\alpha^2) - \frac{\pi}{4\alpha^2} |\mathbf{x}_1 - \mathbf{z}|^2 \\ &+ \sum_{i=2}^{n} \ln \left[\frac{\gamma}{i-1} \sum_{j=1}^{i-1} \frac{1}{4\epsilon^2} \exp\left(-\frac{\pi}{4\epsilon^2} |\mathbf{x}_i - \mathbf{x}_j|^2 \right) + \frac{1-\gamma}{4\alpha^2} \exp\left(-\frac{\pi}{4\alpha^2} |\mathbf{x}_i - \mathbf{z}|^2 \right) \right]. \end{aligned}$$

The maximum likelihood estimate of the parameters are then the values of the parameters that make the likelihood, or equivalently the log-likelihood as large as possible. Clearly these values depend on the model and the precise locations of the crime sites.

The value of AICc is then given by

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

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	$\alpha, \gamma, \epsilon, (z^{(1)}, z^{(2)})$

Table 2
Parameters to be estimated in various models

where L is the value of the likelihood at the maximum likelihood estimate, while k is the number of parameters in the problem and n is the number of data elements- in our case the number of elements in the crime series.

The larger the value of AICc (or AIC) the less evidence the data provides in favor of the model. When comparing multiple models, it should be noted that the relevant information is the difference in the values of AICc (or AIC). Differences in AIC of 10 or more indicate that the model with larger AICc (or AIC) value is essentially unsupported; differences in AIC of 2 or less indicate that the models provide roughly the same explanatory value.

When multiple models are being considered, inference can be drawn from all of the models at the same time through the use of Akaike weights. Given a collection of m models, the weight for model i is given by

$$w_i = \frac{\exp\left(-\frac{1}{2}\text{AICc}_i\right)}{\sum_{j=1}^{m} \exp\left(-\frac{1}{2}\text{AICc}_j\right)}.$$

Together the weights sum to one and provide a relative measure of the value of one model over another. Models with weights close to one are the most supported by the data under consideration while models with weights near zero are the least supported.

Note that the AICc (and AIC) differences as well as the Akaike weights are not absolute numbers but only relative to the collection of models under consideration. If all of the models for a problem under study are fatally flawed, it will still be the case that one of them will be the most supported by the data- despite the fact that it remains fatally flawed.

To compare the normal, near repeat, and general model we will use the data sets from Baltimore County for residential burglary, non-residential burglary, and bank robbery that we have already used. Before we begin the analysis though, we pause to consider the role of the offender's anchor point. The data consists of solved crimes, so that we can consider the anchor point to be known, or we can put ourselves in the position of an analyst working on an unsolved series, and consider the anchor point as an unknown that needs to be estimated. Thus, there are effectively six possible models depending on the precise combination of parameters to be estimated; see table 2.

To even calculate AICc, one needs two more data points than parameters, so we analyzed series with at least seven elements. Moreover, in many crime series an offender commits multiple crimes in the same location on the same day; for example by robbing multiple storage lockers or breaking into multiple offices in a single building. In this analysis, multiple crimes in the same place on the same day were combined into a single incident. Doing so for the Baltimore County data at our disposal, we have 136 residential burglary series, 43 non-residential burglary series and 10 bank robbery series with at least seven crimes.

The process of finding the actual maximum likelihood estimates for the different models is non-trivial as most of the models do not admit of a simple algebraic solution that provides the parameters. Instead, the tool *Mathematica* was used to locate the maximum likelihood using the built in function NMaximize and the Nelder Mead method. Because Nelder Mead is an iterative method, it need not converge to the actual maximum. Care was taken to ensure that the algorithm provided good approximations to the actual maximum likelihood solution by running the algorithm multiple times with multiple random seeds and by using suitably restricted search spaces.

For the non-residential burglaries where the home is considered known, the normal model is by far the least supported by the data; in 38 of the 43 series, the weight accorded to the normal model was essentially zero. Both the near repeat model and the general model were much more supported, with the near repeat model generally preferred to the general model. A complete histogram of the weights for all three models is shown in Figure 77.

The lack of support for the normal model continues with bank robberies; when the offender's home is considered known, then in 8 of the 10 series the weight for the normal model was essentially zero; see Figure 78. Again, the near repeat model was the most supported while showing essentially no support for the general model.

The situation for residential burglaries is more complex, and shown in Figure 79. It is still the case that the normal model is least supported, receiving Akaike weights of roughly zero for 77 of the 136 series. Though the normal model was essentially unsupported almost half the time, this is quite different than the behavior for bank robberies and non-residential burglaries where the lack of support for the normal model was almost complete. Moreover all three models are considered well supported by the data at least some of the time.

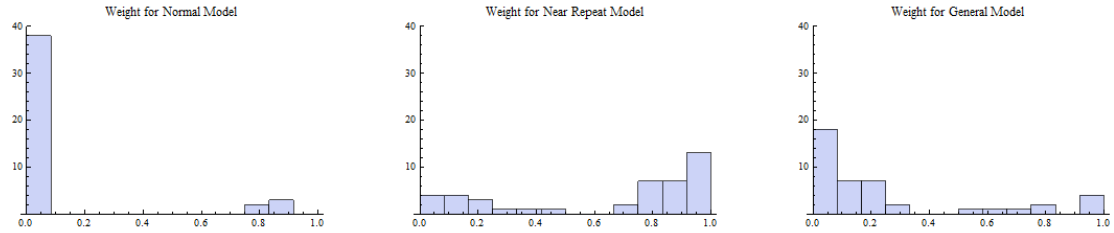


Figure 77. Histograms of Akaike weights for 43 non-residential burglary series in Baltimore County with at least seven crimes where the offender's anchor point is considered unknown.

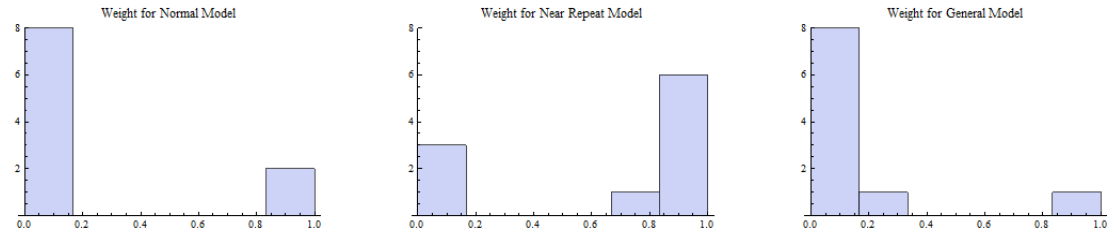


Figure 78. Histograms of Akaike weights for 10 bank robbery series in Baltimore County with at least seven crimes where the offender's anchor point is considered unknown.

Table 3 summarizes which of the three models is most supported, by crime type. For non-residential burglary and bank robbery the near repeat model is seen to be strongest, but the residen-

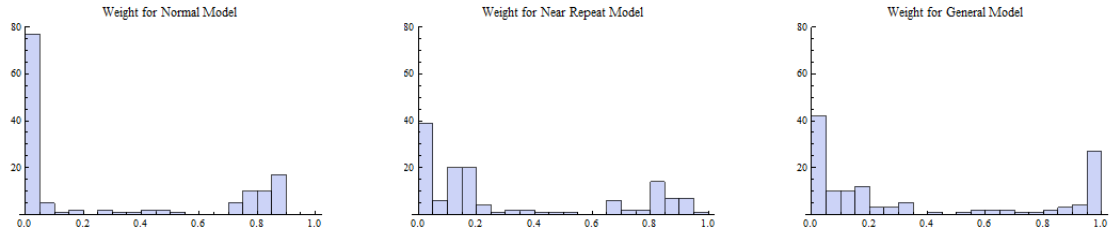


Figure 79. Histograms of Akaike weights for 136 residential burglary series in Baltimore County with at least seven crimes where the offender's anchor point is considered unknown.

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

Table 3
Most supported model, by crime type, where the offender's home is considered known.

tial burglary series data provides support evenly split across the three models.

The mixture parameter γ has particular significance, as it represents the fraction of the offenses that can be explained as a repeat or a near repeat of a previous crime in the series. Figures 80, 81, and 82 show how γ is distributed for different offense types. It is interesting to see that when either the near repeat model or the general model are considered to be the most supported, then usually 20, 30, 40% or more of the data in the series is apparently explained as being better modeled by repeat or near-repeat phenomena than as a new independent crime site. Though the significance of this as a criminological matter is unclear, it does provide additional compelling evidence to suggest that it is essential that researchers begin to examine the dependency of subsequent crime site location choices on the locations of previously successful offenses.

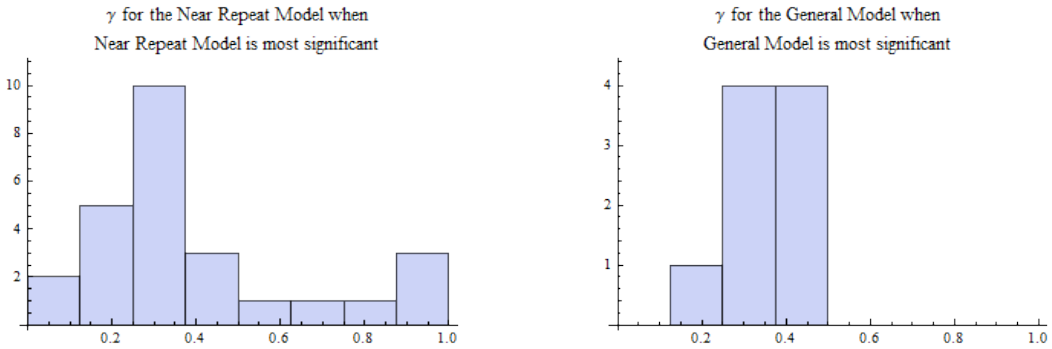


Figure 80. Histograms of the mixture parameter for non-residential burglaries.

As another method of examining the dependence of one crime site location on another, given a series we can simply count the fraction of the crimes in the series that are close- say within a distance $\epsilon_{\text{rep}} = 0.005$ mi. $= 26$ ft. of a previous element in that series. Clearly for each series this number is between zero and one; we plot histograms of this number for all three considered crime

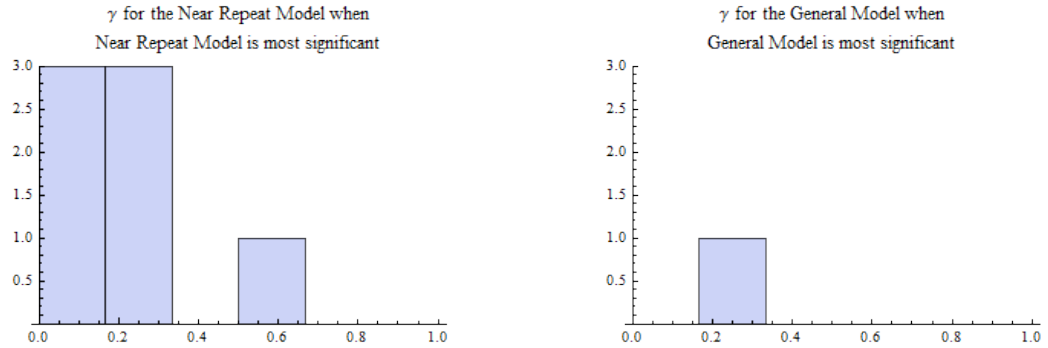


Figure 81. Histograms of the mixture parameter for bank robberies.

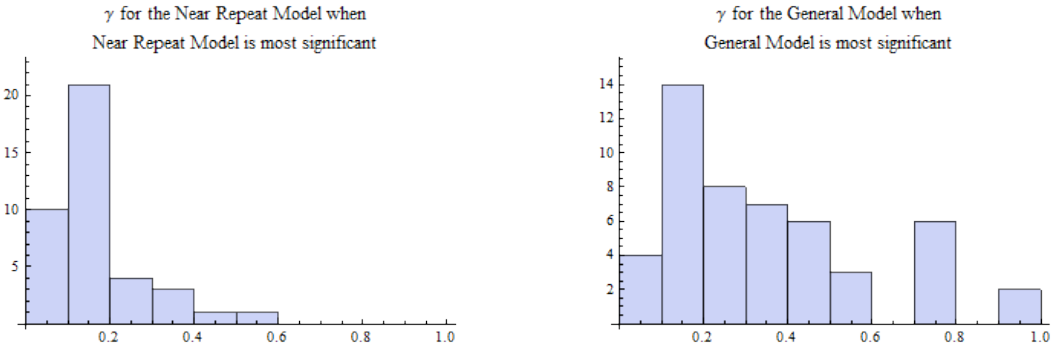
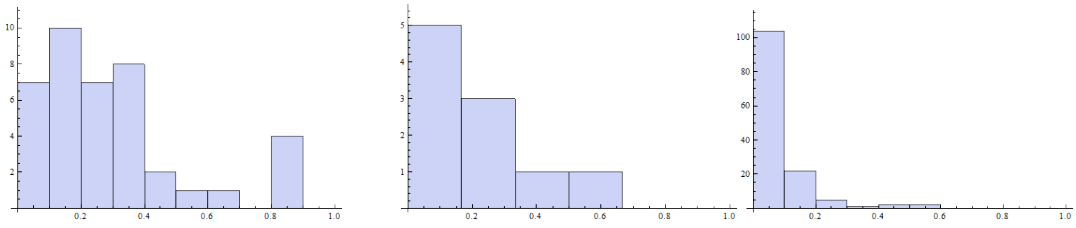


Figure 82. Histograms of the mixture parameter for residential burglaries.

types in Figure 83.



(a) Non-residential burglaries (b) Bank Robberies (c) Residential burglaries

Figure 83. Histogram of the fraction of the crimes within ϵ_{rep} aggregated across series

What is striking in this figure is the fact so many series have so many repeat or near repeat offenses; for non-residential burglaries for four of the 43 series, 80% or more of the crimes were repeats or near repeats of previous crimes. Again, the significance of this result is clearly that it is essential to account for the locations of prior crimes when modeling how offenders select the location of subsequent elements of their crime series.

Although the calculations so far have been performed where the offender's anchor point was considered known, similar results are found when the anchor point is considered unknown. Figures 84, 85 and 86 show the weights for each of the three models. Again for non-residential burglaries

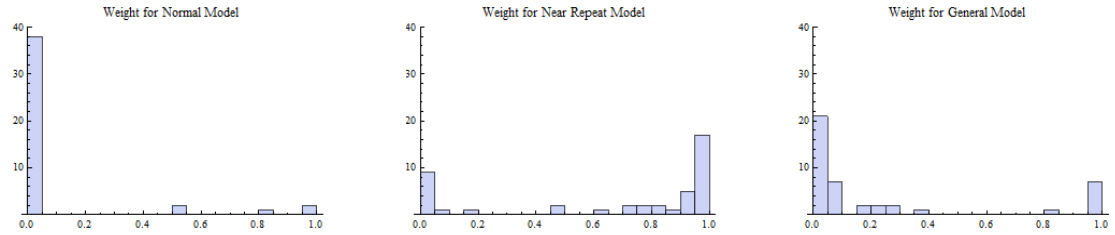


Figure 84. Histograms of Akaike weights for 43 non-residential burglary series in Baltimore County with at least seven crimes where the offender's anchor point is considered known.

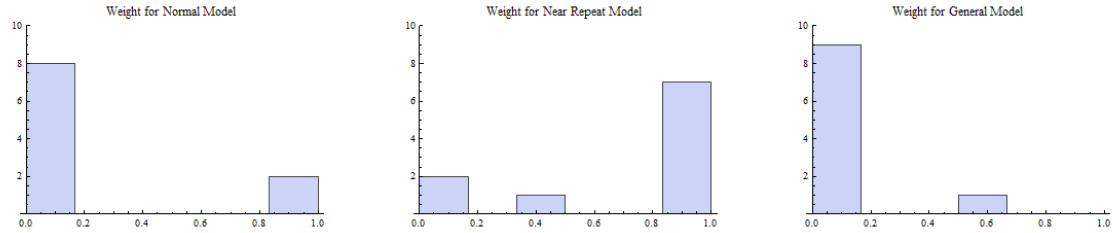


Figure 85. Histograms of Akaike weights for 10 bank robbery series in Baltimore County with at least seven crimes where the offender's anchor point is considered known.

the normal model is essentially unsupported, with the near repeat model most supported; this is also observed for bank robberies. The case of residential burglaries is again different than the previous two cases, but now the near repeat model is least supported, but there are some series where either the normal model or the general model are most supported by the data. Table 4 shows which model is considered the most supported by the data; for non-residential burglaries and bank robberies the near repeat model performed the best, while the general model appeared best most often for residential burglary.

The analysis of models of offender behavior with explicit dependency structure is still in its earliest stages, and much work remains to be done.

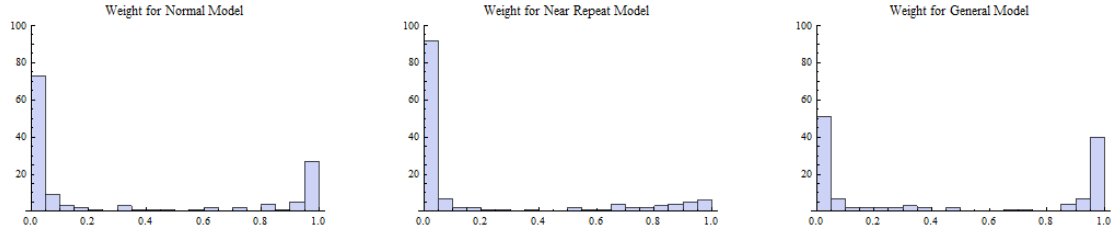


Figure 86. Histograms of Akaike weights for 136 residential burglary series in Baltimore County with at least seven crimes where the offender's anchor point is considered known.

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

Table 4
Most supported model, by crime type, where the offender's home is considered unknown.

Prototype Modification for Non-Independent Behavior

The original prototype was based on the mathematical model (3)

$$P(\mathbf{z} \mid \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) \propto \int \left[\prod_{i=1}^{n} P(\mathbf{x}_i \mid \mathbf{z}, \alpha) \right] H(\mathbf{z}) \pi(\alpha) \, d\alpha \tag{3}$$

that we have already described in detail.

After the round-off error correction was made, the original prototype was tested on 237 residential burglary series in Baltimore County, MD. Analysis of these results showed something very interesting. As the number of crime sites increased, the size of the resulting search area decreased, sometimes quite dramatically. Series with large number of crimes across large swaths of Baltimore County were given offender search areas that could be as small as a few blocks. Consider for example, Figure 87, which shows the search area generated by the original prototype for a large series of crimes that takes place across a wide area. Despite the large activity space of the offender, the original prototype based on (3) produces a comparatively small search area.

In hindsight, this is a correct deduction from the original model and (3). When sampling from an unknown distribution, the larger the number of data points in the sample, the tighter the bounds that can be made on the parameter. Indeed, in the perhaps canonical example, if n values are independently selected from a one-dimensional distribution with mean μ and variance σ^2, then the variance of the sample mean $\bar{X}$ satisfies

$$\operatorname{var} \bar{X} = \frac{\sigma^2}{n}.$$

Thus, as n increases, the size of the region most likely to contain μ decreases like $1/\sqrt{n}$. The prototype was behaving in a similar fashion, though the underlying distributions were more complex.

Though this shrinking of the search area as the number of crimes increases is mathematically correct, it turns out that is incorrect as a criminological predictive tool; indeed the original prototype was only successful 59% of the time for marauders and 13% of the time for commuters.

That fact that the search area should decrease with every new crime site relies on the fundamental assumption that the locations of the crimes are independent on one another, and that each new crime adds an additional piece of information. However we have now seen that there is significant evidence to suggest that this is not the case, and that crime series locations show significant interdependences.

The proper solution to this problem would be to develop an appropriate mathematical theory that better explains the dependencies in the distribution; though this work is ongoing as we have already seen it is far from complete. On the other hand, to develop a tool with practical value for

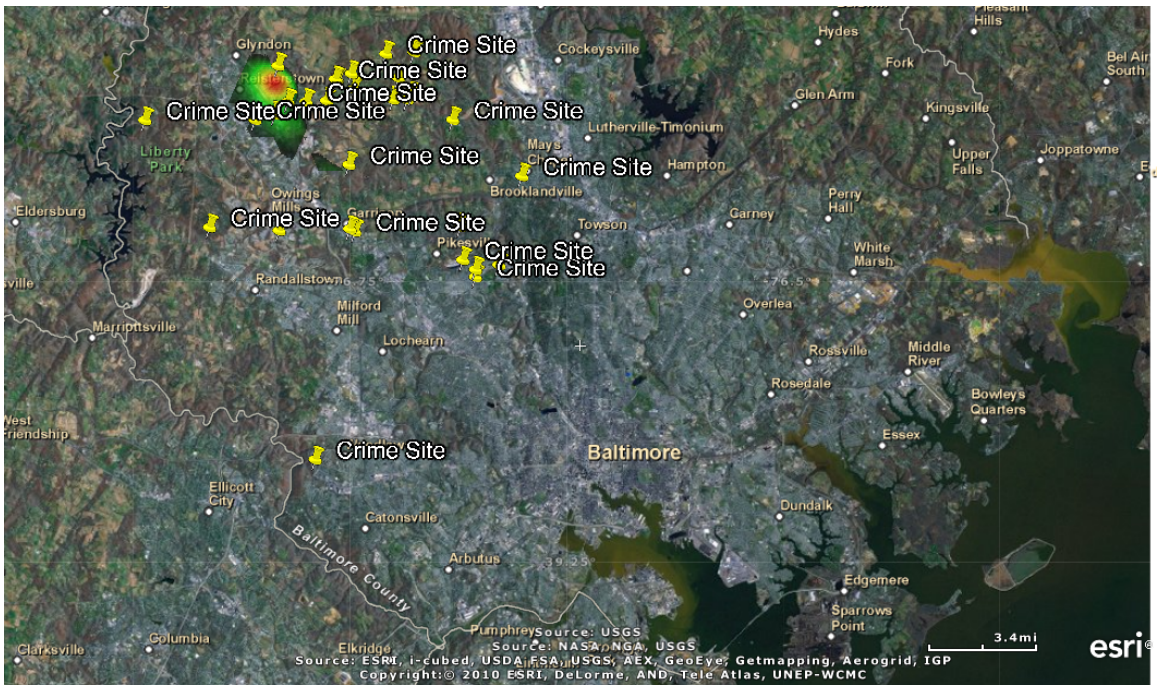


Figure 87. Running the original prototype on a crime series. Notice how, despite a large number of crimes, the search area is quite small. This behavior has been modified in the current prototype.

law enforcement officers, we need to find some way to account for this non-independence in the prototype. The current solution is to try to reduce the effect of adding additional information to the model. In the model generated by (3), each time the location of an additional crime $\mathbf{x}_i$ is added, the amount of information reflected in the number of multiplied terms. To reduce that information, the prototype was re-run, but with the model

$$P(\mathbf{z} \mid \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) \propto \int \left[\prod_{i=1}^{n} P(\mathbf{x}_i \mid \mathbf{z}, \alpha) \right]^{\frac{1}{\ln n}} H(\mathbf{z}) \pi(\alpha) \, d\alpha \tag{5}$$

Mathematically, this cannot (as yet) be rigorously justified. The motivation behind the choice is to assume that additional crime sites provide additional information, but rather than assuming that n crimes provide n pieces of information, we instead assume they provide $n/\ln n$ pieces. Despite the lack of formal justification, the approach does appear to provide significant benefits. Re-running the prototype on the same 237 residential burglary series showed dramatic improvements- the tool was successful on 94% of marauders and 47% of commuters. For this reason, this is the approach that has been taken in the currently released version of the prototype.

Conclusions

We have developed and released a new version of our tool for the geographic profiling tool. Preliminary tests of its accuracy on data residential burglary series and non-residential burglary series from Baltimore County show that the prototype's search area contained the offender's actual

anchor point for 74% of the non-residential burglaries and 70% of the residential burglaries while using a search area that is comparable in size to the Canter circle of the series.

We have also looked deeply into foundational mathematical models of distance decay, and have found significant evidence to suggest that individual distance decay behavior can be well-modeled by a Rayleigh distribution. We also found that the process of offender target selection shows a significant dependence on the locations of previous crime sites.

Implications for Policy and Practice

Because the prototype, its source code, and its underlying mathematical foundation have all been released, this research can benefit policy and practice directly through the use of the prototype directly by interested police agencies, and indirectly, by allowing others with more established tools for the geographic profiling problem like CrimeStat, to easily incorporate these new ideas into their tools.

Implications for Future Research

This project leaves a number of important open research questions, some mathematical, some criminological, and some operational. These questions include the *accuracy problem*, the *resolution problem*, the *effectiveness problem*, the *geography problem* and lastly the *computation problem*. To help tease out the differences between questions about the mathematics, and questions about offenders, let us begin with a simpler foundational example that we can use to illustrate the mathematics without the added complexity of offender behavior.

Foundational Example. Suppose that someone is choosing locations randomly by, for example, throwing darts at a board. Our person is aiming at a particular target, but due to various human and environmental factors, they do not necessarily hit their target. Our goal is to estimate the location of the target from the observed pattern of the darts on the board.

Unlike the case of a real person, we are going to specify the exact distribution of the observations; in particular we will assume that the selected locations $\mathbf{y}$ follow a bivariate normal distribution $\mathbf{y} \sim N(\theta, \Sigma)$ where θ is the center of the distribution and our human's supposed target; the variance matrix is specified simply as $\Sigma = \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Thus we will assume that the probability distribution of the target locations $\mathbf{y}$ follows

$$P(\mathbf{y}|\theta) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{|\mathbf{y}-\theta|^2}{2\sigma^2}\right). \tag{6}$$

Following the argument used to develop the geographic profiling algorithm discussed here, we use Bayes' Theorem to estimate the distribution $P(\theta|\mathbf{y})$ of the target θ given our knowledge of a single known target $\mathbf{y}$; it is given by

$$P(\theta|\mathbf{y}) = \frac{P(\mathbf{y}|\theta)\pi(\theta)}{\iint_{\mathbf{R}^2} P(\mathbf{y}|\psi)\pi(\psi)d\psi}.$$

We again need to provide a specification of the prior distribution $\pi(\theta)$ that provides our knowledge of the location of the selected target point θ before information from the observation $\mathbf{y}$

is included. We specify that θ also follows a bivariate normal distribution $\theta \sim N\left(\mu, \gamma^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$ with known mean μ and known variance matrix $\gamma^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$; then

$$\pi(\theta) = \frac{1}{2\pi\gamma^2} \exp\left(-\frac{|\theta - \mu|^2}{2\gamma^2}\right).$$

Now

$$P(\mathbf{y}|\theta)\pi(\theta) = \frac{1}{4\pi^2\sigma^2\gamma^2} \exp\left(-\frac{|\mathbf{y} - \theta|^2}{2\sigma^2} - \frac{|\theta - \mu|^2}{2\gamma^2}\right),$$

while a little algebra will show

$$\frac{|\mathbf{y} - \theta|^2}{2\sigma^2} + \frac{|\theta - \mu|^2}{2\gamma^2} = \frac{1}{2}\frac{\sigma^2 + \gamma^2}{\sigma^2\gamma^2}\left[\left|\theta - \frac{\mu\sigma^2 + \mathbf{y}\gamma^2}{\sigma^2 + \gamma^2}\right|^2 + K_1\right]$$

for a constant $K_1 = K_1(\mathbf{y}, \sigma, \mu, \gamma)$ independent of θ. We also notice that

$$m(\mathbf{y}) = \iint_{\mathbf{R}^2} P(\mathbf{y}|\psi)\pi(\psi)\, d\psi = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi\sigma\gamma} \exp\left(-\frac{|\mathbf{y} - \psi|^2}{2\sigma^2} - \frac{|\psi - \mu|^2}{2\gamma^2}\right) d\psi^{(1)}\, d\psi^{(2)}$$

is also independent of θ. Thus

$$P(\theta|\mathbf{y}) = \frac{K_2}{2\pi\dfrac{\sigma^2\gamma^2}{\sigma^2 + \gamma^2}} \exp\left(-\frac{\left|\theta - \dfrac{\mu\sigma^2 + \mathbf{y}\gamma^2}{\sigma^2 + \gamma^2}\right|^2}{2\dfrac{\sigma^2\gamma^2}{\sigma^2 + \gamma^2}}\right)$$

for a constant $K_2 = K_2(\mathbf{y}, \sigma, \mu, \gamma)$ independent of θ. By integrating both sides of this relation and using the fact that P is a probability density in θ, we prove that $K_2 = 1$ and conclude

$$\theta|\mathbf{y} \sim N\left(\theta \left| \frac{\mu\sigma^2 + \mathbf{y}\gamma^2}{\sigma^2 + \gamma^2}, \frac{\sigma^2\gamma^2}{\sigma^2 + \gamma^2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right.\right). \tag{7}$$

From this, we see that the probability distribution for the target point θ given knowledge of a single sample $\mathbf{y}$ is simply another bivariate normal distribution. The mean of this distribution is

$$\frac{\mu\sigma^2 + \mathbf{y}\gamma^2}{\sigma^2 + \gamma^2} = \frac{\sigma^2}{\sigma^2 + \gamma^2}\mu + \frac{\gamma^2}{\sigma^2 + \gamma^2}\mathbf{y}$$

which is just a weighted average of the data $\mathbf{y}$ and the center μ of the prior distribution π. Notice that if the variance of the data σ is much smaller than the variance in the prior γ then the estimate of θ is close to the data point $\mathbf{y}$, while if the variance in the prior γ is much smaller than the variance in the data σ then the estimate of θ is much closer to the center μ of the prior π.

Further, the size of the variance of our estimate of the target θ is simply the harmonic mean

$$\frac{\sigma^2\gamma^2}{\sigma^2 + \gamma^2} = \left(\frac{1}{\sigma^2} + \frac{1}{\gamma^2}\right)^{-1}$$

of the size of the variance of the data σ^2 and the size of the variance of the prior γ^2.

Now suppose that we have more data, and that $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ are all independently sampled, so that

$$\mathbf{y}_i \sim N\left(\theta, \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

for each i. Then it is simple to use moment generating functions to show that the mean $\bar{\mathbf{y}} = \frac{1}{n}\sum \mathbf{y}_i$ is also normally distributed, with

$$\bar{\mathbf{y}} \sim N\left(\theta, \frac{\sigma^2}{n} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right).$$

Then arguing as above, we know that

$$\theta|\bar{\mathbf{y}} \sim N\left(\theta \,\middle|\, \frac{\mu(\sigma^2/n) + \bar{\mathbf{y}}\gamma^2}{(\sigma^2/n) + \gamma^2}, \frac{(\sigma^2/n)\gamma^2}{(\sigma^2/n) + \gamma^2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right). \tag{8}$$

We see that the parameter θ is distributed as a bivariate normal, but now the mean is

$$\frac{(\sigma^2/n)}{(\sigma^2/n) + \gamma^2}\mu + \frac{\gamma^2}{(\sigma^2/n) + \gamma^2}\bar{\mathbf{y}}$$

which is a weighted mean of the average of the data $\bar{\mathbf{y}}$ and the center μ of the prior π. As the number of observations increases, we have $\sigma^2/n \to 0$ and so the estimate of θ moves closer to the average of the observations $\bar{\mathbf{y}}$.

Similarly, the variance in our estimate of θ is just the harmonic mean

$$\frac{(\sigma^2/n)\gamma^2}{(\sigma^2/n) + \gamma^2} = \left(\frac{1}{\sigma^2/n} + \frac{1}{\gamma^2}\right)^{-1}.$$

As the number of observations increases we see that this variance tends to zero; indeed we have

$$\frac{(\sigma^2/n)\gamma^2}{(\sigma^2/n) + \gamma^2} = \left(\frac{1}{\sigma^2/n} + \frac{1}{\gamma^2}\right)^{-1} = \frac{\gamma^2}{1 + n(\gamma^2/\sigma^2)} \xrightarrow{n\to\infty} 0.$$

The Accuracy Problem. The *accuracy problem* is to what extent a model actually represents the probability distribution for the anchor point of the offender.

This question can be approached in two different directions. The first is to look at the individual components of the model and validate each one individually. If they are all correct, as they are by hypothesis in our foundational example, then the correctness of the conclusion follows as a matter of logic. So far however, we have only been able to show that there is some support for the form of the model as seen in agreement between the distance decay data and observations. We also know that even so, the models are flawed; this is shown most clearly by the disagreement of the angular distribution of offenses with the observed data.

The most promising area for improvement here appears to be in replacing the assumption that the individual crime sites are selected independently with something more realistic. Today this appears to be a nearly completely open question. Other important areas of further investigation include taking a different approach to the underlying model of offender behavior, like the kinetic

model of Mohler and Short (2012), or to look at incorporating additional data like the temporal sequencing of events along the lines of the work of Porter and Reich (2012).

The question of correctness can also be approached in a different direction by measuring the difference between the predicted distribution and observations. This question is more subtle than it first appears however. If we want to verify that a probability distribution is a good fit for an observable process, the usual approach is to repeat the process a large number of times and perform a goodness of fit test against the hypothesized distribution. Unfortunately, that cannot be done in this case, because the distribution in question is the distribution of the offender's anchor point given the locations of the crimes. We cannot "repeat" this process; an offender only has the one anchor point. We can't directly use information from multiple offenders either, as the distribution for the anchor point depends on the locations of the crime sites; change them and the distribution changes. The approach taken in this project was to examine the percentile rank of the probability density for the offender's actual anchor point (Figure 28) and compare that with expectation. However, it is not clear that this is the best or only way to validate these results. An open question then is to determine better ways to assess the correctness of a hypothesized method for determining a distribution for the offender's anchor point. Such a method would not only be valuable in the present context, it would also be a valuable way to compare different approaches to the geographic profiling problem, an area that has historically been divisive and contentious; see *e.g.* Rich and Shively (2004).

The Resolution Problem. The next open question worth significant inquiry is the question of the *spatial resolution* of the method. Consider an offender who has committed crimes across a very large region- say two or more counties. Regardless of the geographic profiling method selected, one would not expect to be able to develop an accurate geoprofile where the search area is the size of a neighborhood or smaller. Fundamentally, the data is simply too coarse to resolve geographic features on the scale of a neighborhood. The spatial resolution problem then is to understand the relationship between the spatial size of the search area to the spatial size of the crime series for a geoprofiling method, and to develop ways in which the size of the search area can be decreased while retaining accuracy.

To make this idea precise, return to the foundational example. An offender who chooses targets over a large geographic region is analogous to assuming that the σ in the foundational model (6) is large. Then if n points $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ are sampled, then the distribution of the estimate of $\theta|\bar{\mathbf{y}}$ is bivariate normal where the variance in the estimate is

$$\frac{(\sigma^2/n)\gamma^2}{(\sigma^2/n)+\gamma^2}\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} = \frac{\sigma^2/n}{1+(\sigma^2/n\gamma^2)}\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}.$$

The key to notice here is that the size of the resulting variance increases as σ does. There is a fundamental limit to how accurately we can estimate the parameter θ depending on the underlying characteristics of the target selection process through the size of σ. This is the effective spatial resolution of the prediction algorithm; we cannot make predictions for the location of the target θ or equivalently for the offender's anchor point on distance scales significantly smaller than this resolution.

It is important to note that the question of the resolution of the estimate is independent of the question whether or nor the method is correct. In our foundational example, we know that the estimate for the target θ is correct, as it is a direct consequence of the hypotheses imposed as part of the example. The fact that the method is correct is separate from the spatial resolution of the method.

Though a prediction algorithm does have a spatial resolution, it is also important to note that this is not an absolute quantity, but rather depends on the method used to develop the geoprofile. Returning to our foundational example, let us suppose that that instead of using all of the elements in the data set $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ we use only the first $\mathbf{y} = \mathbf{y}_1$. Then (7) tells us that θ is distributed like a bivariate normal with

$$\theta|\mathbf{y}_1 \sim N\left(\theta \left| \frac{\mu\sigma^2 + \mathbf{y}_1\gamma^2}{\sigma^2+\gamma^2}, \frac{\sigma^2\gamma^2}{\sigma^2+\gamma^2}\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right.\right).$$

This model is also exact and accurate, even though it is different than the model (8). The key difference here is that this model relies on a different collection of data, $\mathbf{y}_1$ alone rather than $\bar{\mathbf{y}}$, so it is not surprising that the result is different. Note also that this has a different spatial resolution; by adding data and using $\bar{\mathbf{y}}$ instead of $\mathbf{y}_1$ as the basis for our estimate, we are able to decrease the size of the variance of the estimate for θ from

$$\frac{\sigma^2\gamma^2}{\sigma^2+\gamma^2}$$

to

$$\frac{(\sigma^2/n)\gamma^2}{(\sigma^2/n)+\gamma^2}$$

effectively shrinking the diameter of the search region for θ by roughly $1/\sqrt{n}$.

An important open question then to consider is how to improve the spatial resolution of a geographic profiling algorithm. Perhaps this can be done by considering additional data. In the case of serial crime, can adding information about crime patterns, offense timing, or other data allow us to reduce the size of a search area for a given collection of crimes? Clearly this is a question worthy of additional consideration.

The Effectiveness Problem. Related to the size of the search area is the *effectiveness problem* which is simply the question to what extent does the geoprofile actually aid in an investigation. A geoprofile may be completely accurate, but so large that the result is not useful in an investigation; consider for example the geoprofile for the offender who commits crimes across two or more counties. It is unlikely that a geoprofile would be of significant investigative value, as the underlying spatial resolution of any algorithm would be such that the resulting search are would be enormous. In contrast, if all of the crimes in a series are located in a single small neighborhood, then the geoprofile is also likely to be concentrated in that area; however that does not really provide the investigating officer something new that they did not already know either.

An important open question then, is the extent to which any geographic profiling algorithm provides insight into an investigation.

The Geography Problem. All of the open questions considered so far have an implicit dependence on the underlying geography. It may be the case that a geoprofiling method is accurate or effective in certain geographies and jurisdictions, but inaccurate or ineffective in others. For example, in this work all of the hypotheses were tested and validated against data provided by the Baltimore County police department. There may be geographic features particular to Baltimore county, like its unusual shape and relationship to Baltimore city that have skewed our results in one direction or another. An important open question then is to determine to what extent the conclusions and effectiveness of our tool observed here can be replicated in other geographies and jurisdictions.

The Computation Problem. The final open problem is the most narrow, and focuses on the techniques used in the software prototype to actually calculate the results. A number of assumptions and mathematical simplifications need to be made to actually calculate the quantities that appear in (3). Earlier versions of the prototype used slightly different assumptions; though they were significantly faster there were cases where the assumptions failed and caused the prototype to produce erroneous values. Though those errors have been corrected, the resulting algorithm is now much slower. Are there better ways to perform the calculation?

Other Topics. Internally, I have three different Master's students in mathematics working on research topics that have grown from this grant project. My student Jeremiah Tucker is continuing to look at models for offender behavior that explicitly account for non-independence of crime site locations; he is currently trying to develop an appropriate re-scaling for the distances when the crime sites follow either the near-repeat or the general model to see if they also follow an appropriate theoretical distribution.

A second student, Brian Bielski, is looking to better understand how to evaluate the accuracy of different profiling techniques. For example, though we have been able to show that the offender's actual anchor point was contained in the search area produced by the prototype for 74% of the non-residential burglaries and 70% of the residential burglaries, the search area produced is actually meant to be a full probability distribution for the offender's anchor point. This means that we should be able to generate statistical tests to evaluate the accuracy of the geoprofile and hopefully determine how well, if at all, the probability density produced by the prototype actually matches reality.

Finally, my third student, Brett Fitti-Hafer is looking to construct better models, not for the location of the offender's anchor point, but rather for the location of the next crime site. His current hypothesis is that the offender is likely to offend in regions where the offender is familiar. If an offender has committed crimes at two different locations, then the offender is then likely to not only know the area around the crime site locations, but is also familiar with the road network between the sites. The plan is then to use the different routes between the crime sites as a basis for inference about the next crime. Figure 88 shows a series of non-residential burglaries in Baltimore County. The crime site locations are indicated by pins, while potential routes between crime sites are indicated by lines, with red most likely, yellow less likely, and green less likely still. Notice that the next crime in the series, indicated by a house on the map, was not used to generate the routes yet lies on a route between crime sites.

The work of these students is continuing.

References

Block, R., & Bernasco, W. (2009). Finding a serial burglar's home using distance decay and conditional origin-destination patterns: A test of empirical Bayes journey-to-crime estimation in The Hague. *Journal of Investigative Psychology and Offender Profiling*, *6*, 187–211.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York, NY: Springer Verlag.

Canter, D., Coffey, T., Huntley, M., & Missen, C. (2000). Predicting serial killers' home base using a decision support system. *Journal of Quantitative Criminology*, *16*(4), 457–478.

Canter, D., & Hammond, L. (2006). A comparison of the efficacy of different decay functions in geographical profiling for a sample of US serial killers. *Journal of Investigative Psychology and Offender Profiling*, *3*, 91–103.

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

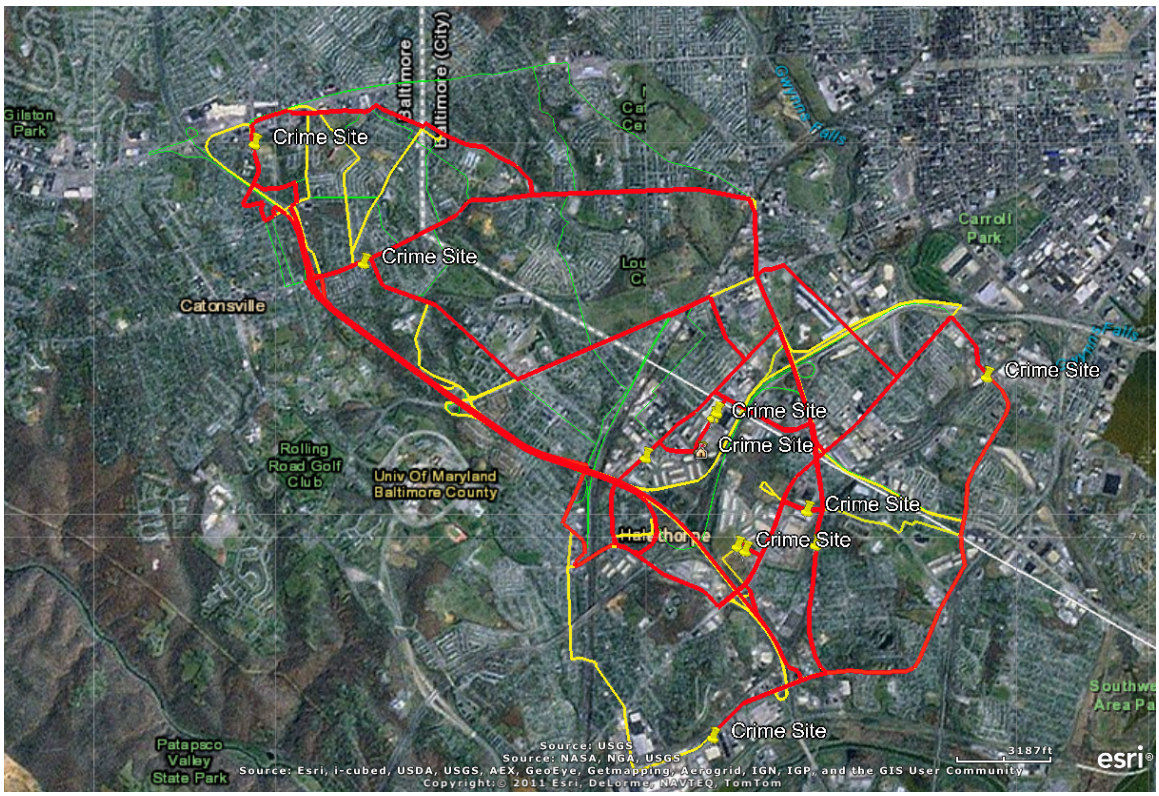


Figure 88. A series of non-residential burglaries in Baltimore County. Possible routes between crime sites are indicated, with red most likely, yellow less likely, and green less likely still. The next crime in the series, indicated by a house, was not used to generate the routes and yet lies on a route between crime sites.

Hald, A. (1952). *Statistical theory with engineering application*. New York: Wiley.

Hendricks, W. A., & Robey, K. W. (1936). Sampling distribution of the coefficient of variation. *The Annals of Mathematical Statistics*, *7*(3), 129–132.

Iglewicz, B., & Myers, R. H. (1970). Comparisons of approximations to the percentage points of the sample coefficient of variation. *Technometrics*, *12*(1), 166–169.

Kent, J. D., & Leitner, M. (2012). Incorporating land cover within Bayesian journey-to-crime estimation models. *International Journal of Psychological Studies*, *4*(2), 120–140.

Kocsis, R. N., Cooksey, R. W., Irwin, H. J., & Allen, G. (2002). A further assessment of "circle theory" for geographic psychological profiling. *The Australian and New Zealand Journal of Criminology*, *35*(1), 43–62.

Kocsis, R. N., & Irwin, H. J. (1997). An analysis of spatial patterns in serial rape, arson and burglary: The utility of the circle theory of environmental range for psychological profiling. *Psychiatry, Psychology, and the Law*, *4*(2), 195–206.

Koopmans, L. H., Owen, D. B., & Rosenblatt, J. I. (1964). Confidence intervals for the coefficient of variation for the normal and log normal distributions. *Biometrika*, *51*(1/2), 25–32.

Laukkanen, M., & Santtila, P. (2006). Predicting the residential location of a serial commercial robber. *Forensic Science International*, *157*(1), 71–82.

Le Comber, S. C., Nicholls, B., Rossmo, D. K., & Racey, P. A. (2006). Geographic profiling and animal foraging. *Journal of Theoretical Biology*, *240*, 233–240.

LeBeau, J. L. (1987). The methods and measures of centrography and the spatial dynamics of rape. *Journal of Quantitative Criminology*, *3*(2), 125–141.

Leitner, M., & Kent, J. (2009). Bayesian journey-to-crime modelling of single and multiple crime-type series in Baltimore County, MD. *Journal of Investigative Psychology and Offender Profiling*, *6*, 213–236.

Levine, N. (2010). *Crimestat: A spatial statistics program for the analysis of crime incident locations (v 3.3).* Ned Levine & Associates, Houston, TX and the National Institute of Justice, Washington, DC. Retrieved December 2010, from `http://www.icpsr.umich.edu/crimestat`

Levine, N., & Block, R. (2011). Bayesian journey to crime estimation: An improvement in geographic profiling methodology. *Professional Geographer*, *63*, 213–229.

Levine, N., & Lee, P. (2009). Bayesian journey-to-crime modelling of juvenile and adult offenders by gender in Manchester. *Journal of Investigative Psychology and Offender Profiling*, *6*, 237–251.

Linhart, H. (1965). Approximate confidence limits for the coefficient of variation of gamma distributions. *Biometrics*, *21*(3), 733–738.

McKay, A. T. (1932). Distribution of the coefficient of variation and the extended "t" distribution. *Journal of the Royal Statistical Society*, *95*(4), 695–698.

Meaney, R. (2004). Commuters and marauders: An examination of the spatial behavior of serial criminals. *Journal of Investigative Psychology and Offender Profiling*, *1*, 121–137.

Mohler, G. O., & Short, M. B. (2012). Geographic profiling from kinetic models of criminal behavior. *SIAM Journal on Applied Mathematics*, *72*(1), 163–180.

O'Leary, M. (2009a). The mathematics of geographic profiling. *Journal of Investigative Psychology and Offender Profiling*, *6*, 253–265.

O'Leary, M. (2009b). *A new mathematical approach to geographic profiling.* NIJ Final Report. Retrieved from `https://www.ncjrs.gov/pdffiles1/nij/grants/237985.pdf`

O'Leary, M. (2010). Implementing a Bayesian approach to criminal geographic profiling. In *COM.Geo '10: Proceedings of the 1st international conference and exhibition on computing for geospatial research & application.* New York, NY, USA: ACM. doi: http://doi.acm.org/10.1145/1823854.1823909

O'Leary, M. (2011). Modeling criminal distance decay. *Cityscape: A Journal of Policy Development and Research*, *13*(3), 161–198.

Paulsen, D. J. (2006). Connecting the dots: assessing the accuracy of geographic profiling software. *Policing: An International Journal of Police Strategies & Management*, *29*(2), 306–334.

Porter, M. D., & Reich, B. J. (2012). Evaluating temporally weighted kernel density methods for predicting the next event location in a series. *Annals of GIS*, *18*(3), 225–240.

Raine, N. E., Rossmo, D. K., & Le Comber, S. C. (2009). Geographic profiling applied to testing models of bumble-bee foraging. *Journal of the Royal Society Interface*, *6*, 307–319.

Rengert, G. F., Piquero, A. R., & Jones, P. R. (1999). Distance decay reexamined. *Criminology*, *37*(2), 427–445.

Rich, T., & Shively, M. (2004). *A methodology for evaluating geographic profiling software* (Tech. Rep.). Cambridge, MA: Abt Associates.

Rossmo, K. (1987). *Fugitive migration patterns*. Unpublished master's thesis, Simon Fraser University. Retrieved July 2012, from `http://summit.sfu.ca/item/5157`

Rossmo, K. (1995). *Geographic profiling : target patterns of serial murderers*. Unpublished doctoral dissertation, Simon Fraser University. Retrieved July 2012, from `http://summit.sfu.ca/item/6820?mode=simple`

Rossmo, K. (2000). *Geographic profiling*. Boca Raton, FL: CRC Press.

Sarangi, S., & Youngs, D. (2006). Spatial patterns of Indian serial burglars with relevance to geographical profiling. *Journal of Investigative Psychology and Offender Profiling*, *3*, 105–115.

Smith, W., Bond, J. W., & Townsley, M. (2009). Determining how journeys-to-crime vary: Measuring inter- and intraioffender crime trip distributions. In D. Weisburd, W. Bernasco, & G. J. Bruinsma (Eds.), *Putting crime in its place* (pp. 217–236). Springer Verlag.

Snook, B. (2004). Individual differences in distance travelled by serial burglars. *Journal of Investigative Psychology and Offender Profiling*, *1*, 53–66.

Snook, B., Zito, M., Bennell, C., & Taylor, P. J. (2005). On the complexity and accuracy of geographic profiling strategies. *Journal of Quantitative Criminology*, *21*(1), 1–26.

Townsley, M., & Sidebottom, A. (2010). All offenders are equal, but some are more equal than others: Variation in journeys to crime between offenders. *Criminology*, *48*(3), 897–917.

van Koppen, P. J., & de Keijser, J. W. (1997). Desisting distance-decay: On the aggregation of individual crime-trips. *Criminology*, *35*(3), 505–515.

Vangel, M. G. (1996). Confidence intervals for a normal coefficient of variation. *The American Statistician*, *50*(1), 21–26.

Warren, J., Reboussin, R., Hazelwood, R. R., Cummings, A., Gibbs, N., & Trumbetta, S. (1998). Crime scene and distance correlates of serial rape. *Journal of Quantitative Criminology*, *14*, 35-59.