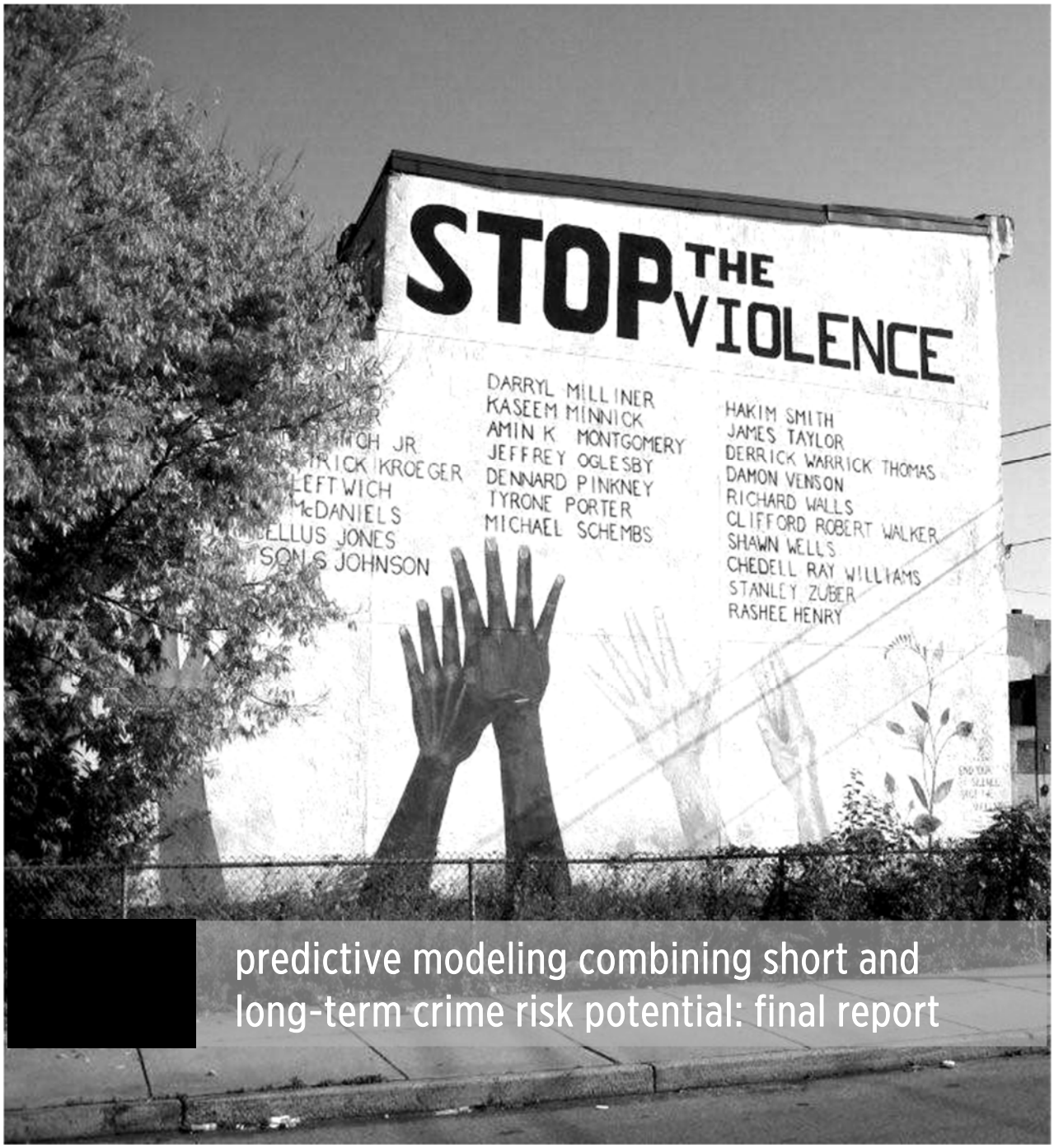
TEMPLE UNIVERSITY®

Center for Security and Crime Science




# predictive modeling combining short and long-term crime risk potential: final report

**Report Title**: Predictive modeling combining short and long-term crime risk potential: Final report
**Award number**: 2010-DE-BX-K004
**Authors**: Jerry H. Ratcliffe, PhD, Ralph B. Taylor, PhD, and Amber Perenzin, MA

**Contact**:
Dr Jerry Ratcliffe
Center for Security and Crime Science
Department of Criminal Justice
Temple University
5th floor, Gladfelter Hall
1115 Polett Walk
Philadelphia PA 19122
215-204 7918

www.cla.temple.edu/cj
jhr@temple.edu

**Notes**
This final report follows the Final Technical Report Guidelines laid out at http://www.nij.gov/funding/pages/final-technical-report-guidelines.aspx

Parts of this report are copied from existing publications written by the research team, and in particular the following publications:

- Taylor, RB, Ratcliffe, JH & Perenzin, A (2015) Can we predict long-term community crime problems? The estimation of ecological continuity to model risk heterogeneity, *Journal of Research in Crime and Delinquency*, 52(3): 635-657.
- Ratcliffe, JH (2016) *Intelligence-Led Policing*, Routledge: London. Second edition.





This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

## ACKNOWLEDGEMENTS

The authors are immensely grateful to Jeremy Heffner from Azavea for his technical assistance throughout the project, and to Jeremy, Bennet Huber and Kenny Shephard for their invaluable efforts creating the PROVE and ACS Alchemist software programs. We are also indebted to John Branigan and Robert Cheetham from Azavea for helping the project reach a successful completion. We would further like to thank Steve Schuetz from the National Institute of Justice for his help and careful stewardship of this project. Opinions or points of view expressed in this report those of the authors and do not necessarily reflect the official position or policies of the agencies, companies or individuals mentioned here.

# ABSTRACT

This research team (Temple University and industry partner Azavea) developed a technology capable of predicting future crime risk potential based on a number of grounded theoretical approaches to understanding localized spatial crime patterns. With regard to long-term crime risk changes, a stable crime niche model assumes that communities occupy crime niches in a broader jurisdiction, niches that are largely stable from year to year and have self-maintaining properties. Thus crime in one year may be predicted best by crime from the previous year. Alternatively, a structural model assumes that key current demographic conditions, such as socioeconomic status and racial composition, generally shape crime levels. Finally, a dynamic ecological and structural model assumes, net of the connections between current crime and demographic structure, that current structural conditions influence future long term changes in crime for a year in the future. The focus here is on ecological crime discontinuities, with priority assigned to demographic factors shaping such crime shifts over time. At the same time, ecological crime continuities also are present to a degree, linking current and future crime levels. These models were compared in the research study.

The research team also examined what role near-repeat crime events, indicative of a short-term change in relative risk, have in modifying this relationship. Near repeats occur when a crime influences the likelihood of another crime within a narrow space and time window after the originator event. In particular, the 'boost' hypothesis (also known as 'event dependency') suggests that subsequent events are conditional on the originator event because (for example) the same offender returns to the area, or there is a retaliatory event.

Using 2009 and 2010 reported crime for the City of Philadelphia, PA (USA) we identified that the demographics-plus-crime was the most parsimonious and accurate for robbery, burglary, aggravated assault, and vehicle theft when predicted from year-to-year in small geographic areas of 500 feet by 500 feet grid cells. Lower volume crime types (homicide and rape) were predicted as well as, or better, by the demographics-only model.

We then added an event-dependency risk surface to the long-term crime risk predictions and estimated what impact this near repeat surface played in changing the accuracy and parsimony of the crime prediction. The best combination of accuracy and model parsimony was estimated by comparing differences in Bayesian Information Criterion (BIC) values. Near repeat patterns were estimated for two week periods across spatial bands of 250 feet width. These near repeat patterns were translated to a mapped risk surface and added to the long-term risk prediction surfaces. In this part of the study, 2012 crime was used to predict 2013 crime in the City of Philadelphia for two of the most frequent types of part 1 crime: robbery and burglary.

With repeated examination of two-week predictions across 500 foot square grid cells, the strongest BIC value was identified with a model that combines crime from the previous year, change in demographic structure, and an adjustment for the near repeat phenomenon. Mixed effects logit models suggest that long-term (year-on-year) crime and demographic changes are more influential in this model than near repeats. Theoretically, this means that long term ecological crime continuities, long term crime *dis*continuities arising from stratification patterns in class and race, and near-term crime continuities in time and space all shape the two week, micro-scale predictions.

In summary, a model combining community structural characteristics, crime counts from the previous year, and an estimate of near repeat activity generated the best results overall. This tells us that small scale, short term crime occurrences reflect a complex mix of near-term crime continuities, ecological crime continuities, and ecological structure which generates ecological crime discontinuities forward in time. The industry partner, Azavea, has created a free software program (PROVE) to perform these calculations for state and municipal police departments.

# CONTENTS

# EXECUTIVE SUMMARY

## Synopsis of the problem and the research purpose

At the First Predictive Policing Symposium held in Los Angeles in 2009, Assistant Attorney General Laurie Robinson noted that "Law enforcement leaders are using predictive techniques in a variety of forms, but we don't necessarily have a handle on all of them". The problem stems in part from an enthusiasm for data-mining that is devoid of theoretical direction. The lack of theory informing current discussions was clear in the breakout sessions at the meeting attended by one of the principal investigators of this report, and in the initial five elements of predictive policing put forward; i*ntegrated information and operations, seeing the big picture, cutting-edge analysis and technology, linkage to performance,* and *adaptability to changing conditions.* The integration of well-tested theory was noticeably absent from this list. Scholars have previously commented on the gulf between crime data mining and criminological theory (Marshall and Townsley 2006). Theoretical development is thus vital to the development of crime prediction, and to policy considerations that follow.

Crime prediction in an emerging technology that combines data mining with (broadly) two contrasting theoretical approaches based on long and short term crime development. With regard to long-term crime risk changes – indicative of a generalized risk heterogeneity - a stable crime niche model assumes that communities occupy crime niches in a broader jurisdiction, niches that are largely stable from year to year and have self-maintaining properties. Thus crime in one year can be predicted best by crime from the previous year. Alternatively, a structural model assumes that key current demographic conditions, such as socioeconomic status and racial composition, generally shape crime levels. Finally, a dynamic ecological and structural model assumes, net of

the connections between current crime and demographic structure, that current structural conditions influence future long term changes in crime for a year in the future. The focus here is on ecological crime discontinuities, with priority assigned to demographic factors shaping such crime shifts over time. At the same time, ecological crime continuities also are present to a degree, linking current and future crime levels. Thus crime and demographic changes best predict crime.

In terms of short-term crime patterns, recent developments in near repeat crime have examined what role near-repeat crime events, indicative of a short-term change in relative risk, have in moderating the long-term crime dynamic? Near repeats occur when a crime influences the likelihood of another crime within a narrow space and time window after the originator event. In particular, the 'boost' hypothesis (also known as 'event dependency') suggests that subsequent events are conditional on the originator event because (for example) the same offender returns to the area, or there is a retaliatory event. The introduction of a short-term risk adjustment based on the near repeat hypothesis increases our understanding of the relative strength of crime predictors. Shane Johnson (2010: 361) has remarked in relation to spatial patterns of crime, "it would seem that neither the risk heterogeneity or boost hypothesis in isolation can explain the observed patterns; both have a part to play. Determining the precise contribution of each and if and how this varies over space and time would be a useful next step". This is one of the central questions that the research project sought to address.

Crime prediction is integral to *predictive policing*. It has been known for more than a decade that police officers do not necessarily know exactly where crime hot spots are located, and that their knowledge is better for some crime types than others (Ratcliffe and McCullagh 2001). There might therefore be a role for theory and software to improve this situation. Predictive policing, at

least in terms of street policing, is 'the use of historical data to create a spatiotemporal forecast of areas of criminality or crime hot spots that will be the basis for police resource allocation decisions with the expectation that having officers at the proposed place and time will deter or detect criminal activity' (Ratcliffe 2014: 4). While it can also be used for predicting offenders, it has attracted the most attention in regards to the geography of crime. Attendees at the National Institute of Justice's first predictive policing symposium in Los Angeles in 2009 identified numerous potential applications of predictive policing, but the primary use was to describe the time and location of future incidents in a crime pattern or series. An additional function of this research project has been to work with an industry partner, Azavea Inc., to develop and release a spatial technology capable of predicting future crime risk potential based on these various grounded theoretical approaches to understanding localized spatial crime patterns.

In summary, the purpose of this project has been to:

1. Resolve is what way fundamental demographic correlates of crime, proven important in community criminology, link to *next* year's crime levels, even after controlling for this year's crime levels? This is answered for small-scale, intro-urban communities.
2. Examine what role near repeat crime events, indicative of a short-term change in relative risk, play in moderating this relationship?
3. Develop – with our industry partner Azavea – a computer program that allows for crime predictions based on these theoretical approaches to be made for cities and jurisdictions across the United States.

## Research design and results

This project employed crime data and Census information for the City of Philadelphia, PA. Philadelphia is the 5th largest city in the country with a population of about 1.5 million people. In 2012 about 15% of Americans were living below the poverty line, but in Philadelphia 25% of residents were living below the poverty line (American Community Survey 2012).

There are two studies conducted within this research project. The first addresses the question of whether fundamental demographic correlates of crime, proven important in community criminology, link to *next* year's crime levels, even after controlling for this year's crime levels, at least in small scale, intra-urban communities? This is a question of *predicting long-term community crime problems*.

2009 crime frequencies for homicide, rape, robbery, aggravated assault, burglary, and vehicle theft were aggregated to over 1,600 census block groups across the city. This study used 2005-2009 five-year data release demographic variables collected through the American Community Survey (ACS). A Socio-Economic Status (SES) index included four variables: percent households reporting income less than $20,000 in 2009 (reversed); percent households reporting income greater than $50,000 in the same year; median house value (natural logged after adding 1, in 2009 dollars); and median household income (natural logged after adding 1, in 2009 dollars). Each variable was z-scored then averaged to create the SES index; higher scores indicate higher SES (Cronbach's $\alpha = .90$). A Race variable measured the percentage of residents in a neighborhood who identified themselves as white non-Hispanic indicated racial composition. Predictions were generated using negative binomial regression models with a spatially smoothed outcome variable.

For four outcomes – robbery, aggravated assault, burglary, and motor vehicle theft – a model that included demographic structure and earlier crime from the previous year provided by far the strongest combination of accuracy and parsimony. In all four of these cases, the Bayesian Information Criterion (BIC) value was at least 10 lower than the next closest model, providing "very strong" evidence that this was the preferred model for these outcomes. In other words, the demographics-plus-crime was the most parsimonious and accurate for robbery, burglary, aggravated assault, and vehicle theft when predicted from year-to-year in the small geographic areas of 500 feet by 500 feet grid cells. Lower volume crime types (homicide and rape) were predicted as well as, or better, by the demographics-only model.

We then added an event-dependency risk surface to the long-term crime risk predictions and estimated what impact this near repeat surface played in changing the accuracy and parsimony of the crime prediction. Parsimony was again estimated using a Bayesian Information Criterion (BIC). Near repeat patterns were estimated for two week periods across spatial bands of 250 feet width. These near repeat patterns were translated to a mapped risk surface and added to the long-term risk prediction surfaces. In this part of the study, 2012 crime was used to predict 2013 crime in the City of Philadelphia for two of the most frequent types of part 1 crime: robbery and burglary.

With repeated examination of two-week predictions across 500 foot square grid cells, the strongest BIC value was identified with a model that combines crime from the previous year, change in demographic structure, and an adjustment for the near repeat phenomenon. Mixed effects logit models suggest that long-term (year-on-year) crime and demographic changes are more influential in this model than near repeats.

In summary, short-term crime (for two weeks) predictions for micro-geographic areas (500 feet grid cells) based on variables derived from an analysis of near repeat patterns outperform crime counts from the preceding calendar year. However, these near repeat estimates do not perform as well as long-term predictions based on community structural variables, and in particular socioeconomic status and race. A model combining community structural characteristics, crime counts from the previous year, and an estimate of near repeat activity generated the best results overall. This tells us that small scale, short term crime occurrences reflect a complex mix of near-term crime continuities, ecological crime continuities, and ecological structure which generates ecological crime discontinuities forward in time. Subsequent analysis using constrained and unconstrained mixed effects logit models suggests that the risk heterogeneity variables (comprised of community structural measures) have more explanatory power for burglary and robbery prediction than short-term near repeat measures.

## Conclusions and implications

It is possible to predict crime for short periods of time (two weeks out) in micro-geographic areas of 500 feet by 500 feet using just crime reports and freely available census data. With this constraint, prediction is best achieved by combining crime from the previous year, changes in particular demographic variables, and an estimation of local near repeat patterns. Two practical considerations intentionally limited the scope of inquiry. Models relied only on data routinely and freely available to crime analysts in local police departments. Second, since general models applicable across a *range* of crime outcomes were of interest, only predictors that consistently worked as theoretically expected across those crimes were included. It did not appear that these limitations on structural variables resulted in more poorly mis-specified models. Comparisons of

observed relative frequencies to predicted relative frequencies showed no improvement when additional predictors (e.g., residential stability) were included in models.

For all crimes examined, save lower volume offenses such as homicide and rape, current work supported the mixing of ecological crime continuities and discontinuities. Crime-plus-demographics models generated the best combination of parsimony and accuracy as reflected in markedly lower BIC scores.

But there are discontinuities as well. After controlling for current crime, current demographic structure linked significantly in all six crime-plus-demographic models. Because current crime was already factored in, demographics were linking to emerging crime changes that were unpredictable and unrelated to current crime levels. It is in this sense that these demographic-crime shift links reflect ecological crime discontinuities. In strong support of the structural perspective broadly, and the basic systemic model of crime (Bursik and Grasmick 1993) and work on the racial spatial divide in particular (Peterson and Krivo 2010), these emerging discontinuities link to community SES and racial composition.

For urban, small-scale communities, crime *and* structural predictors together generate the best crime prediction model for four out of six serious crimes in the long-term (year to year). This performance suggests ecological crime continuities are operative over time while, at the same time, ecological crime discontinuities, linked to current structural conditions, also unfold over time. The use of structural predictor variables will enhance analysts' abilities to inform police executives about which areas in their jurisdiction are most likely to foster criminal activity in the medium to long-term future. Indirectly, the research also suggests some practical value to the yearly estimates from the American Community Survey. Fortunately in the US, demographic

data are freely available to all law enforcement agencies through the U.S. Census Bureau and they are accessible with the development of the Alchemist extraction tool – a valuable byproduct of this funded project.

SES and racial composition prove sturdy crime predictors over all six crime types examined in this research, as would be expected by structural criminologists. Serious work remains ahead identifying the processes maintaining these ecological crime continuities, and the processes that generate the unfolding ecological discontinuities.

It should be noted however, that there are ethical considerations with the use of a racial composition variable in the prediction of crime victimization. There is significant community concern regarding the potential for racial profiling during the delivery of policing services (Welsh 2007). As such, it is understandable that – irrespective of the reliability of the empirical evidence – some segments of the community would have concerns about the use of a race variable in the prediction of crime, especially if coded into a software program used by police departments. It should be stressed at this point that the processes described in this report say nothing about offender characteristics: the report is focused on locations of criminal victimization.

The industry partner, Azavea, has created a free software program (PROVE) to perform the analyses detailed in this report. The target audience for this software is state and municipal police departments. The software program is therefore designed to identify and predict locations where crime victimization occurs. It is an unfortunate reality that in larger urban areas, such as Philadelphia, minority community neighborhoods are over-represented as crime locations and our black and Hispanic citizens are over-represented as crime victims, especially for violent

street crime. Few would disapprove if these findings were used to prioritize the delivery of social services and other long-term community assistance services, and we would certainly promote the use of these long term findings in this manner. It is also however a reality that the software emerging from this research does have a short-term application that police departments are more suited to exploit than other government agencies, based on the ability to quickly reassign patrol officers. Given these ethical concerns, the software program does not exploit the race variable in its default condition. Rather, if the user wishes to improve the accuracy of the predictions by using the race variable within the software, then the user has to change a parameter setting in the program. The user manual explains how to do this.

Based on crime predictability, predictive policing is an emerging tactic relying in part on software predicting the likely locations of criminal events. Predictive policing has been defined as "the application of analytical techniques—particularly quantitative techniques—to identify likely targets for police intervention and prevent crime or solve past crimes by making statistical predictions" (Perry et al. 2013: xiii). At present the field lacks robust evidence to suggest the appropriate policing tactic in predicted areas. Future research would do well to answer the question of whether different varieties of theoretically informed but also operationally realistic police responses to crime predictions estimated by a predictive policing software program can reduce crime. It may be that the ability to predict crime in the short term is of little value if government is insufficiently flexible to capitalize on this ability. If we consider predictive policing strategies to be related to hot spots policing, then two recent observations from Weisburd and Telep are relevant: (1) there are numerous strategies that have not yet been rigorously tested, and (2) much more needs to be learned about the impact of new technology on policing effectiveness (Weisburd and Telep 2014).

# INTRODUCTION

## Statement of the problem

In November 2009, the First Predictive Policing Symposium was held in Los Angeles to discuss this emerging framework and its impact on the future of law enforcement. As Assistant Attorney General Laurie Robinson emphasized in her opening remarks; "Law enforcement leaders are using predictive techniques in a variety of forms, but we don't necessarily have a handle on all of them"[1]. The problem stems in part from an enthusiasm for data-mining that is devoid of theoretical direction, accompanied by a confidence that examining ever-increasing data sets will solve crime prediction problems. The lack of theory informing current discussions was clear in the breakout sessions at the meeting attended by one of the principal investigators of this report, and in the initial five elements of predictive policing put forward[2;] i*ntegrated information and operations, seeing the big picture, cutting-edge analysis and technology, linkage to performance,* and *adaptability to changing conditions.* The integration of well-tested theory was noticeably absent from this list. Scholars have previously commented on the gulf between crime data mining and criminological theory (Marshall and Townsley 2006).

Crime prediction is integral to *predictive policing*. It has been known for more than a decade that police officers do not necessarily know exactly where crime hot spots are located, and that their knowledge is better for some crime types than others (Ratcliffe and McCullagh 2001). There

[1] http://www.ojp.usdoj.gov/nij/topics/law-enforcement/predictive-policing/symposium/opening-robinson.htm

[2] Sourced from Technical Breakout Session report, facilitated by John Morgan, Ph.D., Office Director, Office of Science and Technology, National Institute of Justice.

might therefore be a role for software to improve this situation. Predictive policing, at least in terms of street policing, is 'the use of historical data to create a spatiotemporal forecast of areas of criminality or crime hot spots that will be the basis for police resource allocation decisions with the expectation that having officers at the proposed place and time will deter or detect criminal activity' (Ratcliffe 2014: 4). While it can also be used for predicting offenders, it has attracted the most attention in regards to the geography of crime. Attendees at the National Institute of Justice's first predictive policing symposium in Los Angeles in 2009 identified numerous potential applications of predictive policing, but the primary use was to describe the time and location of future incidents in a crime pattern or series.

It is so new, and the strategies associated with it are so poorly defined, that there have not been any robust evaluations of the effectiveness of predictive policing to reduce crime (Santos 2014, Perry et al. 2013). A predictive policing experiment conducted in 2012 by the Shreveport Police Department (SPD) in Louisiana (the *Predictive Intelligence Led Operational Targeting* program, or PILOT for short) did not find encouraging results. Predictions were distributed at roll-call and discussed at monthly meetings, and commanders moved resources in response to predicted burglary locations. While there were some modest short-term gains, these were not sustained and the evaluators concluding that there was no statistical evidence that the PILOT program reduced crime.

If predictive policing were to reduce crime, a number of pieces have to fall into place (Ratcliffe 2014). First, the predictive software algorithm must be able to predict crime better than existing methods of crime prediction available to the jurisdiction (such as crime analyst or street officer knowledge). Second, if the prediction algorithm is functioning optimally, then the police

command system must be able to identify and deploy an appropriate tactic to reduce crime (Ratcliffe 2016).

Mapping crime and estimating future crime locations based on existing crime locations has long been a staple of crime analysis (Chainey and Ratcliffe 2005). However, researchers from the UK first identified the possibility for improving crime predictions over long-term crime hot spots (Bowers, Johnson, and Pease 2004, Johnson et al. 2009), and that there is a near-repeat crime problem in many places. Repeat victimization tells us that if a person's house is burgled, then they are at increased risk of another burglary for the next few weeks; however, we also know that nearby homes and locations are also at risk. This *near-repeat phenomenon* has been identified for burglaries (Bowers and Johnson 2004, Townsley, Homel, and Chaseling 2003), gunpoint robberies (Haberman and Ratcliffe 2012), shootings (Ratcliffe and Rengert 2008), and even insurgent attacks on coalition forces in Iraq (Townsley, Johnson, and Ratcliffe 2008). Some of the possible explanations include offenders learning about an area's weaknesses, or retaliation for previous nearby incidents.

But which is more important in predicting crime? Long-term crime and demographic changes, or short-term fluctuations based on the near-repeat phenomenon? Criminologist Larry Sherman and his colleagues argue for the long-term, and that 'Predictive policing is premised on the already-falsified claim that hot spots are not stable … many police agencies map hot spots on the basis of far too short a time period, generally less than a year. It is unlikely that crime will either be concentrated or predictable with such short time periods' (Sherman et al. 2014: 108). For now, many commercial software programs combine long-term and short-term crime patterns in their predictions, though few include demographic changes.

Census variables are available from rolling multiyear estimates for community demographic variables for any part of the country. At the census tract (CT) and census block group levels (CBG), new five year estimates are released annually. Our approach has reduced these to a few broad indices and a handful of individual variables. These indices and the selection of individual variables are based on extensive work over several decades on the fundamental demographic structure of communities. These fundamental features of community fabric have been widely linked with community crime rates in a recent high quality meta-analysis (Pratt and Cullen 2005b). These features, when translated to the grid cell level and linked to crime counts at that level create a base long-term '*crime potential'* map surface. This is a theoretically driven estimate of long-term crime potential based on structural community factors; but how does it compare to short term changes in risk?

Relatively recent research (in the last decade) identifying a near repeat phenomenon has been used in this study to create an '*event dependency'* surface that adjusts the base crime potential map and update the overall projected 'risk' surface map with recent events that can suggest short-term changes in crime patterns. The event-dependency surface represents the increased immediate risk of crime due to situational factors associated with crime hotspots and recent flare-ups (or crime spikes) of criminal activity. The combination of these two theoretical approaches to crime emergence incorporates long-term setting conditions and adjusts for recent crime hotspots in a predictive tool that requires only freely available census data for long-term predictions, and recorded crime events from the local police department to update the event dependency map component.

A combined prediction of long and short term risk, derived from theoretically driven variables and distributed freely to the crime analysis community, has been the goal of this project. In an

information-led or intelligence-led policing (Ratcliffe, 2008) environment, proactive policing requires a predictive ability, the opportunity to get ahead of the criminal element. This project has the potential to enhance crime prediction and influence police deployment and preventative patrol strategies across the country. Moreover, it would answer the puzzling criminological conundrum of whether underlying structural variables based on population characteristics are the overwhelmingly dominant force for current crime patterns, or if recent events and crime spikes are strong enough to supersede long-term crime potential to create new and emerging patterns (i.e., (Berk and McDonald 2009) crime regimes). The success of this study to actually estimate the relative weight of both the long- and short-term is, we believe, a sizable step forward in understanding spatial crime patterns.

While we answer the question later in this report, consider the possibilities. If long-term factors are dominant, then this knowledge provides ammunition for police departments to push for (e.g.,) socio-economic or stability changes as a way to begin to tackle endemic and chronic crime problems. If it is found, however, that spatio-temporal local conditions can create crime spikes markedly altering the underlying crime pattern, then this can influence prevention and patrol strategies, deployments, and changes in policing style. The implications for policing are potentially significant.

This project is specifically designed with sufficient analytical clout to provide meaningful and policy-relevant answers without requiring data sets that are only available to a select few or which are expensive and time consuming to gather. Any benefits will therefore accrue to any law enforcement jurisdiction in the country that has geocoded crime data and access to the internet to download Census data.

## Literature citations and review

Although crime prediction is an oft-stated goal for law enforcement agencies in an intelligence or information-led environment (Ratcliffe 2016), it remains a considerable challenge (Quarmby 2009). Even in an era of advanced computer power and data-mining potential, information processing is still most effective when conducted with specific goals and analytical structures in mind, rather than throwing everything into the mix in the hope of striking an often-elusive operational target (Davenport 1997). Scholars have previously commented on the gulf between crime data mining techniques (similar to some predictive policing implementations) and criminological theory (Marshall and Townsley 2006), but a significant body of research demonstrates that crime is unevenly distributed among places and victims (Felson 1987, Sherman, Gartin, and Buerger 1989, Weisburd and Eck 2004), and a number of spatial theories of crime can explain long and short-term changes in crime risk for small, local areas (Chainey and Ratcliffe 2005). Criminological theories, when operationalized, can direct and focus computer processing power and lead to improved crime prediction and prevention[3]. Notwithstanding a wide-ranging literature on city-wide crime reduction studies, we will concentrate in this report on a significant operational policing need – localized, small area crime prediction using 500 foot by 500 foot grid cells.

As Perry and colleagues (2013) point out, the theoretical foundations that allow for crime prediction are ably supported by routine activity theory (Cohen and Felson 1979), rational choice

[3] By focusing on proven criminological theories, we avoid less developed or more theoretically questionable research such as the study that suggested that when a certain number of people began practicing Maharishi Mahesh Yogi's Transcendental Meditation, crime in the UK city of Liverpool declined as a result! (see Hatchard et al. 1996).

theory (Cornish and Clarke 1986) and crime pattern theory (Brantingham and Brantingham 1981-2) - collectively known as opportunity theories. Furthermore, a range of community-level and spatial theories of crime have highlighted the ways in which fundamental dimensions of community demographic structure might set in motion processes which facilitate criminal behavior, crimes or delinquent activities. These models include, *inter alia*, social disorganization (Bursik 1988; Taylor 2000), routine activity theory (Cohen and Felson 1979; Felson 1987) and crime pattern theory (Brantingham and Brantingham 1984, 1993).

One influential meta-analysis (Pratt and Cullen 2005b) has found that all three of the broad dimensions of community demographic structure – socioeconomic status, stability, and race – link strongly to crime, even though the exact reasons for each depend on the theory adopted. Works also have linked demographic structural *changes* with corresponding or later crime changes (Taylor and Covington 1988, Covington and Taylor 1989). Consequently, it is possible to generate models of community-scale crime counts and changes in crime counts relying substantially on the demographic structural components identified as relevant to community crime rates.

The next two sections outline the specific theoretical tenets on which to base an analytical predictive program that incorporates a risk-heterogeneity surface with a shorter-term event-dependency surface.

### *Theoretical foundation for a risk-heterogeneity (crime potential) surface*

If crime can be predicted in the long term (year-on-year) by demographic and community structure indicators, then a vital stage is creating a small number of general indices, and selecting

individual indicators capturing the most important demographic dimensions of small-scale community structure.

Although they have been variously interpreted depending on the theoretical proclivities of the researchers and the theoretical perspectives in vogue at the time, for close to a century sturdy community-level demographic correlates of crime and delinquency rates have been identified in different studies (Kornhauser 1978, Peterson and Krivo 2005, Shaw 1929, Pratt and Cullen 2005a, Taylor 2000). Most of these studies have focused on urban communities, but the correlates seem somewhat similar in suburban jurisdictions (Alba, Logan, and Bellair 1994, Murphy 2007, Sheley and Brewer 1995). Most typically, in decreasing order of importance, studies link lower socioeconomic status (SES), increasingly non-white or increasingly racially mixed populations, household structure, and instability with higher crime rates (Pratt and Cullen 2005b: Table 2).

Some background on these factors is warranted. Repeated community-level analyses from different decades in cities on different continents have confirmed the existence of three broad, relatively independent community-level demographic dimensions: SES, racial/ethnic composition, and stability (Berry 1965, Berry and Kasarda 1977, Janson 1980, Wyly 1999). Generally, these three components are remarkably robust (Hunter 1971), though please note that the use of factorial ecology to identify relevant broad structural variables is different from social ecology or social area analysis as a theoretical perspective on community structure and dynamics. No social ecology or social area analysis assumptions are adopted in the current work.

In communities and crime work in the U.S. these dimensions have been used along with the recent addition of indicators for emerging household structures such as single parent households

with children, or single parent households in poverty with children. These aspects of household structure may contribute to either crime or victimization rates independent of SES, race, and stability dimensions (Sampson and Lauritsen 1994). The emerging separate relevance of household structure to crime and victimization rates reflects in part changing family structures and connections of family structure to other community dimensions. For example, the stability dimension used to tap into variables such as the percentage of married couple households, and the percentage of households with young children. Earlier factorial ecology called it a stability/familism dimension. But after the middle of the 20th century, the familism variables seemed to pull away from the stability indicators as a function of changing incarceration rates (Sampson and Wooldredge 1987), declines in married couple households generally, and shifts in birth rates between white and non-white urban US households (Bursik 1984, Cherlin 1981, Hunter 1971, Taylor and Covington 1988).

Additional components worth mentioning but not previously explored include the following. First, racial segregation and economic inequality within communities by racial or ethnic group have generated interest as they relate to crime (Lee and Ousey 2005, Peterson and Krivo 2005); however meta-analyses have not established their contribution net of the factors already mentioned, and their calculation requires analytic sophistication beyond the goals of developing and delivering the simple technology described here. Second, land use patterns contribute to crime. For example, recent research by Stucky and Ottensmann (2009) examined the contribution of land use to five year violent crime counts in 1,000 foot by 1,000 foot grid cells in Indianapolis. Land use is not included here for two reasons, one theoretical and one practical. First, even sophisticated research like Stucky and Ottensman (2009) *rarely* finds *consistent* main effects of land use across a number of crimes (out of the many land use variables examined,

those researchers found only two variables, residential density [often a proxy for position in the city and theoretically difficult to understand, see (Verbrugge and Taylor 1980)], and major road length [perhaps also a proxy for city position or sector] that consistently and significantly linked in the same direction with all their crime outcomes). Second – and more importantly - the purpose of the current project has not been to develop a comprehensive understanding of small-scale crime counts or crime count changes. The purpose has been to bring together predictors from two well-established theoretical veins of crime and place research to construct a geospatial technology that is effective, cost effective, and easy to use and implement. The required gathering of land use data would in most municipalities not fit these project constraints. Put simply, too many jurisdictions and crime analysts do not have accurate or even available land use data at the level required to convert into a useful predictive tool.

Indicators of disorder and incivilities are missing as well. Simply put, studies of long term community level crime changes find such indicators do not strongly and consistently predict such changes (Taylor 2001).

Two key ideas from the human ecological framework are relevant here. Different communities in a broader ecosystem like a city or a metropolitan area are interdependent. Further, these communities serve different functional niches relative to one another. In effect, different communities play different roles for populations throughout the region. “Ecological organization pertains to the total fabric of dependences that exist within a population” (Hawley 1950: 179). These niches can be stable from year to year or even decade to decade under some conditions. For example, with regard to delinquency “Shaw and McKay concluded that the local community areas of a city maintained an ongoing, consistent role in the dynamics of the urban system” (Bursik 1986: 39). This is acceptable “if the ecological structure of an urban system is in a state

of equilibrium" (Bursik 1986: 41). Of course, research has shown that over a longer period, such as a decade, ecological structures, crime and delinquency, and perhaps the connections between structure and crime or delinquency can shift (Bursik 1986, Taylor and Covington 1988, Velez and Richardson 2012). But for the look-ahead period of interest here, one year, if a large urban system is not afflicted with a major natural or man-made event like a Hurricane Katrina or 9/11 attacks, "local community areas" *generally* should be expected to maintain "an ongoing consistent role" to some degree.

One set of roles concerns crimes taking place within those communities. "Illicit or criminal occupations," and perhaps the patterns of their targets, can be part of those differentiated ecological functions (Hawley 1950:217). This is perhaps most readily grasped for crime functions like open air drug markets (Johnson, Taylor, and Ratcliffe 2013), but may apply to other major property and personal crimes as well. Therefore, next year's community crime levels may be largely shaped by this year's levels. Weisburd et al. (2012) powerfully demonstrated this for many of the streetblock trajectories they followed in Seattle. If this is largely true, then the only long-term risk factor needed to reliably estimate next year's crime risk level is this year's crime level. Ecological continuity of community crime niches will dominate, assuming stability in the broader ecosystem.

Alternatively, the key premise of structural criminology is that "the meaning and explanation of crime is to be found in its structural foundations" (Hagan and Palloni 1986: 432). Further "structural relations organized along vertical, hierarchical lines of power are of greatest interest to criminologists ... Structural criminology is distinguished by its attention to power relations and by the priority it assigns them in addressing criminological issues" (Hagan and Palloni 1986: 432) . For community criminology in the United States, when considering communities at the

intra-urban scale, the two dimensions of community fabric most clearly reflecting "lines of power" are socioeconomic status (SES) and racial—or, depending on the region of the country, ethnic—composition. Although in many cities these two threads correlate negatively and substantially (Peterson and Krivo 2010: 58), creating a racial spatial divide, the two are conceptually distinct and associated with distinct covariates and impacts (Massey 1998).

If models with just structural conditions outperform models with only current crime, this would suggest two points. First, the structural setting conditions prove broadly applicable, shaping future crime more strongly than current crime. These setting conditions may better reflect current and future power differentials than does observed crime. (As an aside, what observed community crime rates reflect may be far more tangled than current scholarship has acknowledged (Taylor 2015: 25-68)). Second, numerous crime niches at the community level, i.e., relative crime levels, may be shifting over time and thus demonstrating ecological discontinuities. Such shifts may reflect responses to changing inter-community power relations. The latter may be connected to temporal instability in the broader urban system.

Relatively small community units are examined here. Since smaller ecological units have greater potential for sizable change in shorter time frames (Abbott 2001), the crime functions that communities serve relative to one another may shift substantially in short time frames.

In terms of long-term crime prediction, the third prediction possibility combines Hawley's consideration of ecological continuity with the structural idea that power differentials shape ecological discontinuities. This frame expects that next year's community crime levels represent a mix of ecological crime continuities *and* discontinuities. If this is the case, next year's levels would be best predicted by this year's crime levels, *and* structurally-driven crime discontinuities.

If a model controls for current crime, the only portion of future crime remaining in the outcome reflects crime shifts *unrelated* to current crime levels (Bohrnstedt 1969, Bursik and Webb 1982). This portion reflects temporal discontinuities in the crime niches occupied by communities. If any non-crime predictors have a significant effect on future crime it is these discontinuities that structural factors are forecasting. Thus, these ecological crime discontinuities are *emerging from* current structural conditions. These conditions link not only to current crime; they also have generative impacts, unfolding over time, on crime. The current crime/future crime link is building on ecological continuities of crime niches over time, and the current demographics/future crime link is building on structurally-driven, temporally lagged ecological discontinuities in those same crime niches. Current structural relations are shaping elements of next year's crime, elements not detectable given this year's crime levels. If this mix of ecological continuity and discontinuity is the perspective that applies to year-ahead crime level predictions at the community level, then both current crime and current community structural data are needed.

In sum, we have identified three broad dimensions and a small number of additional variables that can capture most of the demographic variation relevant to community crime or victimization rates: SES, racial/ethnic composition, stability, household/supervisory structure, and racial/ethnic heterogeneity. These dimensions, based on earlier work, are broadly applicable to crime and victim risk heterogeneity at the community and sub-community scales, and hopefully to changes in crime and victim risk heterogeneity. This small set of indicators is theoretically relevant and generally complete, as proven by the factorial ecology literature, and crime relevant as shown by recent meta-analysis (Pratt and Cullen 2005a).

Which model outperforms which other models carries important theoretical implications. If the crime functions, or *niches*, which communities hold relative to one another and as reflected in their crime levels, (a) remain largely static from one year to the next, i.e., are operating within a largely stable urban system; and (b) are functionally more important than ongoing structural setting conditions, the crime only model would offer the simplest, most accurate model for next year's crime levels. By contrast, if the crime niches are (a) largely static from one year to the next but (b) are trumped in empirical importance by current setting conditions reflecting power relations, then the demographic model will offer the simplest, most accurate model for next year's crime levels. Finally, if the crime-plus-demographics model offers the simplest, most accurate model for the coming year's crime, this suggests that (a) crime niches are changing substantially from year to year and in ways not entirely predictable from their current crime levels. Additional implications follow if this last model proves preferable. Specific implications will depend on specific findings. (i) Should current demographic conditions significantly shape future crime, *after controlling* for current crime, this means that these current structural features play a role in generating forthcoming ecological crime discontinuities. The forthcoming shifts represent discontinuities because they are *unrelated* to current crime levels, since the latter are controlled. Structural consequences continue to unfold over time in ways not predictable given current crime. (ii) Should current crime also significantly link to later crime after controlling for structure, it means that next year's crime levels reflect a mix of ecological crime continuities, captured with the link to current crime, as well as ecological crime discontinuities, captured with the link between current structure and future crime after controlling for current crime.

### *Theoretical foundation for an event-dependency surface*

We now turn to a consideration of short term crime risk factors.

A significant body of research demonstrates that crime is unevenly distributed among places and victims (Cohen and Felson 1979, Sherman, Gartin, and Buerger 1989, Weisburd et al. 2004, Weisburd and Eck 2004). Importantly, a number of spatial theories of crime can explain short-term changes in crime risk for small, local areas (Chainey and Ratcliffe 2005). Two particular examples of measurable short-term heightened risk are repeat and near-repeat victimization. The former research has clearly demonstrated that victims experience an elevated risk of re-victimization in the months that follow an initial crime (Farrell, Chenery, and Pease 1998, Pease 1998), a feature of criminality that has implications for crime prevention (Forrester, Chatterton, and Pease 1988, Laycock 2001) and is highly amenable to identification through geospatial technology (Ratcliffe and McCullagh 1998). Indeed as pointed out by Townsley et al (2003), the last twenty years or so have seen an explosion in our understanding of the phenomenon of repeat victimization. Particularly applicable to residential burglary, it has been recognized that once a property has been the target of an offense, the risk of a repeat offense at that location is elevated. In some cases, this elevated level of risk has been determined to be 12 times the expected risk for the first month following a burglary, and that "the chance of a repeat burglary over the period of one year was around four times the rate to be expected if the events were independent" (Polvi et al. 1991: 412). The repeat victimization problem has also been identified for non-residential premises and, in particular, sporting facilities and educational establishments have been identified as victimized locations with a rapid repeat victimization cycle (Bowers, Hirschfield, and Johnson 1998).

More recently, research has identified that risks cluster in space and time (Townsley, Homel, and Chaseling 2003, Bowers and Johnson 2004). Using epidemiological techniques, it was found that in the aftermath of a burglary, the risk to nearby houses was temporarily elevated. This increased

level of risk varied from place to place, but in every study the elevated risk showed distinct and measurable spatial and temporal constraints. In other words, there was an elevated risk to nearby locations for a few weeks and for a distance of usually a few hundred feet (Bowers and Johnson 2004, Bowers, Johnson, and Pease 2004, Johnson et al. 2007). Although this important spatio-temporal crime pattern was first discovered for burglary, it has been found to also exist for violent crime (Ratcliffe and Rengert 2008) and even insurgent attacks on coalition forces in Iraq (Townsley, Johnson, and Ratcliffe 2008). Further examination of the near-repeat phenomenon to other crime types is underway (see Johnson, Summers, and Pease 2009 for a theft from vehicle example), partly encouraged by an NIJ grant supporting software development to create a stand-alone tool to identify the near-repeat phenomenon in any spatio-temporal data set (see grant 2006-IJ-CX-K006, awarded to JH Ratcliffe, 2006-2008).

Near repeat victimization has a strong theoretical foundation within environmental criminology. Burglars return to the same location drawn by the likelihood of similar properties to ones they have successfully targeted, and simply because the area is within their awareness space; robbers repeatedly target familiar areas within their awareness space, drawn by the value of a knowledge of local streets should they need to evade police (Rengert and Wasilchick 1985, 2000); and drug markets are drawn to particular sites (such as drug treatment facilities and transit hubs) because the routine activities of the local people allow dealers to blend in and find customers (St. Jean 2007). Violent crime events can also be predictive of further events, potentially overriding the long-term crime potential of an area with the more immediate need within some offenders for retribution and retaliation (Ratcliffe and Rengert 2008).

Knowing that there is a repeat victimization risk for a targeted location for a number of weeks *and* for a finite distance from an initial event allows the creation of an event dependency surface

that can be continually updated at regular short-term intervals, on the order of every one or two weeks. Further, these short term changes in the risk surface based on event dependency dynamics can be used to adjust the long-term crime potential or risk heterogeneity map surface. This combination of underlying risk and localized correction for short-term flare-ups of crime has the potential to provide significant predictive capacity for policing. It bears emphasizing that these short-lived, localized periods of higher victimization risk are *not* hot spots (Sherman, Gartin, and Buerger 1989). Hot spots are generally thought to be small scale locations where crime counts, relative to the surround, are elevated for an extended period, on the order of a year or two. The short term shifts discussed here represent something more dynamic.

Each of these surfaces, long and short term risk heterogeneity, may be of interest on their own, for different crime analysis and policy assessment purposes.

### *Combining long- and short-term risk surfaces*

The current study built on the underlying idea that crime analysts want to know simultaneously about what is happening with *both* of these risk surfaces. But there is a practical challenge about how to combine a long-term risk heterogeneity map surface with a short-term event dependency surface. That challenge arises simply because, to these authors' knowledge, this has not been attempted in the past.

Deepening the challenge is the lack of clear theoretical guidance on how these two risk dynamics connect, or might be conditional one upon the other. Bowers and Johnson (2004) attempted to use modus operandi to identify if repeat victimization was the result of one of two competing hypotheses. The 'flag' hypothesis suggests that certain locations effectively advertise their longstanding vulnerability to crime, thus attracting any passing opportunistic criminal (Pease

1998). In other words, the underlying socio-demographic conditions of the area create risk heterogeneities and thus drive criminal events. Connections between incidents emerge largely from long-term setting conditions creating these risk heterogeneities, i.e., crime potential differences. Alternatively, the 'event-dependency' hypothesis (known in the UK as the 'boost' hypothesis), suggests that a subsequent event is conditional—in space and time—on the first. Therefore because of an initial event, the follow-up crime is dependent on and related to the first (Johnson, Summers, and Pease 2009) [Alternative terms for these two phenomena are risk heterogeneity and state dependence (Taylor 1998)]. Mechanisms that might cause this include an offender returning to a previous burglary location (Rengert and Wasilchick 1985), retaliation in the local area driving a spike in shootings (Ratcliffe and Rengert 2008), or disruptions to local networks or gang dynamics as a result of an earlier shooting. The flag hypothesis examines how a community or community section contrasts with a broader environment, while the boost hypothesis addresses how a more micro-scale location contrasts with a more immediate environment for a brief period. The flag model is for the longer term, and larger or smaller areas, whereas the boost thesis is for the shorter terms and spatially restricted areas.

These two hypotheses indicate the challenge of merging the two risk surfaces. If risk-heterogeneity, i.e., the underlying crime potential, is the dominant mechanism driving crime counts, then the risk-heterogeneity surface should be given the majority weight. By contrast, if event-dependency dynamics are dominant, the major weighting should be given to that surface. The proposed work will derive an empirical solution to this combination problem.

The only partial test of these two competing hypotheses is found in recent work of Johnson et al. (Johnson et al. 2009, Johnson, Summers, and Pease 2009); but their [2009a] analysis drew on the similarities of modus operandi variables only. This is simply not viable for a program that aims

to be available across the whole of the U.S. Many jurisdictions record modus operandi in different ways, and it is an approach that is probably useful only for burglary. Turning to land use data, the ProMap analysis of Johnson et al (2009) employed data sets such as road lengths, numbers of buildings and a calculation for the number of barriers existing in the built environment; not only did these add minimal predictive capacity, again these are data sets that are unavailable to many analysts. We therefore avoid such exotic data sets. We have preferred a different mechanism to compare these two hypotheses, but in doing so actually have a methodology that appears to be easier to implement and maintain. The construction and combination of the surfaces are described later.

## Statement of hypothesis or rationale for the research

In small scale, intra-urban communities, do fundamental demographic correlates of crime, proven important in community criminology, link to *next* year's crime levels, even after controlling for this year's crime levels? If they do, it would imply that shifting ecologies of crime apparent after a year are driven in part by dynamics emerging from structural differentials.

What role do near repeat crime events, indicative of a short-term change in relative risk, play in modifying this relationship? If near repeats are the dominant mechanism by which crime patterns emerge, this would have significant implications for police deployment.

Finally, the project has developed – through our industry partner Azavea – a computer program that allows for crime predictions based on both approaches to be made for cities and jurisdictions across the United States.

# METHODS

There are two studies conducted within this research project. The first addresses the question of whether fundamental demographic correlates of crime, proven important in community criminology, link to *next* year's crime levels, even after controlling for this year's crime levels, at least in small scale, intra-urban communities? This is a question of *predicting long-term community crime problems*.

The second question examines whether short-term near repeat events modify the impact of long-term crime predictors. Does adding a short-term event dependency measure affect crime prediction? This is a question of *comparing long-term and short-term crime predictions*. We start with the question of predicting long-term community crime problems.

## Predicting long-term community crime problems

This study uses 2009 crime frequencies, demographic data reflecting these locations during the period 2005-2009, or both, as predictors of 2010 crime counts. The study location is Philadelphia, PA. Philadelphia is the 5$^{th}$ largest city in the country with a population of about 1.5 million people. In 2012 about 15% of Americans were living below the poverty line, but in Philadelphia 25% of residents were living below the poverty line (American Community Survey 2012). From 2009 to 2010 there were modest increases in the number of reported rapes, aggravated assaults and motor vehicle thefts (MVT) while burglary, robbery and homicide incidents all decreased slightly (see Table 1).

*Table 1 Descriptive Statistics (N=1,771 Block Groups)*

| Crime | 2009 | Mean | S.D | Min | Max | Sum |
|---|---|---|---|---|---|---|
| | Homicide | .2 | .48 | 0 | 4 | 256 |
| | Rape | .45 | .80 | 0 | 6 | 804 |
| | Robbery | 4.76 | 5.07 | 0 | 62 | 8,433 |
| | Aggravated Assault | 4.63 | 4.65 | 0 | 41 | 8,201 |
| | Burglary | 5.83 | 5.25 | 0 | 110 | 10,320 |
| | MVT | 3.72 | 3.12 | 0 | 26 | 6,587 |
| | 2010 | Mean | S.D. | Min | Max | Sum |
| | Homicide | .2 | .49 | 0 | 4 | 346 |
| | Rape | .47 | .83 | 0 | 6 | 834 |
| | Robbery | 4.42 | 4.93 | 0 | 49 | 7,820 |
| | Aggravated Assault | 4.70 | 4.75 | 0 | 43 | 8,201 |
| | Burglary | 5.75 | 4.66 | 0 | 34 | 10,189 |
| | MVT | 5.83 | 5.25 | 0 | 24 | 6,607 |
| City block group community structure | | Mean | S.D. | Min | Max | |
| | SES Index | -.06 | .89 | -2.91 | 2.60 | |
| | Percent WNH | 35.13 | 36.03 | 0 | 100 | |
| | Population | 864.55 | 583.95 | 4 | 4,548 | |

Note: Census block group data for Philadelphia. The analysis excluded census block groups with no reported population, leaving a total of 1,771 census block groups in the analysis. Structural variables based on the 2005-2009 American Community Survey estimates. S.D. =standard deviation. WNH=white non-Hispanic.

## *Unit of analysis*

As Johnson et al. (2009) explain, it is important that the unit of analysis used in a study matches the social process under investigation. The spatial unit in this study is the census block group. Numerous studies on crime, drugs and reactions to crime using census block groups can be found in the literature (Gorman et al. 2001, Jennings et al. 2012, Harries 1995, McCord et al. 2007). It seems a reasonable approximation of a community although, of course, community exists at smaller and larger scales than this (Suttles 1972). In a developed city a census block group usually includes four contiguous census blocks, with each census block having four sides. 1,771 census block groups in the city of Philadelphia were included in the analysis. We excluded 45 census block groups in the city because these areas had no residential population, and therefore,

no demographic data associated with them. There were no crimes geocoded to these areas in 2010, therefore no crime data were eliminated from the analysis by this exclusion.

### *Structural data*

This study uses demographic variables collected through the American Community Survey (ACS). The ACS is administered every year by the U.S. Census Bureau. ACS data are published every year for counties with populations of 65,000 people or more, every three years for populations of 20,000 people or more and every five years at the census block group level. The data that are used in this paper are from the 2005-2009 five-year data release. These data were downloaded with the Alchemist tool (see page 82 for details of where to download this tool). Our operationalization of these variables is described in the following section.[4]

The Socio-Economic Status (SES) index included four variables: percent households reporting income less than $20,000 in 2009 (reversed); percent households reporting income greater than $50,000 in the same year; median house value (natural logged after adding 1, in 2009 dollars); and median household income (natural logged after adding 1, in 2009 dollars). Each variable was z-scored then averaged to create the SES index; higher scores indicate higher SES (Cronbach's $\alpha = .90$). Descriptive statistics appear in Table 1.

.

[4] Details on the construction of the demographic indices can be found in an online appendix located at http://www.rbtaylor.net/crime_continuity_online_appendix.pdf. It provides names of specific ACS variables used and how each variable was modified to construct each index.

In the current study, race is represented with a variable measuring the percentage of residents in a neighborhood who identified themselves as white non-Hispanic indicated racial composition. This variable ranged from 0% to 100%, with a mean value of 35% (median = 20%). See Table 1.

The population variable summed the number of males and the number of females. This variable was natural log transformed and entered as a predictor. This is a recommended approach for a generalized count model (Maddala 1983: 51, 53, King 1988: 857) and does *not* assume marginal impacts.

### *Geographically smoothed outcome counts*

Predictions were generated using negative binomial regression models with a spatially smoothed outcome variable. The spatially lagged outcome variable was generated using OpenGeoda (v. 1.0.1). Crime counts for each census block group were averaged with the 6 nearest neighbors to that block group. Alternate versions of the outcome, using 7, 8, or 9 neighbors also were created and analyzed. These alternate analyses showed similar results. Models were also run after rounding the outcome variable to a whole number. The pattern of significant differences across models was unchanged. Generating a spatially smoothed outcome variable also helped correct for potential geocoding imprecision in the dataset. The data used in this study may have slight inaccuracies for street segments that cross census block group boundaries because the specific location on one side of the street segment is approximated. The spatial smoothing reduced the number of census block groups that experienced no crime over the outcome period (calendar year 2010). Using a lagged outcome variable also helped to reduce potential modifiable areal unit problems (Openshaw 1984), a useful trait given that one goal of this analysis is to generate a model that accurately and simply predicts crime counts in a general area. Please note that after

spatial smoothing, the number and percent of census block groups with crime counts of zero in the outcome year were as follows: burglary 0 (0%); motor vehicle theft 0 (0%); aggravated assault 1 (.06%); robbery 4 (0.23%); rape 271 (15.3%); homicide 729 (41.2%). Comparisons with theoretical expectations showed these distributions matched a *non*-zero inflated count model.

### *Model sequence and identification*

Three different negative binomial models were generated for each crime type. Throughout, each model is used for all six crime types. Model 1 represents the crime-only model which uses prior crime counts to generate predicted counts for the following year. Model 2 includes the two consistently-linked, theoretically-most-central and empirically-most-important demographic predictors, the SES index and racial composition, and the population variable (natural logged). Model 3 contains both demographic variables, population, and 2009 crime counts.

When assessing forecast quality, no one statistic can determine which model performs the best. Whether the model is used to predict the weather, flu outbreaks, future sales or crime, multiple measures are needed to assess various aspects of the model performance (Ebert 2003). In the current study, models are assessed relative to one another based primarily on a measure that considers both model fit to the data and model simplicity. This is the Bayesian Information Criterion (BIC). Standard forecast indicators of accuracy and bias are reported as well.

With regard to goodness of fit and parsimony, the Bayesian Information Criterion (BIC) "has become quite popular for model selection in sociology" especially for generalized models (Raftery 1995: 112). BIC values take into account both model fit to the data and model parsimony. When comparing across models, the strength of the evidence is determined by the

difference of the BIC values: the model with a lower BIC value is preferred. If the absolute difference between the two BIC values is greater than 10, this is interpreted as "very strong" evidence that one model is preferred over another. Differences of 6-10 provide "strong" evidence to prefer one model over another, and differences of 2-6 provide "positive" evidence that one model is preferred. Differences less than 2 are interpreted as "weak" evidence for preferring one model (Raftery 1995: 138-141).

Model accuracy was measured with two statistics commonly applied in forecasting models (Pepper 2008). Mean Absolute Error (MAE) measures the magnitude of the error values without considering whether errors in prediction arise from over- or under-prediction. The absolute value of the error term is calculated for each census block group. These values are then averaged together across the dataset. The Root Mean Square Error (RMSE) is more sensitive to substantial prediction errors (Pepper 2008). Residual values (observed – predicted) for each census block group are squared. The squared residuals are averaged over the data set, and then the square root of that average is calculated to produce the RMSE.

Given that each crime outcome is modeled three times, for individual predictors a Bonferroni-adjusted alpha level of $p < .01$ is adopted.

## Comparing long-term and short-term crime predictions

This part of the study examined two specific research questions:

1. Is it necessary to incorporate both long-term and short term (event dependency) variables in crime prediction models?

2. Which variables (long-term or short-term) carry more predictive weight in explaining future crime patterns?

To answer these questions, independent variables for the long-term (risk heterogeneity) component were created as described above. An estimate of the strength of the event dependency in a region was estimated using a near-repeat analysis.

With regard to the long-term variables, one model (referred to as *flag 1*) was generated using demographic data reflecting the premise of structural criminology. This represents the 'enduring' risk associated with a criminogenic area (Tseloni and Pease 2003). This model assumes that demographic data, such as socioeconomic status and racial composition, can be used to model long-term crime risk. A second version of the variable (*flag 2*) was created which used both structural data and one full year of temporally-aggregated crime data. Modeling the (long term) risk heterogeneity using both structural and crime data will been shown in the results section to improve predictive accuracy and model fit; however, this conceptualization of long-term crime patterns is theoretically different from a purely structural model (as will be demonstrated in the results related to the question of predicting long-term community crime problems).

To compare the relative influence of each of these variables, the risk heterogeneity mapped crime surface and the two variations on the event dependency risk surface are used to predict violent crime and property crime in micro areas over two-week time periods in 2013. Two separate analyses will be used to answer the research questions in this section. To answer the first research question, the first analysis tests the degree to which the combination of prediction surfaces improve model fit compared to a baseline model. The purpose of this is to determine if it is necessary to use both long-term and short-term variables to explain crime patterns. We

compare models using the Bayesian Information Criterion (BIC) which penalizes the goodness of fit measure when you add more variables to the model.

We anticipate that including both the short and long term surfaces together will be preferred over models that include only the risk heterogeneity surface or short term event dependency surface. If this is the case, the second part of the analysis will examine which prediction surface is more important when it comes to explaining future crime.

### *Estimating the near repeat measure for event dependency*

The event dependency (boost) variable is intended to measure the short term boost in crime risk generated after a nearby crime incident. According to this theory, the risk of crime reduces over time if no new incidents occur. In essence near term, near in space and time, crime continuities are being generated.

To model crime risk generated by recent nearby crime, a near-repeat analysis was run using an R script generated by Azavea Inc. This nearrepeat.R script identified the rates of spatial and temporal crime risk decay following a crime event. Using 2012 crime data from January 1st through December 31st, near repeat patterns were calculated separately for burglary and robbery. The near-repeat pattern for both crime types appear in Table 2.

*Table 2 Near-repeat odds ratios for 2012 in Philadelphia, PA.*

| Burglary | | | | |
|---|---|---|---|---|
| | 0 to 7 days | 8 to 14 days | 15 to 21 days | 22 to 28 days |
| 0 to 250 feet | 1.66 | n.s. | n.s. | n.s. |
| 250 to 500 feet | 1.15 | 1.08 | n.s. | n.s. |
| 500 to 1000 feet | 1.06 | 1.05 | n.s. | 1.05 |
| 1000 to 1500 feet | 1.03 | 1.03 | 1.04 | 1.04 |

| Robbery | | | | |
|---|---|---|---|---|
| | 0 to 7 days | 8 to 14 days | 15 to 21 days | 22 to 28 days |
| 0 to 250 feet | n.s. | 1.14 | n.s. | n.s. |
| 250 to 500 feet | 1.10 | n.s. | 1.06 | n.s. |
| 500 to 1000 feet | 1.04 | n.s. | 1.03 | 1.06 |
| 1000 to 1500 feet | n.s. | n.s. | n.s. | n.s. |

Note: Results are odds ratio values reflecting the increased risk of crime. Only statistically significant results are shown ($p<.05$). n.s. = not significant

The values in Table 2 are odds ratios that reflect the odds of a future crime event following the initial event. Odds above 1 reflect heightened risk.

In 2012, there was an elevated risk of burglary 0-7 days and 8-14 days following the initial burglary incident. This risk decreased as distance from the initial incident increased (an exception is the 0-250 foot bandwidth in the 8-14 day window which was not significant).

The robbery data generated a different near-repeat pattern. Using these data, crime risk 0-7 days from the initial event was only significant in the 250-500ft and 500-100ft spatial bands. The highest risk of a repeat robbery was within 0-250 ft of the initial incident extending 8-14 days.

The odds ratio values generated by the near repeat analysis were statistically significant in some distance ranges up to 28 days after the initial crime incident. Thus, in 2012 burglary and robbery events generated some degree of elevated crime risk for up to 28 days later. Therefore, it is necessary to use 28 days of prior crime events to generate the boost variable.

The odds ratio values that were generated in the near repeat analysis using 2012 data were carried forward and used to calculate the predicted odds of a near repeat crime occurring in 2013. To account for the short-term fluctuations in crime risk, the crime event values were updated every 2 weeks using the period starting 28 days prior to the start of the 2 week time period being predicted.

Short term crime risk is calculated using 500ft by 500ft grid cells as the unit of analysis. Values for this variable are calculated at a finer unit of analysis to increase the precision of the results. Each 500 foot grid cell was divided into 9 smaller cells. These smaller cells are referred to as the fishnet cells. The fishnet cells were assigned an odds ratio value based on the cell's proximity in space and time to previous crime events based on the results from the near-repeat analysis. In many cases, these fishnet cells were assigned more than one odds ratio value because the cell was spatially and temporally proximate to more than one incident over the 28 day period. When this happened, overlapping odds ratio values were added together. Note that before summing the values, a value of 1 was subtracted from each odds ratio value. Once the values were summed, a value of 1 was added back in at the end. For example, if the overlapping odds ratio values were 1.13 and 1.17, the summed value would be 1.30, based on (1-1.13) + (1-1.17). For a visual example, see Figure 1.

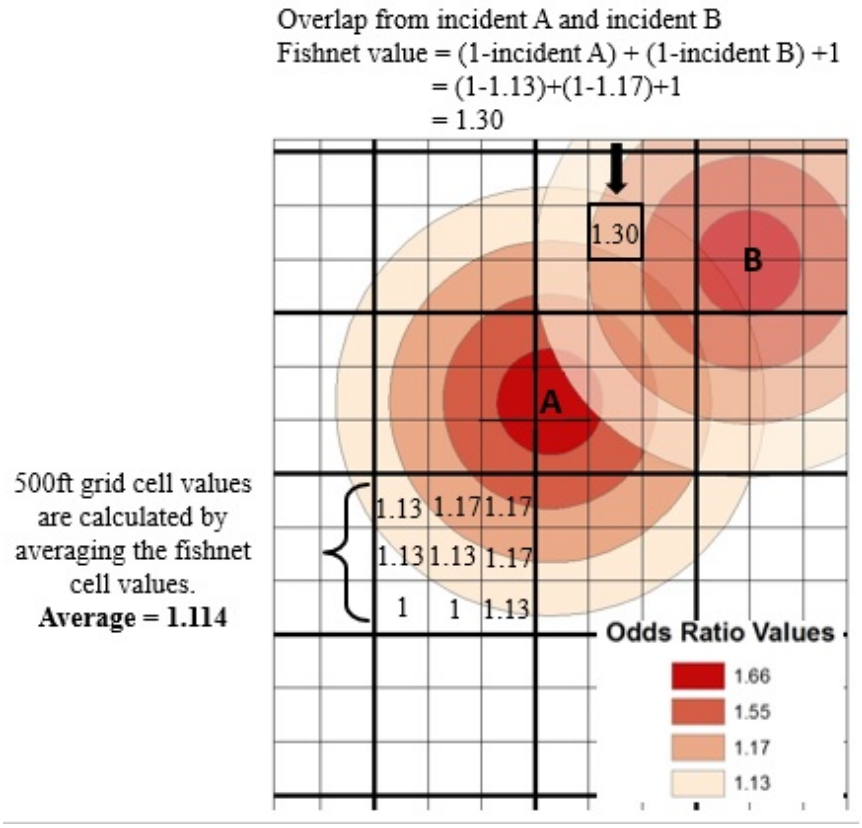



*Figure 1 Boost Calculation. "A" and "B" represent separate crime incidents in the 28 day period.*

Note: Heavy grid lines represent the 500 foot grid cells, while the thin grid lines represent the fine grained fishnet cells used to make the results more precise.

Fishnet cell values were based on the centroid of each fishnet. The 9 fishnet values within each 500ft grid cell were averaged together (see Figure 1 for an example). Again, these averaged odds ratio values were updated in each 14 day prediction window based on crime from the previous 28 days, thus reflecting changing crime risk over time.

### *Modeling the (long-term) risk heterogeneity*

To model the risk heterogeneity, two separate variables were created reflecting long-term crime potential. The first variable is generated using only community level demographic data published by the American Community Survey. The second uses both demographic data and prior annual

crime counts. These two versions are referred to as *flag 1* and *flag 2*, respectively. These variables were created using the R script longterm-build.R within the PROVE utility.

Negative binomial regression models generated each of the risk heterogeneity variables as follows. The outcome variable was the spatially smoothed crime count in 2012. This became the training data set for the modeling approach, in order to predict crime in 2013. Predictor variables included either demographic data (flag 1) or both demographic and crime data (flag 2). The demographic variables that were used included socio-economic status, race and residential population (The online appendix from Taylor, Ratcliffe & Perenzin (2015) is available as an appendix to this report and it contains a detailed explanation of how the SES index was constructed). Parameter estimates for these models are displayed in Table 3.

*Table 3: Training model risk heterogeneity results.*

| | Burglary | | Robbery | |
|---|---|---|---|---|
| | Flag 1 | Flag 2 | Flag 1 | Flag 2 |
| constant | .373 | 1.967 | 0.887 | 1.475 |
| SES $_{2006\text{-}2010}$ | -0.178* | -0.107* | -0.298* | -0.217* |
| Race $_{2006\text{-}2010}$ | 0.05 | 0.035 | -0.25* | -0.189* |
| population $_{2006\text{-}2010}$ | 0.104* | -0.015 | 0.125* | 0.003* |
| Crime count $_{2011}$ | | 0.026* | | 0.036* |

Note: N=1,400 census block groups. Results of GLM negative binomial regression at the census block group level. The dependent variable is the crime count in 2012, spatially smoothed with 6 nearest neighbors. *p<0.01

To generate the risk heterogeneity variables for predictions in 2013, the coefficients from the training model are rolled forward and applied to census and crime data that are available by the end of 2012 in order to be used. Census data published in 2012 reflect the 2007-2011 5 year estimates. By multiplying each 2012 variable by the coefficient from the training model, we generate predicted crime counts for 2013. See Equation 1 for an example using burglary (flag 2).

*Equation 1: Generating predicted crime counts*

$$Predicted\ Burglary\ Count_{2013} = 1.967 - 0.107(SES_{2007-2011}) + 0.035(Race_{2007-2011}) - 0.015(population_{2007-2011}) + 0.026(Burglary_{2012})$$

Predicted crime counts in 2013 are created by applying the regression output from the training model to the most recently available crime and census data. These predicted counts reflect the estimated number of crimes that are expected to occur over the course of the 2013 calendar year in each census block group. These annual counts were disaggregated to two-week time periods and to 500 foot grid cells to match the spatial and temporal scales of the boost variable.

The spatial disaggregation process involved dividing each 500 foot grid cell into 9 smaller fishnet cells, just as was done to create the boost variable. This improves the precision of the analysis and allows us to aggregate data when one 500 foot grid cell contains parts of more than one census block group. The spatial disaggregation process assumes that the annual predicted crime risk is homogeneous within each census block group across space and time.

The first part of the disaggregation process requires the calculation of the area of each fishnet cell. Since we are using 500 foot by 500 foot grid cells split into 9 smaller fishnet cells, each fishnet cell has an area of 27,778 square feet.

$$\frac{500\ ft\ gridcell}{3} = 166.67ft\ fishnet\ cell\ edge\ length$$

$$fishnet\ 166.67ft^2 = 27{,}778\ sq\ ft$$

Using the area of each census block group, we calculated how many fishnet cells it would take to perfectly cover each census area. A census block group that measures 2,000,000 square feet, for an example, would be completely covered using 72 fishnet cells.

$$\frac{census\ area}{fishnet\ area} = \frac{2{,}000{,}000 sq\ ft}{27{,}778\ sq\ ft} = 72\ fishnet\ cells\ to\ completely\ cover\ the\ cbg$$

Next, the annual predicted crime count (generated by rolling forward the regression output) is divided by the number of fishnet cells. If the long term crime prediction estimated a census block group would experience 20 crimes over the course of a year, this means there would be about .28 crimes per fishnet over the course of a year.

$$\frac{20\ crimes\ per\ year\ in\ cbg}{72\ cells} = 0.28\ crimes\ per\ fishnet\ per\ year$$

Calculating the number of crimes per fishnet allowed us to combine output from multiple census block groups into one 500 foot grid cell. These fishnet values are aggregated back up to the 500ft grid cells based on the location of the fishnet centroids. Cell values within a 500ft grid cell are summed. These summed values reflect the number of crime incidents that are predicted to occur inside that grid cell over the course of a year. An example of this process is depicted in Figure 2 and Figure 3. If all 9 fishnet cells in this example are contained within the same 500ft grid cell, the grid cell value would be 2.52 crimes per grid cell.

$$.28\ crimes\ per\ fishnet\ *\ 9\ fishnet\ cells = 2.52\ crimes\ per\ year\ per\ grid\ cell$$

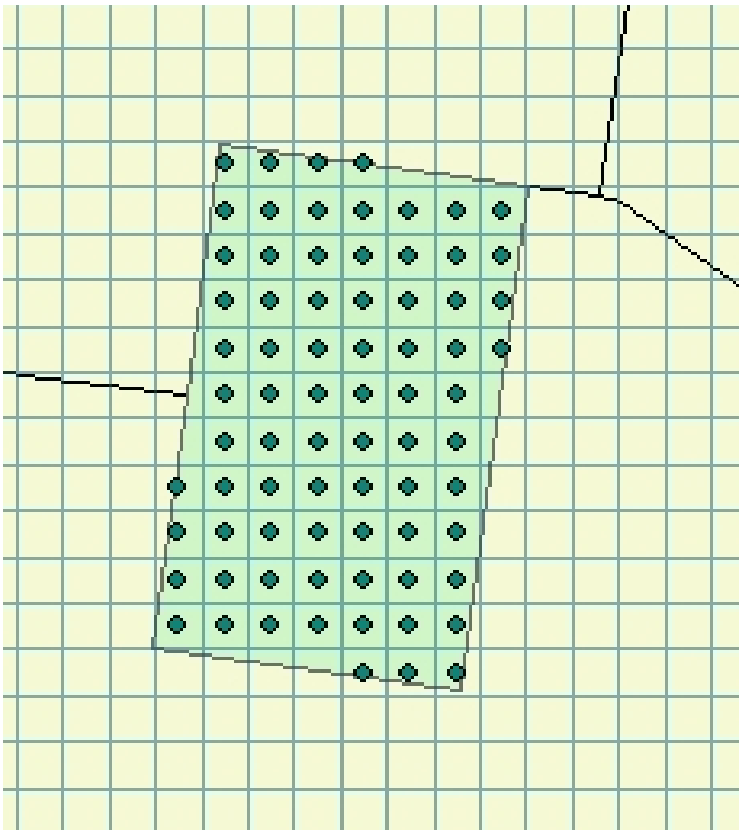


*Figure 2 Fishnet centroid locations.*

At this stage the risk heterogeneity variables are in a spatial unit that matches the boost variable. The temporal unit, however, still does not match. The flag variables reflect crime counts per year while the boost variable reflects two week time periods. To disaggregate the flag variable into bi-week time periods, each value is divided by 26 weeks. In this example, our model would estimate there will be 0.097 crimes in the grid cell in each 2 week time period in 2013; stated differently, one crime every 10 bi-weeks.

$$\frac{2.52\ crimes\ per\ year}{26\ biweeks} = 0.097\ cries\ per\ cell\ per\ biweek$$

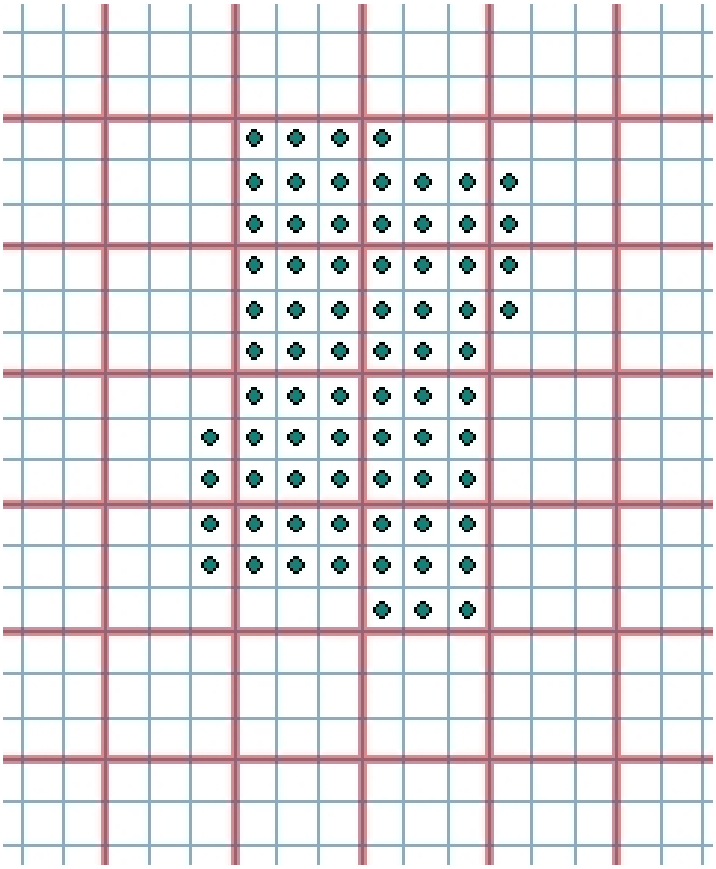


*Figure 3 Centroids with respect to 500 foot grid cells.*

It is important to reiterate that the risk heterogeneity variables do not update in each bi-week time period like the boost variables do. Instead, the same long-term crime prediction is applied to each bi-week throughout the course of the calendar year. This reflects the temporal stability of crime risk that is assumed under the risk heterogeneity hypothesis.

### *Descriptive statistics*

The long-term predicted counts and short-term risk values are used as independent variables in the analyses which follows. Descriptive statistics for the event dependency and risk heterogeneity variables are displayed in Table 4.

*Table 4 Descriptive statistics for event dependency (boost) and risk heterogeneity (flag 1 and flag 2) variables.*

| | N | min | max | mean | st. dev | skew |
|---|---|---|---|---|---|---|
| Burglary | | | | | | |
| Flag 1 | 440700 | 0.000 | 0.193 | 0.026 | 0.030 | 1.628 |
| Flag 2 | 440700 | 0.000 | 0.284 | 0.026 | 0.030 | 1.627 |
| Boost | 440700 | 1.000 | 11.084 | 1.051 | 0.094 | 12.133 |
| Robbery | | | | | | |
| Flag 1 | 440700 | 0.000 | 0.162 | 0.017 | 0.022 | 1.911 |
| Flag 2 | 440700 | 0.000 | 0.316 | 0.017 | 0.022 | 2.058 |
| Boost | 440700 | 1.000 | 1.392 | 1.011 | 0.026 | 3.183 |

Note: descriptive statistics are for 500 foot grid cell values for 26 bi-weeks in 2013. The total number of grid cells for the city of Philadelphia was N=16,950 (16,950 cells*26 bi-weeks = 440,700 observations).

Two separate analyses are run to answer the two primary research questions in this part of the study. In the first analysis, output from a series of seven mixed effects logit models are compared to test the first research question—do the event dependency and risk heterogeneity variables perform better or worse compared to a baseline model? The baseline model uses only previous crime counts to predict future crime. This method assumes that high crime locations at one time period will persist into later time periods due to the dominance of ecological crime continuity.

In the second analysis, we examine if the event dependency and risk heterogeneity variables are differentially relevant in explaining crime patterns; stated differently, if one variable is carrying more predictive weight compared to the other. We test this research question after standardizing the variables, then comparing output from constrained and unconstrained models.

# RESULTS

## Predicting long-term community crime problems results

Three prediction models were generated for each crime type resulting in a total of 18 regression models. The output generated by these models can be seen in Table 5.

*Table 5 Model comparisons by crime type (N=1,771 block groups)*

| | Homicide | | | Rape | | | Robbery | | |
|---|---|---|---|---|---|---|---|---|---|
| | Crime (model 1) | Demog (model 2) | Both (model 3) | Crime (model 1) | Demog (model 2) | Both (model 3) | Crime (model 1) | Demog (model 2) | Both (model 3) |
| 2009 Crime | .278** | | .113 | .189** | | .121** | .042** | | .034** |
| Population | | .186 | .162 | | .096 | .047 | | .141* | .052 |
| SES Index | | -.346** | -.341** | | -.234** | -.217** | | -.176** | -.136** |
| Percent WNH | | -.009** | -.009** | | -.006** | -.006** | | -.004** | -.003** |
| BIC | -11628 | -11687 | -11681 | -10522 | -10581 | -10583 | -5,289 | -5,289 | -5,481 |
| MAE | 0.323 | 0.310 | 0.310 | .615 | .600 | .592 | 2.782 | 3.178 | 2.777 |
| RMSE | 0.491 | 0.481 | .480 | .808 | .804 | .794 | 3.909 | 4.759 | 4.001 |

| | Aggravated Assault | | | Burglary | | | Motor Vehicle Theft | | |
|---|---|---|---|---|---|---|---|---|---|
| | Crime (model 1) | Demog (model 2) | Both (model 3) | Crime (model 1) | Demog (model 2) | Both (model 3) | Crime (model 1) | Demog (model 2) | Both (model 3) |
| 2009 Crime | .055** | | .032** | .020** | | .017** | .058** | | .055** |
| Population | | .196** | .042 | | .232** | .165** | | .198** | .045 |
| SES Index | | -.237** | -.176** | | -.095** | -.078** | | -.143** | -.116** |
| Percent WNH | | -.006** | -.005** | | .000147 | .0000659 | | .000659 | .0009501 |
| BIC | -5,137 | -5,417 | -5,556 | -5,206 | -5,134 | -5,239 | -6,429 | -6,238 | -6,451 |
| MAE | 2.822 | 3.034 | 2.770 | 3.239 | 3.315 | 3.163 | 2.076 | 2.288 | 2.060 |
| RMSE | 3.765 | 4.259 | 3.810 | 4.395 | 4.506 | 4.286 | 2.762 | 3.058 | 2.742 |

Note: Results of a negative binomial regression. Outcome is spatially smoothed 2010 crime (6 nearest neighbors). WNH=white non-Hispanic. BIC=Bayesian Information Criterion. MAE=mean absolute error. RMSE=root mean square error. ROC=receiver operating characteristic. Demog=Demographics model. *p < .01; **p < .001. Model with most negative BIC preferred.

### *Preferred models*

For four outcomes – robbery, aggravated assault, burglary, and motor vehicle theft – Model 3 with demographic structure and earlier crime provided by far the strongest combination of accuracy and parsimony. In all four of these cases, the BIC value was at least 10 lower than the next closest model, providing "very strong" evidence that this was the preferred model for these outcomes. Results proved different for homicide and rape. For rape, Model 3 (crime-plus-demographics) did not do appreciably better than Model 2 (demographics). For homicide, Model 2 (demographics) was preferred, generating a BIC value six units smaller than the next best model, Model 3 (crime-plus-demographics). This represents "positive" evidence that Model 2 was preferred.

These conflicting results about preferred model type, Model 2 or 3 for rape and Model 2 for homicide vs. Model 3 (crime-plus-demographics), may have arisen in part from the relatively infrequent nature of homicide and rape. In 2009 and 2010, there were less than 400 homicides and less than 850 rapes reported citywide for each calendar year. For other crime types, at least 6,000 incidents were reported yearly. Those lower yearly totals for homicide and rape, when disaggregated to the census block group level, may have affected the strength of the connection between 2009 and 2010 crime counts at this level.

### *Model accuracy differences*

Turning to the accuracy measures, Model 3 (crime-plus-demographics) models generated the lowest mean absolute error (MAE) for all crimes save homicide. For the latter, Model 3 and Model 2 (demographics) proved equally accurate. We gain a closer idea of what this means for model performance if we compare MAE to observed values, and fitted counts to observed counts, for a crime like robbery. An MAE for Model 3 (crime plus demographics) of 2.777

compares to a mean observed count of 4.39 suggesting on average predicted counts were off by about 63 percent. Although this is a sizable number, it should be borne in mind that these are crime counts for very small areas, and thus the average yearly count per area was also quite small.

If we turn to the absolute difference in relative frequencies, results were more encouraging. The mean absolute relative frequency differences in (observed minus predicted relative frequencies) was .0168 for counts of 0; for counts of 1, it was.0163; for counts of 2 it was .0183; and for counts of 3 or more it was .0148. Inspection of these same absolute difference patterns across counts for other outcomes suggested similarly sized differences in observed minus predicted relative frequencies.

### *Individual predictors*

Looking just at Model 3 (crime-plus-demographics), SES was always significant ($p < .01$) in the expected direction for all six crimes. This aligns with Pratt and Cullen's (2005b) conclusions about the primacy of SES for community crime.

Racial composition was significant in the expected direction ($p < .01$) for all four personal crimes, but not the two property crimes. This discrepancy for race aligns to some extent with Peterson & Krivo's (2010) finding of more complicated links between racial composition and property crime than for race and violent crime.

Earlier crime linked significantly ($p < .01$) to later crime for all crimes save homicide.

The relative impacts of crime, racial composition, and SES can be brought into closer focus by examining the impacts associated with standard deviation shifts in each of these predictors while

holding other predictors constant. [5] We use robbery as an example. Communities one standard deviation higher (sd=.88) on SES had an expected robbery count that was lower by a factor of .89, i.e., an expected robbery count 11 percent lower. Locales one standard deviation (sd=.36) higher on percent white non-Hispanic had expected robbery counts that were also lower by a factor of .89, i.e., 11 percent lower. So, for this crime, racial composition and socioeconomics proved comparably influential. Earlier crime, however, proved somewhat more potent. Places one standard deviation (sd=5.07) higher on robbery in 2009 had an expected robbery count a year later that was 19 percent higher.

## Comparing long-term and short-term crime predictions results

### *First analysis: model sequence*

Seven mixed effects logit models, with bi-weeks nested within grid cells were generated for each crime type. The focus shifted from trying to predict crime counts to simply predicting whether or not any crime of a particular type occurred inside each 500 foot grid cell during each of bi-week period in 2013. Given the short time frame and small scale cells, and more importantly the fact that the pertinent question has now switched to which model makes the best look-ahead forecast, the more pragmatic focus on whether a crime took place or not seemed warranted.

The model sequence merits mention. Not shown is the null or ANOVA mixed effects model, that only includes intercepts for each grid cell capturing the proportion, over all bi-weeks, of periods

[5] These can be generated by hand, or automatically using the *sppost* command *listcoef* (Long and Freese 2006).

when there was at least one crime event of the type in question. All of the models shown here doing significantly better, as witnessed by markedly lower BIC values, than the ANOVA model.

Model 1 reflects ecological crime continuity. If it does the best, compared to the other models, across all the bi-weeks, it is saying ecological crime continuity, as reflected in last year's crime counts, exerts the most powerful influence on crime occurrence in the future, at small spatial and temporal scales.

Model 2 reflects only *near-term* temporal and spatial crime continuity. It is based solely on what crime of the same type happened recently in or near the locale. If this model generates the best fit/parsimony combination then this is saying that future crime occurrences are shaped most strongly by what just happened here, in the recent past, in the location of interest, and nearby.

Model 3 blends longer-term ecological crime continuity and near-term crime continuity, allowing each to contribute to crime occurrences in the next two weeks. Model 1 says long term crime levels here in the locale of interest best predict upcoming crime. Model 2 says recent crime here in this cell and quite close by best predicts upcoming crime. Model 3 allows both of these dynamics to operate. In short, in all these models, if they do better than other models, the take away lesson is that past crime best predicts future crime; in the contribution ecological crime continuity may dominate (model 1), near term crime continuities may dominate (model 2), or the best crime model might combine the two.

Models 4-7 begin to introduce community structure. Model 4 uses just community demographic fabric (flag1). As was noted earlier when discussing the long term predictions, the idea is simply that impacts of locale, as reflected in residential composition, continue to produce crimes in short time-and-space windows in the future. If this model does the best, it is saying that what types of

people and households are living there drive crime occurrences within that context even in restricted spatial and temporal crime windows.

Model 5 leaves structure in but also introduces near-term crime continuities with the boost variable. If it does the best it is telling us that the crime-generating features of community structure, and recent crime nearby, present a more powerful prediction model than either of these alone.

Model 6 ignores near-term crime continuities, but combines long term impacts of community structure and crime. If this model does best it is telling us that short term, small scale, crime occurrences reflect both longer term community structure generating ecological crime *dis*continuities and longer-term ecological crime continuities.

Model 7 then brings in the near-term crime continuities, i.e., the near repeat dynamics, ignored in the preceding model. If this model does the best it is telling us that small scale, short term crime occurrences reflect a complex mix of near-term crime continuities, ecological crime continuities, and ecological structure which generates ecological crime *dis*continuities forward in time.

The baseline model, model 1, uses only 2012 crime counts from the previous two week time period as a predictor variable. Each subsequent model builds on this baseline model by using some combination of boost and flag variables. Each of these models are summarized in

Table 6.

*Table 6 Model sequence for mixed effects logit models predicting 2013 crime.*

| | model 1 | model 2 | model 3 | model 4 | model 5 | model 6 | model 7 |
|---|---|---|---|---|---|---|---|
| 2012 crime counts | X | | X | | | | |
| Boost | | X | X | | X | | X |
| Flag 1 | | | | X | X | | |
| Flag 2 | | | | | | X | X |

We compare these models using the Bayesian Information Criterion (BIC) and pseudo $R^2$ values. The latter values were calculated by comparing the variance of the predicted probability outcome with the variance of the binary outcome variable (none vs. presence of one or more crimes during the 2 week time period). This is in essence the Efron version of pseudo $R^2$.

Regression models using the boost variable would not converge. To overcome this issue, boost values were Winsorized at the highest 6 values. These 6 values were given the value of the $7^{th}$ highest value (value = 5.8). After the boost variable was Winsorized, the models converged.

Model comparisons can be found in Table 7. The lowest BIC value reflects the best combination of model accuracy and simplicity. Going from one model to the next, a lowering of the BIC value by more than 10 indicates "very strong" (Long 1997: 112) improvement in model fit while controlling for model complexity. Pseudo $R^2$ values are small across all models, but it is important to remember these models are predicting crime in micro areas (500ft grid cells) over micro time periods (2 weeks) using only two variables (boost and flag). Improvement in pseudo $R^2$ values is calculated with the percent change from one model to the next.

These results indicate that all models perform better than the baseline model. (The baseline model is not shown. It has no predictors, just intercepts for each grid cell. In mixed models these are termed the null or ANOVA models.) The best improvement in model fit and pseudo $R^2$ is achieved in Model 4 when the long-term (flag 1: structure only) variable is used as a predictor variable. However, model fit continues to improve throughout the model sequence. The best model in the sequence is model 7 which includes both boost and flag variables (flag 2: crime and structure). Model 7 is the best model for violent crime as well as property crime.

*Table 7 Comparing model fit across seven logit models.*

| | Burglary | | | | Robbery | | | |
|---|---|---|---|---|---|---|---|---|
| model | BIC | BIC change | Pseudo $R^2$ | $R^2$ change | BIC | BIC change | Pseudo $R^2$ | $R^2$ change |
| 1 | 84,900 | n.a. | 0.0307 | n.a. | 65,908 | n.a. | 0.0404 | n.a. |
| 2 | 84,208 | 692 | 0.0328 | 6.40% | 65,406 | 502 | 0.0410 | 1.50% |
| 3 | 84,106 | 102 | 0.0335 | 2.10% | 65,382 | 24 | 0.0413 | 0.70% |
| 4 | 80,279 | 3,827 | 0.0383 | 12.50% | 62,040 | 3,342 | 0.0456 | 9.40% |
| 5 | 80,060 | 219 | 0.0397 | 3.50% | 61,929 | 111 | 0.0462 | 1.30% |
| 6 | 79,900 | 160 | 0.0398 | 0.30% | 61,311 | 618 | 0.0507 | 8.90% |
| 7 | 79,732 | 168 | 0.0410 | 2.90% | 61,279 | 32 | 0.0510 | 0.60% |

Note: N=440,700 (16,950 grid cells x 26 bi-weeks). BIC=Bayesian Information Criterion. Pseudo R2 values shown. Pseudo R2=variance of predicted probability/variance of outcome. This follows Efron's Pseudo R2 formulation. 6 highest boost values for burglary are Winsorized.

Model definitions:

1. 2012 bi-week crime counts
2. Boost
3. Boost + 2012 bi-week crime counts
4. Flag 1
5. Flag 1 + Boost
6. Flag 2
7. Flag 2 + Boost

## *Second analysis: relative weighting*

Having determined that both event dependency and risk heterogeneity variables are necessary to achieve optimal model fit, the second research question becomes relevant: which variable is

more important? The second part of the analysis investigates the relative contribution of the event dependency and risk heterogeneity variables.

Mixed effects logit models were used to test this research question as well, but this time the event dependency and risk heterogeneity variables were log transformed then standardized before being entered into the models. This reduced potential leverage problems due to extreme predictor values. Most importantly, since the variables were standardized with the same standard deviation, an equivalent b weight indicates a comparable impact. Testing the comparable impact of the boost and flag variables is the question here. Descriptive statistics for the log transformed and standardized variables are in Table 8.

*Table 8: Descriptive statistics for standardized and log transformed variables*

| | | N | Min | max | mean | st. dev | skew |
|---|---|---|---|---|---|---|---|
| Burglary | | | | | | | |
| | Flag1 | 440,700 | -0.873 | 5.256 | -5.16E-13 | 1 | 1.550 |
| | Flag2 | 440,700 | -0.896 | 7.956 | -1.38E-12 | 1 | 1.533 |
| | Boost* | 440,700 | -0.591 | 29.50# | -4.78E-11 | 1 | 3.573 |
| Robbery | | | | | | | |
| | Flag1 | 440,700 | -0.791 | 6.261 | 4.638E-13 | 1 | 1.839 |
| | Flag2 | 440,700 | -0.788 | 12.063 | 1.938E-13 | 1 | 1.926 |
| | Boost | 440,700 | -0.420 | 13.580 | -1.52E-10 | 1 | 3.098 |

Note: Boost variable for burglary is winsorized at the highest 6 values. After winsorizing, the variable was log transformed and standardized. #=see text that follows.

Please note that in Table 8, one value is marked with a hash mark (#). This boost value of 29.50 seemed exceptionally high, but after reviewing the raw crime data we found that 3 storage facilities in Philadelphia experienced multiple break-ins on a single night. A separate incident report was generated for each unit that was broken into. Leverage, Cook's Distance and residual values were generated for each of the grid cells that were affected by these repeat break-ins. None of these statistics flagged these high boost values as being statistically problematic.

Leverage values were all less than .002, Cook's Distance values were all less than .004 and residual values were all near -1. The analysis was re-run without these repeat offenses in the dataset. Doing so did not change any of the results.

Each of the mixed effects logit models were run two different ways. First, we placed a constraint on the model which forced the coefficients of the standardized event dependency and risk heterogeneity variables to be equal. We then ran the models again without the constraint. The constrained and unconstrained models were compared using a Wald test. This does not directly test if the coefficients were significantly different from one another. Rather it looks at the entire model results and reports if results are significantly worse should the two coefficients be forced to share the same value. In Stata both *test* and *lrtest* can be used to compare models. While the *test* command can be used after any estimation command, the *lrtest* is more appropriate when maximum-likelihood estimation is used. Both tests are used here. Results indicate that the constrained model where the flag and boost variables were constrained to have equivalent impacts performed significantly more poorly. This suggests that the best model is one where the two coefficients can be different, one variable can have a greater impact than another. For both crime types, the flag variables have larger coefficients than the boost variables. This suggests the risk heterogeneity variables have more explanatory power compared to the event dependency variable (Tables 9 and 10).

*Table 9: Comparing constrained and unconstrained models (burglary).*

| | Burglary | | | |
|---|---|---|---|---|
| | Flag 1 | | Flag 2 | |
| | unconstrained | constrained | unconstrained | constrained |
| Constant | -4.714 | -4.407 | -4.717 | -4.398 |
| | (0.009) | (0.012) | (0.009) | (0.012) |
| Flag 1 | 0.837 | 0.355 | | |
| | (2.309) | (1.426) | | |
| Flag 2 | | | 0.869 | 0.357 |
| | | | (2.385) | (1.429) |
| Boost | 0.136 | 0.355 | 0.12 | 0.357 |
| | (1.146) | (1.426) | (1.127) | (1.429) |
| BIC | 79,923 | 81,534 | 79,591 | 81,417 |
| $R^2$ | 0.041 | 0.043 | 0.041 | 0.044 |
| test $\chi^2$ | 1352 | | 1443 | |
| | p<.001 | | p<.001 | |
| lrtest $\chi^2$ | 1625 | | 1839 | |
| | p<.001 | | p<.001 | |

Note: Results from mixed effects logit models. All variables are log transformed and standardized. Exponentiated coefficients are in parenthesis.

*Table 10 Comparing constrained and unconstrained models (robbery).*

| | Robbery | | | |
|---|---|---|---|---|
| | Flag 1 | | Flag 2 | |
| | unconstrained | constrained | unconstrained | constrained |
| Constant | -5.329 | -4.893 | -5.309 | -4.867 |
| | (0.005) | (0.007) | (0.005) | (0.008) |
| Flag 1 | 0.925 | 0.358 | | |
| | (2.522) | (1.430) | | |
| Flag 2 | | | 0.988 | 0.36 |
| | | | (2.686) | (1.433) |
| boost | 0.103 | 0.358 | 0.061 | 0.36 |
| | (1.108) | (1.430) | (1.063) | (1.433) |
| BIC | 61,855 | 63,427 | 61,130 | 63,202 |
| R2 | 0.046 | 0.047 | 0.051 | 0.049 |
| test $\chi^2$ | 1185 | | 1288 | |
| | p<.001 | | p<.001 | |
| lrtest $\chi^2$ | 1584 | | 2041 | |
| | p<.001 | | p<.001 | |

Note: Results from mixed effects logit models. All variables are log transformed and standardized. Exponentiated coefficients are in parenthesis.

In summary, short-term crime (for two weeks) predictions for micro-geographic areas (500 feet grid cells) based on variables derived from an analysis of near repeat patterns outperform crime counts from the preceding calendar year. However, these near repeat estimates do not perform as well as long-term predictions based on community structural variables, and in particular socioeconomic status and race. A model combining community structural characteristics, crime counts from the previous year, and an estimate of near repeat activity generated the best results overall. This tells us that small scale, short term crime occurrences reflect a complex mix of near-term crime continuities, ecological crime continuities, and ecological structure which generates ecological crime discontinuities forward in time. Subsequent analysis using constrained and unconstrained mixed effects logit models suggests that the risk heterogeneity variables (comprised of community structural measures) have more explanatory power for burglary and robbery prediction than short-term near repeat measures.

# CONCLUSIONS

## Discussion of findings

### *Predicting long-term community crime problems*

This part of our research compared the relative abilities of three theoretically-grounded, risk heterogeneity models to predict one-year, look-ahead future crime counts at the community level. Two practical considerations intentionally limited the scope of inquiry. Models relied only on data routinely and freely available to crime analysts in local police departments. Second, since general models applicable across a *range* of crime outcomes were of interest, only predictors that consistently worked as theoretically expected across those crimes were included. It did not appear that these limitations on structural variables resulted in more poorly mis-specified models. Comparisons of observed relative frequencies to predicted relative frequencies showed no improvement when additional predictors (e.g., residential stability) were included in models.

The three different model types examined here made different theoretical assumptions about community crime levels in the broader urban system. The crime only model can be derived from an ecological perspective. Crime levels reflect ecological niches (Hawley 1950), functional roles served by communities relative to other communities in the ecological system. If the broader urban system is in a relatively stable state from one year to the next, communities will not shift crime roles relative to one another, ecological crime continuity will predominate, and this year's crime should do the best job of predicting next year's crime. The demographics only model can be derived from structural criminology and the focus on power relations (Hagan and Palloni 1986, Logan and Molotch 1987, Logan 1978). Communities are constantly in conflict with one

another, and thus are continually sorting and re-sorting. Power differentials arise in part from different structural conditions at the community level, most notably SES and racial composition. These differentials shape future crime levels, especially if the broader urban system is in flux and community crime niches are shifting. Finally, a demographics-plus-crime model suggests that future crime levels in part reflect ongoing ecological crime continuities, leading to current crime significantly shaping future crime, and in part reflect ecological crime discontinuities over time, crime shifts unrelated to current crime but related to current ecological power differentials.

For long term models of all crimes, save homicide and rape, current work supported the mixing of ecological crime continuities and discontinuities. Crime-plus-demographics (Model 3) generated the best combination of parsimony and accuracy as reflected in markedly lower BIC scores. To some extent future community crime levels represent a continuation of current crime levels; current crime connected significantly to future crime in all versions of Model 3 save the model for homicide. Crime levels from one year to the next reflect significant ecological continuity.

But there are ecological discontinuities as well. After controlling for current crime, current demographic structure linked significantly in all six crime-plus-demographic models. Because current crime was already factored in, demographics were linking to emerging crime changes that were unpredictable and unrelated to current crime levels. It is in this sense that these demographic-crime shift links reflect ecological crime discontinuities. In strong support of the structural perspective broadly, and the basic systemic model of crime (Bursik and Grasmick 1993) and work on the racial spatial divide in particular (Peterson and Krivo 2010), these emerging discontinuities link to community SES and racial composition.

The relatively poor performance of the models that only used demographic data (Model 2) would indicate that future crime counts cannot be predicted with structurally-driven factors alone. While these models were generally accurate in generating 2010 crime predictions, they consistently underperformed relative to crime only (Model 1) and crime-plus-demographics models (Model 3).

Particularly small counts seem to shift the picture. The demographics-only models did best for homicide and rape. This may simply reflect the weakness of the current-crime indicators for these two variables, given their low counts. Results based on small numbers are analogous to results based on small samples. The latter are more variable from sample to sample than is commonly believed (Tversky and Kahneman 1971). Analogously, small numbers like yearly murder or rape counts in small areas like communities are also highly variable from year to year, even after spatial smoothing, compromising the predictive impact of current crime.

But for the four most frequently occurring serious crimes, the main takeaway lesson at the community level tell a two-part story about place distinctiveness in terms of crime levels (Molotch, Freudenburg, and Paulsen 2000: 792), while at the same time raising questions. One part of the story is what Hawley (1950) and Bursik (1986) would see as ongoing ecological continuity, or what Molotch et al. (2000: 792) would see as "tradition." Current crime shapes future crime. But a key question is "how the continuity works" (Molotch, Freudenburg, and Paulsen 2000: 793). In community criminology broadly, work has concentrated more on understanding community determinants of crime levels rather than impacts of community crime levels. More insight is needed into the dynamics, whether those are within the community or nearby, that maintain either high or low crime levels from year to year. The second part of the story is ecological discontinuity in a Hawley/Bursik frame. Molotch et al. (2000: 792) would call

these impacts of place "character" while structural criminologists would see these as reflections of ongoing power differentials and related conflicts (Hagan and Palloni 1986). As we think about how structural setting conditions shape cultural dynamics, including social and political processes, the basic systemic model of community crime rates (Bursik and Grasmick 1993) presents one set of possibilities about how all this might work. Other models offer different suggestions. There are numerous challenges to figuring all this out.

Another future challenge, and one where there may be less theoretical guidance, is determining whether sub-city, regional discrepancies are at work, shaping the dynamics described here differently in different places. For example Graif and Sampson (Graif and Sampson 2009) found language diversity and foreign born composition had differently signed significant impacts on homicide in different parts of Chicago (Graif and Sampson 2009). There are tools for considering such dynamics (Anselin 1988, Fotheringham, Brunsdon, and Charlton 2002). But whether geographically weighted regression or a spatial Chow test (or equivalent) is used, the key questions are (1) how much can accuracy be improved?; (2) where is our theoretical guidance on how these extra-community influences operate (Taylor 2015: 117-119)?; and, from the policy-oriented perspective of crime analysts, (3) are the model accuracies gained significant enough and durable enough to justify the additional modeling complexities?

The most significant limitation of the current work is the inability of these long term models to remove spatial autocorrelation from the outcomes. Because the outcome already was spatially smoothed, a further spatially smoothed crime predictor was too diffuse theoretically. Further, such a predictor sometimes created "beta bounce" problems (Gordon 1968) in the rest of the model. We cannot simultaneously test the net impact of current structure and crime while also introducing a doubly spatially lagged crime outcome as a predictor.

One thing that might seem to be a significant limitation but, we would argue, is not, is the sparseness of the predictor space. This does not mean that we have created a mis-specified model. Tests with additional predictors such as residential stability did not alter the significance pattern of the predictors reported here or the patterns of (observed minus predicted) relative frequencies for different counts.

Limits of the long term models are perhaps partially counterbalanced by study strengths which include: a focus on the theoretically most relevant and empirically most supported community crime demographic correlates; tests of model robustness by repeating models using different amounts of spatial smoothing for the outcome; a focus on predictors that worked as expected across six serious crimes*;* and constraining the predictor space to items readily available *at no cost* to crime analysts (it is worth noting that models with spatially smoothed outcomes based on 7-9 nearest neighbors provided closely comparable results).

### *Comparing long-term and short-term crime predictions*

Building on the results of the first study, we rolled the prediction period forward so that we were predicting 2013 crime based on census data from 2006-2011 and crime from 2012. We also added a value for grid cells based on the near-repeat hypothesis, a value that summed near-repeat odds ratios to create an event dependency (short-term risk) surface.

The near repeat odds ratios (Table 2 on page 37) showed a significant burglary risk to properties from zero to 1500 feet within seven days of an originator event, with lessening – but still elevated – risk for the subsequent week. This is entirely in agreement with previous research, both in Philadelphia and other locations. That work has demonstrated a reliably robust near repeat burglary pattern across jurisdictions (Johnson et al. 2007, Townsley, Homel, and

Chaseling 2003). Robbery, while demonstrating smaller odds ratio values, also demonstrated a risk from 250 to 1,000 feet within seven days of an originator event.

Researchers raised on interpretation of $R^2$ values from traditional regression models will likely look upon the pseudo $R^2$ results in Table 7 on page 61 with concern. Several points, however, should be borne in mind. Pseudo $R^2$ values, like the ones shown here which are in essence Efron's pseudo $R^2$, are difficult to interpret and are not directly comparable to OLS $R^2$. "There is no clear interpretation of values other than 0 or 1, nor is there any standard by which to judge if the value is 'large enough'" (Long 1997: 105). Second, the challenge of the test being undertaken is significant – to predict the likelihood of a burglary or robbery in a grid cell approximately one city block in length during a two week window. This small chance of success contributes to the low pseudo $R^2$ values; how much is hard to say. What is of particular importance here is the relative change in pseudo $R^2$ values as we move through the models. Combined with a change in the BIC score, they give us an indication of how the models improve with the additional variables. Finally, in some ways BIC is the most relevant because it is simultaneously considering model accuracy and complexity. The penalty for complexity makes sense in a practical sense because additional indicators mean additional work for analysts who must find and process those data.

The jump in improvement from the short-term event dependency to the long-term risk heterogeneity model (that is from model 3 to model 4) is substantial. This demonstrates, *we believe for the first time*, empirical support for Sherman's contention that the operationalization of predictive policing on short-term changes in crime 'is premised on the already-falsified claim that hot spots are not stable … many police agencies map hot spots on the basis of far too short a time period, generally less than a year. It is unlikely that crime will either be concentrated or

predictable with such short time periods' (Sherman et al. 2014: 108). The relatively dramatic improvement in BIC (reducing by in excess of 3,000 points) and change in pseudo $R^2$ values for both robbery and burglary are indicative that long-term crime potential remains the dominant mechanism by which areas, even very small scale ones, provide evidence, even in very small time windows, to  of serious crimes. The consequences of socioeconomic stratification and racial settlement patterns generate later crime occurrences on a regular basis.

A model combining community structural characteristics, crime counts from the previous year, and an estimate of near repeat activity generated the best results overall. Subsequent analysis using constrained and unconstrained mixed effects logit models confirmed that the risk heterogeneity variables—community structure—have more explanatory power for burglary and robbery prediction than short-term near repeat measures reflecting near-term crime continuities.

## Implications for policy and practice

In an information-led or intelligence-led policing (Ratcliffe, 2008) environment, proactive policing requires a predictive ability, the opportunity to get ahead of the criminal element. One operational challenge has been the puzzling criminological conundrum of whether underlying structural variables based on population characteristics are the overwhelmingly dominant force for current crime patterns, or if recent events and crime spikes are strong enough to supersede long-term crime potential to create new and emerging patterns (i.e., Berk and MacDonald's (2009) crime regimes). The ability of this project to actually estimate the relative weight of both the long- and short-term has been, we believe, a sizable step forward in understanding spatial crime patterns.

For urban, small-scale communities, crime *and* structural predictors together generate the best crime prediction model for four out of six serious crimes in the long-term (year to year). This performance suggests ecological crime continuities are operative over time while, at the same time, ecological crime discontinuities, linked to current structural conditions, also unfold over time. The use of structural predictor variables will enhance analysts' abilities to inform police executives about which areas in their jurisdiction are most likely to foster criminal activity in the medium to long-term future. Indirectly, the research also suggests some practical value to the yearly estimates from the American Community Survey. Fortunately in the US, demographic data are freely available to all law enforcement agencies through the U.S. Census Bureau and they are accessible with the development of the Alchemist extraction tool – a valuable byproduct of this funded project.

SES and racial composition prove sturdy crime predictors over all six crime types examined in this research, as would be expected by structural criminologists. Serious work remains ahead identifying the processes maintaining these ecological crime continuities, and the processes that generate the unfolding ecological discontinuities.

It should be noted however, that there are ethical considerations with the use of a racial composition variable in the prediction of crime victimization. There is significant community concern regarding the potential for racial profiling during the delivery of policing services (Welsh 2007). As such, it is understandable that – irrespective of the reliability of the empirical evidence – some segments of the community would have concerns about the use of a race variable in the prediction of crime, especially if coded into a software program used by police departments. It should be stressed at this point that the processes described in this report say nothing about offender characteristics: the report is focused on locations of criminal

victimization. The software program that builds on the research herein is therefore designed to identify and predict locations where crime victimization occurs. It is an unfortunate reality that in larger urban areas such as Philadelphia, minority community neighborhoods are over-represented as crime locations and our black and Hispanic citizens are over-represented as crime victims, especially for violent street crime.

Few would disapprove if these findings were used to prioritize the delivery of social services and other long-term community assistance services, and we would certainly promote the use of these long term findings in this manner. It is also a reality, however, that the software emerging from this research does have a short-term application that police departments are more suited to exploit than other government agencies. This is because these agencies have the ability to quickly reassign patrol officers. Given these considerations, the software program does not exploit the race variable in its default condition. Rather, if the user wishes to improve the accuracy of the predictions by using the race variable within the software, then the user has to change a parameter setting in the program. The user manual explains how to do this.

The introduction of a short-term risk adjustment based on the near repeat hypothesis increases our understanding of the relative strength of crime predictors. Shane Johnson (2010: 361) has remarked in relation to spatial patterns of crime, “it would seem that neither the risk heterogeneity nor boost hypothesis in isolation can explain the observed patterns; both have a part to play. Determining the precise contribution of each and if and how this varies over space and time would be a useful next step”. This is one of the central questions that the second part of the analysis sought to address.

In terms of the precise contribution of each to community criminology, the evidence from our analysis overwhelmingly favors the contribution of long-term risk heterogeneity as a dominant mechanism in the production of geographic patterns of robbery and burglary. With that statement in place however, the inclusion of a measure of near-repeat boost criminal activity does make a significant additional contribution to the explanatory power of the combined metrics. The near-repeat hypothesis was found to exist in both robbery and burglary patterns in Philadelphia, and incorporation of these patterns in the prediction models improved their ability to predict future criminality for the next two weeks.

In sum then, from a theoretical perspective, there are three dynamics operating to generate these short term and small scale crime patterns: long term ecological crime continuities, reflected in the impacts of a previous year's crime levels, ecological crime discontinuities spawned by socioeconomic and racial differences in community settlement patterns, and near term – near in both space and time -- crime continuities, which generate transient and localized impacts on future crime.

At this point, it is useful to recap the project objectives. This project defined the following original objectives from the project application.

1. Create a stand-alone software program that can incorporate theoretically-relevant and crime-linked census-based demographic indices and indicators, and their spatial relationships to geocoded crime events, to create a crime potential surface map of long-term crime-relevant setting conditions.

- This objective has been met through the creation of the PROVE software program that incorporates the risk heterogeneity indicators and creates a mappable potential crime surface.

2. Have the program integrate recent crime events entered by the user; these update the crime potential map surface with additional risk parameters drawn from recent theoretical advances in repeat victimization and near-repeat victimization.
    - The PROVE software program is indeed able to integrate crime events entered by the user, not just in terms of a year-to-year measure of crime hot spots, but also in terms of short-term event dependency predictions.
3. Generate an enhanced predictive map of short-term future crime risk for an area, on a user-determined temporal frequency, by combining both the crime potential and event dependency surfaces.
    - The PROVE software program generates for the user the required short-term future crime risk for an area.
4. Test this using multiple years of Philadelphia (PA) recorded crime data, varying the surface weights to identify an optimal distribution for crime prediction.
    - The report above reports the results of the application of the research methods across various years of crime data for Philadelphia, PA, stretching from 2009 to 2013.
5. Create a user manual and a web site to support the user community with advice and assistance for optimum program application.
    - At the time of writing this report, the user manual is in a draft format and is included in this report as an appendix. There is also the beginnings of a web

site that will support the distribution of the software as well as update users on different research products that have been developed. The draft web site can be found at:

- *http://www.cla.temple.edu/cj/center-for-security-and-crime-science/prove-predictive-policing-software/* though there is also a short url available at: *http://bit.ly/PROVEsoftware*

6. Distribute the result as a free stand-alone desktop computer program for use by US law enforcement agencies (and other agencies interested in crime prevention).
    - This software program exists with the PROVE software. See *http://bit.ly/PROVEsoftware*

As a review of the above project objectives indicates, the project has succeeded in each of the program objectives. That being said, there are some limitations that should be recognized. The empirical results were based on the city of Philadelphia and there is an external validity question regarding the applicability of the results to other locations. The external validity of the empirical results in an empirical question that can only be answered by replication in other locations. In short, external validity is not a study limitation per se; rather, it is just a future question waiting to be answered. Fortunately, the PROVE software program is designed to adjust metrics and program parameters for different locations. Therefore even if other demographic coefficients are discovered, the software has a degree of flexibility to make necessary adjustments.

Second, the software program does make use of the Alchemist program that enables easy downloading of census data from the US Census website. This website is outside of our control and it may be that future changes to the structure of the website or the data structure may necessitate changes to the software program so that PROVE can continue to easily access census

information. At the time of writing, Azavea are committed to supporting the software program, but it is unreasonable to expect an open-ended commitment from them.

## Implications for further research

The research component of this project has identified a combination of near repeat (event dependency) and risk heterogeneity measures that can predict a degree of future criminality. This project was grounded in a certain reality – that many police department analysts do not easily have access to a range of data sets that would enhance the predictability of crime. For example, risk terrain modeling requires that crime analysts have at their fingertips a range of additional datasets such as pawn shops, check-cashing, fast food restaurants, on-site and off-site alcohol establishments, independent grocery stores, department discount stores, convenience stores, tobacco shops, tattoo parlors, and motel/hotel/motor home parks (Drawve in press). These data sets are time consuming to research and maintain, difficult to verify, and sometimes may be expensive to source from outside agencies. By comparison, the PROVE program deliberately only uses crime data (which it is assumed the agency already possesses) and freely accessible Census information. Future research may help the analysis field by determining if any additional accuracy in terms of crime prediction that might stem from the use of risk terrain modeling (or other analytical processes) is a cost benefit relative to the additional effort and cost of data set maintenance.

Based on crime predictability, predictive policing is an emerging tactic relying in part on software predicting the likely locations of criminal events. Predictive policing has been defined as “the application of analytical techniques—particularly quantitative techniques—to identify likely targets for police intervention and prevent crime or solve past crimes by making statistical

predictions" (Perry et al. 2013: xiii). At present the field lacks robust evidence to suggest the appropriate policing tactic in predicted areas. Future research would do well to answer the question of whether different varieties of theoretically informed but also operationally realistic police responses to crime predictions estimated by a predictive policing software program can reduce crime. It may be that the ability to predict crime in the short term is of little value if government is insufficiently flexible to capitalize on this ability. If we consider predictive policing strategies to be related to hot spots policing, then two recent observations from Weisburd and Telep are relevant: (1) there are numerous strategies that have not yet been rigorously tested, and (2) much more needs to be learned about the impact of new technology on policing effectiveness (Weisburd and Telep 2014).

# THE PROVE SOFTWARE UTILITY

The PROVE software program was created by Azavea using findings described in this research project. The software program's name was designed to loosely relate to the Prediction of Repeat Offending and Victimization in the Environment. To access the software please follow the links accessible through *http://www.cla.temple.edu/cj/center-for-security-and-crime-science/prove-predictive-policing-software/* though there is also a short url available at *http://bit.ly/PROVEsoftware*

The draft manuals for the software are available in the appendix of this report; however, please note that at the time of writing the software is being beta tested by a community of crime analysts. The eventual program that will be publicly released may differ from the draft version that we have access to at the time of writing. The draft manuals are only included in the appendix here to allow reviewers of the final report to gain an understanding of the likely structure of the final program. We strongly recommend that updated manuals are downloaded from the web site.

# DISSEMINATION OF RESEARCH FINDINGS

Project results have been distributed in the following ways.

## Journal articles

Taylor, RB, Ratcliffe, JH & Perenzin, A (2015) Can we predict long-term community crime problems? The estimation of ecological continuity to model risk heterogeneity, *Journal of Research in Crime and Delinquency*, 52(3): 635-657.

> In small-scale, intra-urban communities, do fundamental demographic correlates of crime, proven important in community criminology, link to next year's crime levels, even after controlling for this year's crime levels? If they do, it would imply that shifting ecologies of crime apparent after a year are driven in part by dynamics emerging from structural differentials. To the best of the authors' knowledge, this question has not yet been addressed. For Philadelphia (PA) census block groups, 2005 to 2009 data from the American Community Survey and 2009 crime counts were used to predict spatially smoothed 2010 crime counts in three different models: crime only, demographics only, and crime plus demographics. Models are tested for major personal (murder, rape-aggravated assault, and robbery) and property (burglary and motor vehicle theft) crimes. For all crime types investigated except rape and homicide, crime plus demographics resulted in the best combination of prediction/simplicity based on the Bayesian Information Criterion. Socioeconomic status (SES) and racial composition linked as expected theoretically to crime changes. Intercommunity structural differences in power relationships, as reflected in SES and racial composition, link to later crime shifts at the same time that ongoing crime continuities link current and future crime levels. The main

practical implication is that crime analysts tasked with long-term, one-year-look-ahead forecasting may benefit by considering demographic structure as well as current crime.

Further articles are in preparation.

## Conference presentations

Ratcliffe, J.H., Taylor, R.B., & Perenzin, A.R., (2013, March) *Keeping One Step Ahead: Generating Risk Heterogeneity Map Surfaces to Predict Long-Term Crime Patterns*. Paper presented at the Academy of Criminal Justice Sciences Annual Meeting. Dallas, TX. Presenter: Amber Perenzin, MA.

## Software programs

ACS Alchemist. *ACS Alchemist* is an open source tool that enables the extraction of up to 100 variables of the American Community Survey (ACS). The data is extracted directly into a format convenient for display on maps or for use in advanced spatial analysis and modeling. The source code for ACS Alchemist has been released under a GNU General Public License.

PROVE. *PROVE* is a freely-available software program for the prediction of crime. It uses components of ACS Alchemist to access data from the American Community Survey (ACS). The PROVE utility generates crime predictions by combining short-term and long-term crime indicators. First, a model is calibrated by combining short-term and long-term crime indicators for a given year during a model building stage. Using properties of the calibrated models, predictions are calculated for the desired timeframe.

# Web sites

ACS Alchemist is available for download at: *https://github.com/azavea/acs-alchemist*

PROVE is available through *http://www.cla.temple.edu/cj/center-for-security-and-crime-science/prove-predictive-policing-software/* though there is also a short url available at *http://bit.ly/PROVEsoftware*

# REFERENCES

Citations used in this report:


Abbott, Andrew. 2001. *Time Matters: On Theory and Method*. Chicago: University of Chicago Press.

Alba, R. D., J. R. Logan, and P. E. Bellair. 1994. "Living with Crime - the Implications of Racial Ethnic-Differences in Suburban Location." *Social Forces* 73 (2):395-434.

Anselin, Luc. 1988. *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer.

Berk, Richard, and John McDonald. 2009. "The Dynamics of crime regimes." *Criminology* 47 (3):971-1008.

Berry, B. J. L., and J. D. Kasarda. 1977. *Contemporary urban ecology*. New York: Macmillan.

Berry, B.J.L. 1965. "Internal structure of the city." *Law and Contemporary Problems* 30 (111-119).

Bohrnstedt, George W. 1969. "Observation on the measurement of change." In *Sociological methodology*, edited by Edward. F. Borgatta and George W. Bohrnstedt, 113-133. San Francisco: Jossey-Bass.

Bowers, Kate J., Alex Hirschfield, and Shane D. Johnson. 1998. "Victimization revisited." *British Journal of Criminology* 38 (3):429-452.

Bowers, Kate J., and Shane D. Johnson. 2004. "Who commits near repeats? A test of the boost explanation." *Western Criminology Review* 5 (3):12-24.

Bowers, Kate J., Shane D. Johnson, and Ken Pease. 2004. "Prospective hot-spotting: The future of crime mapping?" *British Journal of Criminology* 44 (5):641-658.

Brantingham, Patricia L., and Paul J. Brantingham. 1981-2. "Mobility, notoriety, and crime: A study in the crime patterns of urban nodal points." *Journal of Environmental Systems* 11 (1):89-99.

Bursik, Robert J., and Harold G. Grasmick. 1993. *Neighborhoods and crime*. New York: Lexington.

Bursik, Robert J. Jr. 1984. "Urban dynamics and ecological studies of delinquency." *Social Forces* 63:393-413.

Bursik, Robert J. Jr. 1986. "Ecological stability and the dynamics of delinquency." In *Communities and crime*, edited by Albert J. Reiss and Michael Tonry, 35-66. Chicago: University of Chicago Press.

Bursik, Robert J. Jr., and Jim Webb. 1982. "Community change and patterns of delinquency." *AJS* 88 (1):24-42.

Chainey, Spencer, and Jerry H. Ratcliffe. 2005. *GIS and Crime Mapping*. London: John Wiley and Sons.

Cherlin, A. J. 1981. *Marriage, divorce, remarriage*. Cambridge: Harvard University Press.

Cohen, Lawrence E., and Marcus Felson. 1979. "Social change and crime rate trends: A routine activity approach." *American Sociological Review* 44:588-608.

Cornish, Derek, and Ron Clarke. 1986. *The Reasoning Criminal: Rational Choice Perspectives on Offending*. New York: Springer-Verlag.

Covington, J. C., and R. B. Taylor. 1989. "Gentrification and crime: Robbbery and larceny changes in appreciating Baltimore neighborhoods in the 1970's." *Urban Affairs Quarterly* 25:142-172.

Davenport, Thomas H. 1997. *Information Ecology: Mastering the Information and Knowledge Environment*. New York: Oxford University Press.

Drawve, Grant. in press. "A metric comparison of predictive hot spot techniques and RTM." *Justice Quarterly*.

Ebert, Beth. 2003. What is the best statistic for measuring the accuracy of a forecast?. Centre for Australian Weather and Climate Research. [ONLINE: http://cawcr.gov.au/bmrc/wefor/staff/eee/verif/BestStatistic.html ;.

Farrell, Graham, Sylvia Chenery, and Ken Pease. 1998. Consolidating police crackdowns: findings from an anti-burglary project. London: Policing and Reducing Crime Unit, Research, Development and Statistics Directorate, Home Office.

Felson, Marcus. 1987. "Routine activities and crime prevention in the developing metropolis." *Criminology* 25 (4):911-932.

Forrester, D, M Chatterton, and K Pease. 1988. The Kirkholt Burglary Prevention Project, Rochdale. London: Crime Prevention Unit (Home Office).

Fotheringham, A. Stewart, Chris Brunsdon, and Martin Charlton. 2002. *Geographically Weighted Regression*. Chichester (UK): John Wiley.

Gordon, Robert A. 1968. "Issues in multiple regression." *American Journal of Sociology* 73:592-616.

Gorman, Dennis M., Paul W. Speer, Paul J. Gruenewald, and Erich W. Labouvie. 2001. "Spatial dynamics of alcohol availability, neighborhood structure and violent crime." *Journal of Studies on Alcohol and Drugs* 62 (5):628-636.

Graif, Corina, and Robert J. Sampson. 2009. "Spatial Heterogeneity in the Effects of Immigration and Diversity on Neighborhood Homicide Rates." *Homicide Studies* 13 (3):242-260. doi: 10.1177/1088767909336728.

Haberman, Cory P., and Jerry H. Ratcliffe. 2012. "The predictive policing challenges of near repeat armed street robberies." *Policing: A Journal of Policy and Practice* 6 (2):151-166.

Hagan, J., and A. Palloni. 1986. "Toward a Structural Criminology - Method and Theory in Criminological Research." *Annual Review of Sociology* 12:431-449.

Harries, Keith. 1995. "The ecology of homicide and assault: Baltimore City and County, 1989–91." *Studies on Crime & Crime Prevention* 4 (1):44-60.

Hawley, Amos H. 1950. *Human Ecology: A Theory of Community Structure*. New York City: Ronald Press.

Hunter, A. 1971. "The ecology of Chicago: Persistence and change, 1930 - 1960." *American Journal of Sociology* 77:425-443.

Janson, C-G. 1980. "Factorial social ecology: An Attempt at summary and evaluation." *Annual Review of Sociology* 6:433-456.

Jennings, J. M., R. B. Taylor, R. A. Salhi, C. D. M. Furr-Holden, and J. M. Ellen. 2012. "Neighborhood drug markets: A risk environment for bacterial sexually transmitted infections among urban youth." *Social Science & Medicine* 74 (8):1240-1250. doi: 10.1016/j.socscimed.2011.12.040.

Johnson, Lallen T., Ralph B. Taylor, and Jerry H. Ratcliffe. 2013. "Need drugs, will travel?: The distances to crime of illegal drug buyers." *Journal of Criminal Justice* 41 (3):178-187.

Johnson, S.D., W. Bernasco, K.J. Bowers, H. Elffers, J.H. Ratcliffe, G.F. Rengert, and M. Townsley. 2007. "Space-time patterns of risk: A cross national assessment of residential burglary victimization." *Journal of Quantitative Criminology* 23 (3):201-219.

Johnson, S.D., K.J. Bowers, D. Birks, and K. Pease. 2009. "Predictive mapping of crime by ProMap: Accuracy, units of analysis and the environmental backcloth." In *Putting Crime in its Place: Units of Analysis in Geographic Criminology*, edited by D. Weisburd, W. Bernasco and G.J.N. Bruinsma, 171-198. New York: Springer-Verlag.

Johnson, Shane D. 2010. "A brief history of the analysis of crime concentration." *European Journal of Applied Mathematics* 21 (4/5):349-370.

Johnson, Shane D., Lucia Summers, and Ken Pease. 2009. "Offender as forager? A direct test of the boost account of victimization." *Journal of Quantitative Criminology* 25 (2):181–200.

King, Gary. 1988. "Statistical Models for Political Science Event Counts: Bias in Conventional Procedures and Evidence for the Exponential Poisson Regression Model." *American Journal of Political Science* 32 (3):838-863. doi: 10.2307/2111248.

Kornhauser, Ruth Rosner. 1978. *Social sources of delinquency*. Chicago: University of Chicago Press.

Laycock, Gloria. 2001. "Hypothesis-based research: The repeat victimization story." *Criminology and Criminal Justice* 1 (1):59-82.

Lee, M. R., and G. C. Ousey. 2005. "Institutional access, residential segregation, and urban black homicide." *Sociological Inquiry* 75 (1):31-54.

Logan, John R. 1978. "Growth, Politics, and the Stratification of Places." *American Journal of Sociology* 84 (2):404-416. doi: 10.2307/2777855.

Logan, John R., and Harvey Molotch. 1987. *Urban fortunes*. Berkeley: University of California Press.

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks: Sage.

Long, J. Scott, and Jeremy Freese. 2006. *Regression Models for Categorical Dependent Variables Using Stata*. Second ed. College Station, Texas: Stata Press.

Maddala, G.S. 1983. *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Marshall, B., and M. Townsley. 2006. Needles or needless: The Applicability of data mining to the academic criminological community. London: Jill Dando Institute, University College London.

Massey, Douglas S. 1998. "Review: Back to the Future: The Rediscovery of Neighborhood Context." *Contemporary Sociology* 27 (6):570-572.

McCord, E.S., J.H. Ratcliffe, R. M. Garcia, and R. B. Taylor. 2007. "Nonresidential crime attractors and generators elevate perceived neighborhood crime and incivilities." *Journal of Research in Crime and Delinquency* 44 (3):295-320.

Molotch, H, W. R. Freudenburg, and K. E. Paulsen. 2000. "History repeats itself, but how?: City character, urban tradition, and the accomplishment of place." *American Sociological Review* 65 (6):791-823.

Murphy, Alexandra K. 2007. "The Suburban ghetto: The Legacy of Herbert Gans in understanding the experience of poverty in recently impoverished American suburbs." *City & Community* 6 (1):21-37.

Openshaw, Stan. 1984. "The modifiable areal unit problem." *Concepts and Techniques in Modern Geography* 38:41.

Pease, Ken. 1998. "Repeat victimisation: Taking stock." *Police Research Group: Crime Detection and Prevention Series* Paper 90:1-48.

Pepper, John V. 2008. "Forecasting crime: A City-level analysis." In *Understanding Crime Trends: Workshop Report*, edited by National Research Council, 177-210. Washington, D.C.: The National Academies Press.

Perry, Walter L., Brian McInnis, Carter C. Price, Susan C. Smith, and John S. Hollywood. 2013. Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations. Washington DC: Rand Corporation.

Peterson, Ruth D., and Lauren J. Krivo. 2005. "Macrostructural analyses of race, ethnicity, and violent crime: Recent lessons and new directions for research." *Annual Review of Sociology* 31:331-356.

Peterson, Ruth D., and Lauren J. Krivo. 2010. *Divergent Social Worlds: Neighborhood Crime and the Racial-Spatial Divide*. New York: Russell Sage

Polvi, N., T. Looman, C. Humphries, and K. Pease. 1991. "The time course of repeat burglary victimization." *British Journal of Criminology* 31 (4):411-414.

Pratt, Travis C., and Francis T. Cullen. 2005a. "Assessing macro-level predictors and theories of crime: A meta-analysis." In *Crime and Justice: a Review of Research*, 373-450.

Pratt, Travis C., and Francis T. Cullen. 2005b. "Assessing macro-level predictors and theories of crime: A meta-analysis." *Crime and Justice* 32:373-450.

Quarmby, Neil. 2009. "Futures work in strategic criminal intelligence." In *Strategic Thinking in Criminal Intelligence (2nd edition)*, edited by Jerry H Ratcliffe. Sydney: Federation Press.

Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." *Sociological Methodology* 25:111-163.

Ratcliffe, Jerry H. 2014. "What is the future... of predictive policing?" *Translational Criminology* 2014 (Spring):4-5.

Ratcliffe, Jerry H. 2016. *Intelligence-Led Policing*. Abingdon, Oxon.: Routledge.

Ratcliffe, Jerry H., and Michael J. McCullagh. 1998. "Identifying repeat victimisation with GIS." *British Journal of Criminology* 38 (4):651-662.

Ratcliffe, Jerry H., and Michael J. McCullagh. 2001. "Chasing ghosts? Police perception of high crime areas." *British Journal of Criminology* 41 (2):330-341.

Ratcliffe, Jerry H., and G.F. Rengert. 2008. "Near repeat patterns in Philadelphia shootings." *Security Journal* 21 (1-2):58-76.

Rengert, George F., and J. Wasilchick. 1985. *Suburban burglary: A time and place for everything*. Springfield, IL: C.C. Thomas Publishing.

Rengert, George F., and J. Wasilchick. 2000. *Suburban Burglary: A Tale of Two Suburbs*. Second ed. Springfield, IL: C.C. Thomas Publishing.

Sampson, Robert J., and J. L. Lauritsen. 1994. "Violent victimization and offending: individual, situational- and community-level risk factors." In *Understanding and preventing violence (Volume 3: Social Influences)*, edited by A. J. Jr Reiss and J. A. Roth, 1-114. Washington, DC: National Academy Press.

Sampson, Robert J., and J.D. Wooldredge. 1987. "Linking the micro- and macro-level dimensions of lifestyle-routine activity and opportunity models of predatory victimization." *Journal of Quantitative Criminology* 3 (4):371-393.

Santos, Rachel Boba. 2014. "The effectiveness of crime analysis for crime reduction: Cure or diagnosis?" *Journal of Contemporary Criminal Justice* 30 (2):147-168.

Shaw, C. R. 1929. *Delinquency areas: A study of the geographic distribution of school truants, juvenile delinquents and adult offenders in Chicago*. Chicago: University of Chicago Press.

Sheley, J. F., and V Brewer. 1995. "Possession and carrying of firearms among suburban youth." *Public Health Reports* 110 (1):18.

Sherman, Lawrence W., P. Gartin, and M. E. Buerger. 1989. "Hot spots of predatory crime: Routine activities and the criminology of place." *Criminology* 27 (1):27-55.

Sherman, Lawrence W., Stephen Williams, Barak Ariel, Lucinda R. Strang, Neil Wain, Molly Slothower, and Andre Norton. 2014. "An integrated theory of hot spots patrol strategy: Implementing prevention by scaling up and feeding back." *Journal of Contemporary Criminal Justice* 30 (2):95-112.

St. Jean, Peter K. B. 2007. *Pockets of Crime: Broken Windows, Collective Efficacy, and the Criminal Point of View*. Chicago: University of Chicago Press.

Stucky, Thomas D., and John R. Ottensmann. 2009. "Land use and violent crime." *Criminology* 47 (4):1223-1264.

Suttles, G. D. 1972. *The social construction of communities*. Chicago: University of Chicago Press.

Taylor, R. B. 1998. "Crime in small scale places: What we know, what we can do about it." In *Research and Evaluation Conference 1997*, 1-20. Washington, DC: National Institute of Justice.

Taylor, R. B. 2000. "Crime and human ecology." In *Explaining criminals and crime*, edited by Ray Paternoster and Ronet Bachman. Los Angeles: Roxbury.

Taylor, R. B. 2001. *Breaking Away from Broken Windows: Evidence from Baltimore Neighborhoods and the Nationwide Fight Against Crime, Grime, Fear and Decline*. New York: Westview Press.

Taylor, R. B., and J. Covington. 1988. "Neighborhood changes in ecology and violence." *Criminology* 26:553-589.

Taylor, Ralph B. 2015. *Community Criminology: Fundamentals of Spatial and Temporal Scaling, Ecological Indicators and Selectivity Bias*. New York: New York University Press.

Townsley, Michael, Ross Homel, and Janet Chaseling. 2003. "Infectious burglaries: A test of the near repeat hypothesis." *British Journal of Criminology* 43 (3):615-633.

Townsley, Michael, Shane D. Johnson, and Jerry H. Ratcliffe. 2008. "Space time dynamics of insurgent activity in Iraq." *Security Journal* 21 (3):139-146.

Tseloni, Andromachi, and Ken Pease. 2003. "Repeat personal victimization: 'Boosts' or 'Flags'?" *British Journal of Criminology* 43 (1):196-212.

Tversky, Amos, and Daniel Kahneman. 1971. "Belief in the law of small numbers." *Psychological Bulletin* 76 (2):105-110. doi: 10.1037/h0031322.

Velez, Mara B, and Kelly Richardson. 2012. "The political economy of neighbourhood homicide in chicago: The role of bank investment." *British Journal of Criminology* 52 (3):490-513.

Verbrugge, L. M., and R. B. Taylor. 1980. "Negative and positive effects of metropolitan population density." *Urban Affairs Quarterly* 16:135-160.

Weisburd, David, Shawn Bushway, Cynthia Lum, and Sue-Ming Yang. 2004. "Trajectories of crime at places: A longitudinal study of street segments in the City of Seattle." *Criminology* 42 (2):283-321.

Weisburd, David, and John Eck. 2004. "What can police do to reduce crime, disorder, and fear?" *The Annals of the American Academy of Political and Social Science* 593 (1):43-65.

Weisburd, David, Elizabeth R. Groff, and Sue-Ming Yang. 2012. *The Criminology of Place: Street Segments and Our Understanding of the Crime Problem*. Oxford: Oxford University Press.

Weisburd, David, and Cody Telep. 2014. "Hot Spots Policing: What we know and what we need to know." *Journal of Contemporary Criminal Justice* 30 (2):200-220.

Welsh, Kelly. 2007. "Black criminal stereotypes and racial profiling." *Journal of Contemporary Criminal Justice* 23 (3):276-288.

Wyly, E. K. 1999. "Continuity and change in the restless urban landscape." *Economic Geography* 75 (4):309-338.

# APPENDICES

## *Construction of indices used in this report*

Source: Online appendix to Taylor, RB, Ratcliffe, JH & Perenzin, A (2015) Can we predict long-term community crime problems? The estimation of ecological continuity to model risk heterogeneity, *Journal of Research in Crime and Delinquency*, 52(3): 635-657.

## *PROVE Manual*

This is a draft document. Please refer to *www.bit.ly/PROVEsoftware* for latest version.

## *PROVE quick start guide*

This is a draft document. Please refer to *www.bit.ly/PROVEsoftware* for latest version.

# Online Appendix
# Taylor, Ratcliffe & Perenzin

This appendix provides details for the construction of two indices described in the paper: socioeconomic status, and residential stability, and how specific variables cross-reference with 2005-2009 American Community Survey Census Block Group variables. Data are for Philadelphia (PA).

**Table 1**

Socioeconomic Status Index: Name of ACS variables

| Variable | ACS Variable Name | Study Variable Name |
|---|---|---|
| households with income less than $20K | B19001002 | nhhlt10k |
| | B19001003 | nhh10k |
| | B19001004 | nhh15k |
| | B19001001 | nhh_inc |
| households with income greater than $50K | B19001011 | nhh50k |
| | B19001012 | nhh60k |
| | B19001013 | nhh75k |
| | B19001014 | nhh100k |
| | B19001015 | nhh125k |
| | B19001016 | nhh150k |
| | B19001017 | nhh200kp |
| | B19001001 | nhh_inc (**denominator)** |
| median house value | B25077001 | medhval |
| median household income | B19013001 | medhinc |

**Table 2**

Socioeconomic Status Index: Transformation of ACS Variables to Create Variables for Index

| Variable | ACS Variables | Study Variable Name |
|---|---|---|
| Percent households with income less than $20,000 in 2009 dollars | 100*(B19001002+003+004) / B19001001 | phhlt20k |
| Percent households with income greater than $50,000 in 2009 dollars | 100*(B19001011+012+013+014+015+016+017) / B19001001 | phhgt50k |
| natural log 1+Median house value | ln(1+B25077001) | lnmdval |
| natural log 1+median household income | ln(1+B19013001) | lnhhinc |

**Table 3**

Descriptive statistics: SES variables and index

| Variable | Variable name | N | Mean | Median | Standard Deviation | Minimum | Maximum | American Community Survey Table ID Cross reference for constituent variables | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Value or numerator | Denominator (if applicable) |
| Percent households with income less than $20,000 in 2009 dollars (reversed when factored into index) | phhlt20k | 1766 | 33.17 | 30.39 | 21.67 | 0.00 | 100.00 | (B19001002 + 003 + 004) | B19001001 |
| Percent households with income greater than $50,000 in 2009 dollars | phhgt50k | 1766 | 34.51 | 31.90 | 22.17 | 0.00 | 100.00 | (B19001011 + 012 + 013 + 014 + 015 + 016 +017) | B19001001 |
| natural log 1 + median house value | lnmdval | 1647 | | 11.56 | .73 | 8.97 | 13.82 | B25077001 | na |
| natural log 1 + median household income in 2009 dollars | lnhhinc | 1749 | | 10.42 | .62 | 7.82 | 12.43 | B19013001 | na |
| Socioeconomic status index (higher score = higher status) | sesindx | 1771 | -0.06 | -0.06 | 0.88 | -2.91 | 2.60 | na | na |

Note. Units = Philadelphia census block groups. Results weighted by the number of households providing income for household income calculations. Averages not shown for variables based on median values. Index score calculated if scores on at least three items out of four available. Cross reference identifies starting variables. na = not applicable. All variables z scored and then averaged to create index scores. Reversed variables multiplied by -1. Data from 2005-2009 American Community Survey (ACS) for Philadelphia.

**Table 4**

Residential Stability Index: Name of ACS Variables

| Variable | ACS Variable Name | Study Variable Name |
|---|---|---|
| Percent Owner Occupied Housing | B25003002<br>B25003001 | noohu_2<br>nocc_hu2 (**denominator**) |
| percent occupied housing units where current residents moved in before 2005 | B25038003<br>B25038010<br>B25038001 | omv2005<br>rmv2005<br>tfammvin (**denominator**) |
| percent occupied housing units where current residents moved in before 2000 | B25038003<br>B25038004<br>B25038010<br>B25038011<br>B25038001 | omv2005<br>omv2000<br>rmv2005<br>rmv2000<br>tfammvin (**denominator**) |

**Table 5**

Residential Stability Index: Transformation of ACS Variables to Create Variables for Index

| Variable | ACS Variables | Variable Name |
|---|---|---|
| Percent owner occupied housing units | 100*(B25003002) / B25003001 | pownochh |
| Percent occupied housing units where current residents moved in before 2005 | 100 - [ 100*(B25038003+010) / B25038001 ] | plong |
| Percent occupied housing units where current residents moved in before 2000 | 100 – [ 100*(B25038003+004+010+011) / B25038001 ] | plonger |

**Table 6**

Descriptive statistics: Residential Stability Variables and Index

| Variable | Variable name | N | Mean | Median | Standard Deviation | Minimum | Maximum | American Community Survey Table ID Cross reference for constituent variables | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Numerator | Denominator |
| Percent owner occupied housing units | pownochh | 1766 | 57.30 | 58.92 | 25.18 | 0.00 | 100.00 | B25003002/ | B25003001 |
| Percent occupied housing units where current residents were there before 2005 | plong | 1766 | 73.92 | 75.69 | 18.12 | 0.00 | 100.00 | (B25038003 + B25038010) | B25038001 |
| Percent occupied housing units where current residents were there before 2000 | plonger | 1766 | 49.00 | 49.01 | 20.90 | 0.00 | 100.00 | (B25038003 + B25038004 + B25038010 + B25038011) | B25038001 |
| Stability index (higher = more stability) | stabindx | 1771 | -0.02 | 0.03 | 0.85 | -2.92 | 1.84 | | |

Note. Units = Philadelphia census block groups. Results weighted by the number of housing units where tenure status was determined (B25038001). Higher score = higher stability. For pownochh, ratio shown multiplied by 100 to create percentages. For plong and plonger, ratio shown in table was multiplied by 100 to create percentages, then subtracted from 100. All variables z scored and then averaged to create index scores.

# PROVE Manual

**DRAFT – Please check the PROVE website for the latest version.**

**www.bit.ly/PROVEsoftware**

# Contents

List of Figures

# Overview

This manual describes the process of generating crime predictions using the PROVE software. The PROVE utility generates crime predictions by combining short-term and long-term crime indicators. First, a model is calibrated by combining short-term and long-term crime indicators for a given year during a model building stage. Using properties of the calibrated models, predictions are calculated for the desired timeframe.

To install the program, the execution file (.exe) will need to be downloaded from this website:

http://s3.azavea.com.s3.amazonaws.com/temp/deleteafter/2016-01/PROVE-Utility-install.exe

# Data preparation

The PROVE utility requires the user to have 2 files:

1. A csv file containing 3 years of specially formatted crime data (.csv file)
2. A shapefile of a jurisdiction boundary (.shp file)

## Crime data

In order to use the PROVE software, crime data need to be saved in a format the software will recognize.

The user will need to have a minimum 3 complete years of geocoded crime data to run the program. For an example, if a user would like to predict burglaries in the 2013 calendar year, a .csv that contains burglary incidents from all of 2010, 2011 and 2012 will be needed. A fourth year of crime data will be needed if the user wishes to test the accuracy of a prediction for a given time period.

The user will need to generate separate .csv files for *each crime type*. Each .csv file needs to have the following fields in this order:

1. id (number/string)—This field should contain a unique identifier for each crime event in your dataset.
2. pointx (number)—The data used in the program need to be geocoded. This field should contain the x coordinate for the location of each crime incident.
3. pointy (number)—The y coordinate for the location of the crime incident.
4. eventtime (ISO datetime)—This is the date and time that each event occurred. This field requires the following format: yyyy-mm-dd hh:mm:ss.
5. class (string)—The type of crime event you are analyzing, ex: burglary, robbery, etc.

An example of the .csv file format is below:

| id | pointx | pointy | eventtime | class |
|---|---|---|---|---|
| 201135003729 | 2702545.003 | 265037.0758 | 2011-01-08 00:00:00 | AgAssault |
| 201125004221 | 2698695.005 | 257534.6473 | 2011-01-15 00:00:00 | AgAssault |
| 201112004021 | 2676429.744 | 225424.8352 | 2011-01-16 00:00:00 | AgAssault |

### Jurisdiction shapefile

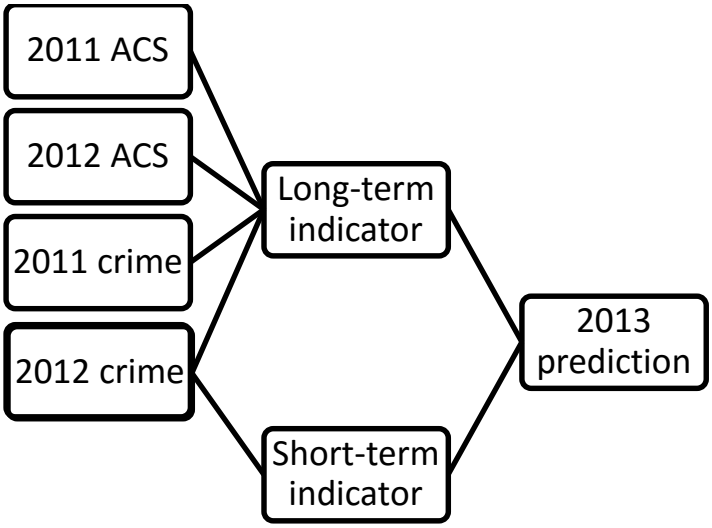
The jurisdiction shapefile should be one polygon. For jurisdictions with unincorporated land, the polygon may be multi-part. If the polygon is multi-part those parts do not necessarily have to be contiguous. Multiple single-part polygons should be dissolved into a single, multi-part jurisdiction polygon.

## Summary

The remainder of this manual will detail how the software generates crime predictions.

The PROVE utility approximates the locations of future crime hotspots using both long-term and short-term crime indicators. The short-term indicator is calculated with the most recently available crime data and is informed by a near-repeat analysis for the previous year. The long-term crime indicator is calculated with two years of prior crime data and two years of census data from the American Community Survey (ACS).

The software generates a crime prediction in a two-step process. In the first step, census and crime data are analyzed in a model building stage. During this stage, the data are calibrated to determine how much weight should be given to the long-term crime indicator relative to the short-term indicator. If the user wishes to generate crime predictions for 2014, this model building stage would test the data to make a prediction for the previous year, 2013.

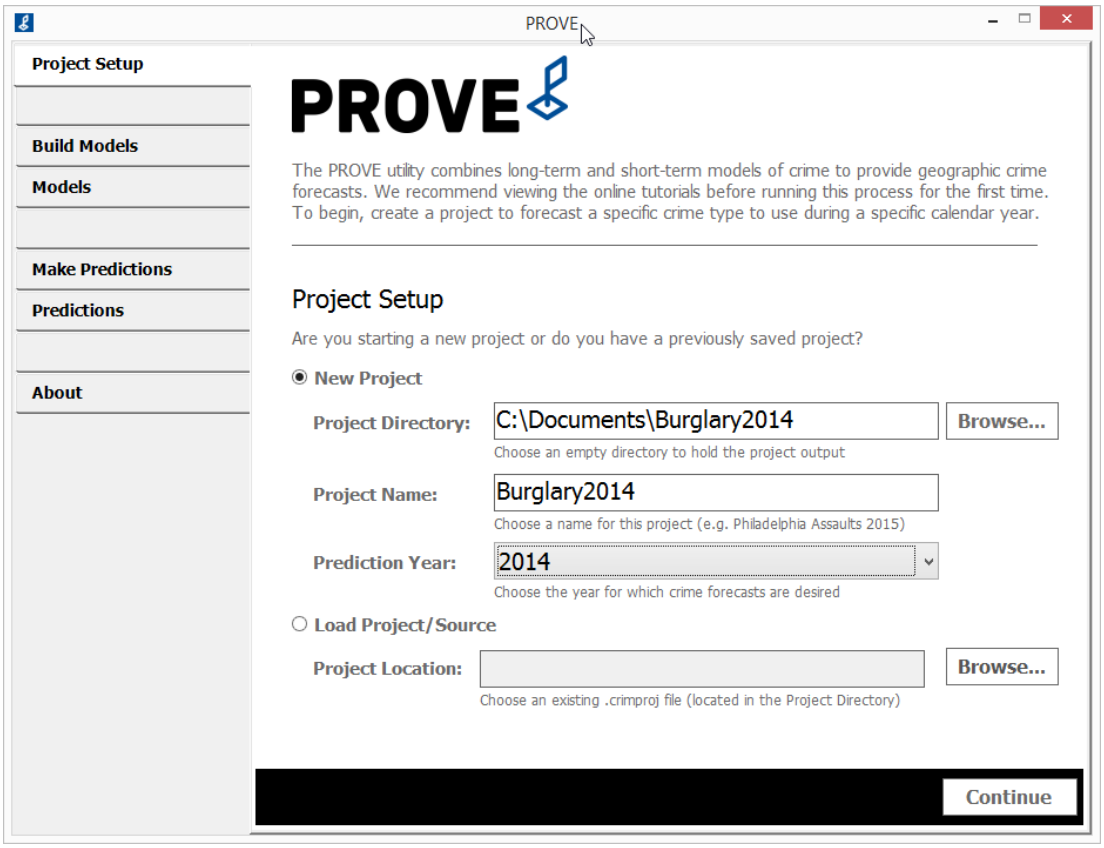



In the second phase of the program, the data from the model building stage are used to inform the crime prediction model. Using the previous example, the 2014 predictions would be informed by the models that were used to make 2013 predictions in the model building stage.

## Using the PROVE utility

When the utility is opened, a project setup window will open. In this window, there are options to start a new project or load a previous project. To start a new project, choose a project directory, a project name and the desired prediction year. To continue working on a previous

project, click on the Browse… button and select the appropriate project directory, then click Continue.




The PROVE utility creates crime predictions using both long-term and short-term crime indicators. The next set of options will be used to generate the long-term crime indicator. Long-term crime indicators include census data and prior crime data.

To extract the appropriate census data, select the state in which the jurisdiction is located. Next, select the jurisdiction boundary shapefile and the crime data .csv file. The jurisdiction boundary is used to select the census block groups that overlap the crime jurisdiction. For a list of census variables that are extracted during this process, see Appendix A. These variables are used to create 2 demographic variables reflecting socio-economic status and the total population. In research mode, a race variable is also created.

Three years of crime data are also used to generate the long-term crime indicator. Crime data need to be specified as unprojected or projected. If the data are projected, they should match the projection of the jurisdiction boundary shapefile.

To generate the long-term crime indicator, the 2 census variables (3 if you are using research mode) and 1 year of prior crime data are used as independent variables in a regression model to predict later crime using census block groups as the unit of analysis (For more detailed information about this process, see Appendix B). There are settings that can be adjusted to modify how the long-term indicator is generated. These settings include the type of regression

model and the amount of spatial smoothing used in the analysis. These options are explained in more detail in the following section.

## Regression method

The default regression method is called **GLM** (the generalized linear model). If GLM is selected, the long-term crime indicator will be calculated using a negative binomial regression model. Negative binomial regression models are appropriate for many count models. The default method does not place constraints on the direction of the link between each predictor and the crime count outcome. BIC scores can be used to select the best model.

**GAM** (Generalized Additive Model) does not assume a linear relationship between the predictor variables and the outcome variable like the GLM models do. Instead, GAM models uses smoothing functions which are summed, or added, to generate B coefficients. This method does not place constraints on the direction of the link between each predictor and the crime count outcome.

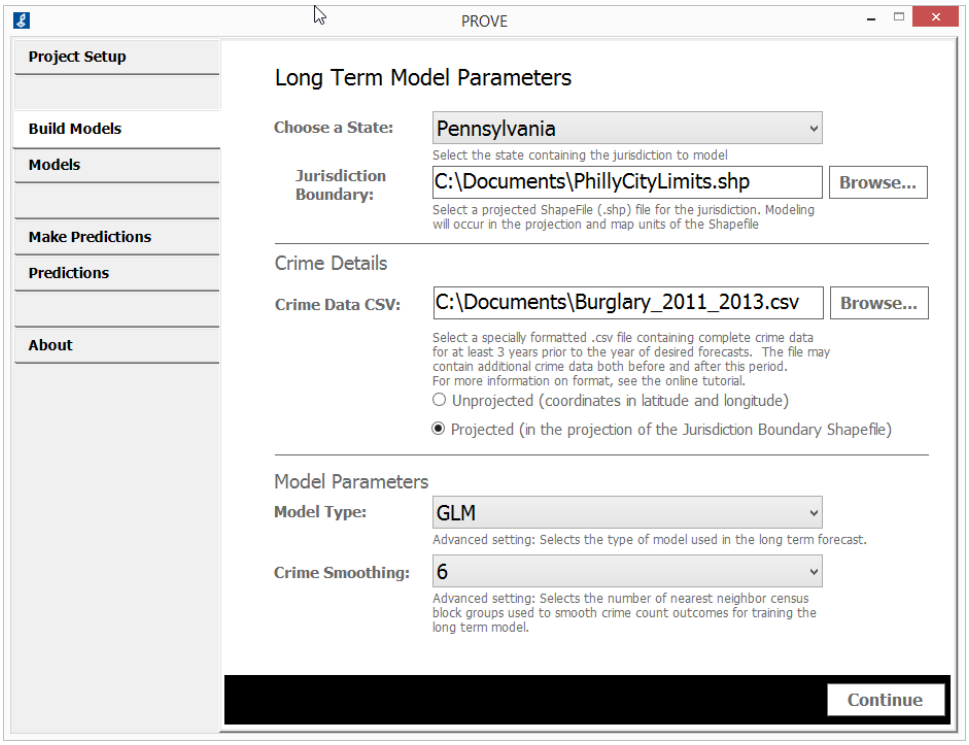
**SCAM** (Shape Constrained Additive Model) allows you to enforce the shape of the relationship that each variable in the model has with the crime count outcome. For example, with this model you can force a positive relationship between prior crime and future crime (monotonic increasing). If a predictor variable does not have a relationship with the crime count outcome that aligns with the specified shape constraint, it is automatically dropped from the regression model.

There are a number of shape constraints that can be specified with this model. By default, the utility constrains prior crime and the total population using the monotone increasing function, while the SES index is constrained using the monotone decreasing function. These parameters can only be changed by modifying the longterm-build.R script.

**GLM Net** models are linear penalized regression (elastic net) models that provide automatic variable selection. This method eliminates variables from the model using cross validation methods to reduce error. This model forces the direction of the coefficients for each variable and drops a variable from the analysis if the coefficient is in the wrong direction. The coefficients for prior crime and total population are forced to be positive, while the coefficient for the SES variable is forced to be negative. The software automatically explores how strongly the LASSO penalty should be used in relationship to the ridge penalty – the two components that make up the elastic net penalty.

## Smoothed outcome

The regression models can be generated with a spatially smoothed version of the outcome variable. By default, the script will not spatially smooth the outcome variable (future crime) in the model calibration stage. This parameter can be set to a value of 2 so spatial smoothing will be done with 2 nearest neighbors. Setting the value to 3 will smooth values with 3 nearest neighbors, and so on.




The short-term crime indicator is based on the near-repeat phenomenon. Research on near-repeat victimization has found that in the aftermath of a burglary, nearby houses are temporarily at a heighted risk of being burgled. There is an elevated risk of burglary to nearby locations for a few weeks and for a distance of usually a few hundred feet. This crime pattern has been found to also exist for violent crime.

The PROVE utility uses a Monte Carlo simulation to identify the near-repeat pattern in 1 year of prior crime data. The analysis is based on Euclidean distances. A .csv file identifying the near repeat pattern is produced. If there is no near repeat pattern in the data, the program will output a blank .csv file and the analysis will stop. The crime prediction software will not generate predictions if there is no near-repeat pattern in the crime data. The optional features in this analysis are listed below.

## Spatial band size

Spatial band size is measured in the map units of the jurisdiction boundary .shp file used (usually feet or meters). This parameter is used to determine the spatial extent of the near-repeat pattern in the dataset. Values for spatial band size should depend on how far the user expects the near-repeat pattern to extend. The default setting is to identify a near repeat pattern in 200 foot bands.

## Spatial band count

This is the number of spatial bands the user would like to analyze. The program will look for the presence of a near-repeat pattern in each band separately. The default option is 4 spatial bands. If the options for the spatial band size and count are left at the default values, the program will look for a near-repeat pattern in four separate 200ft spatial bands. In other words, the program will search for a near-repeat pattern that extends 0-200ft, 200ft-400ft, 400ft-600ft and 600ft-800ft. If

the spatial band count setting were changed from 4 to 5, the program would also analyze data in the 800ft-1000ft band.

## Temporal band size

The temporal band size is measured in days. Values entered for this setting should depend on how long the user expects the near-repeat pattern to persist. The default setting is 7 days.

## Temporal band count

The temporal band count is the number of time periods the user would like to analyze. The program will look for the presence of a near-repeat pattern in each temporal band separately. The default option is 4 temporal bands. If the options for the temporal band size and count are left at their default values, the program will look for a near-repeat pattern in 4 separate 7 day periods. In other words, the program will search for a near-repeat pattern that extends 0-7 days, 7-14 days, 14-21 days and 21-28 days. If the temporal band count setting were changed from 4 to 5, the program would also analyze data in the 28-35 day time period.

## Simulation counts

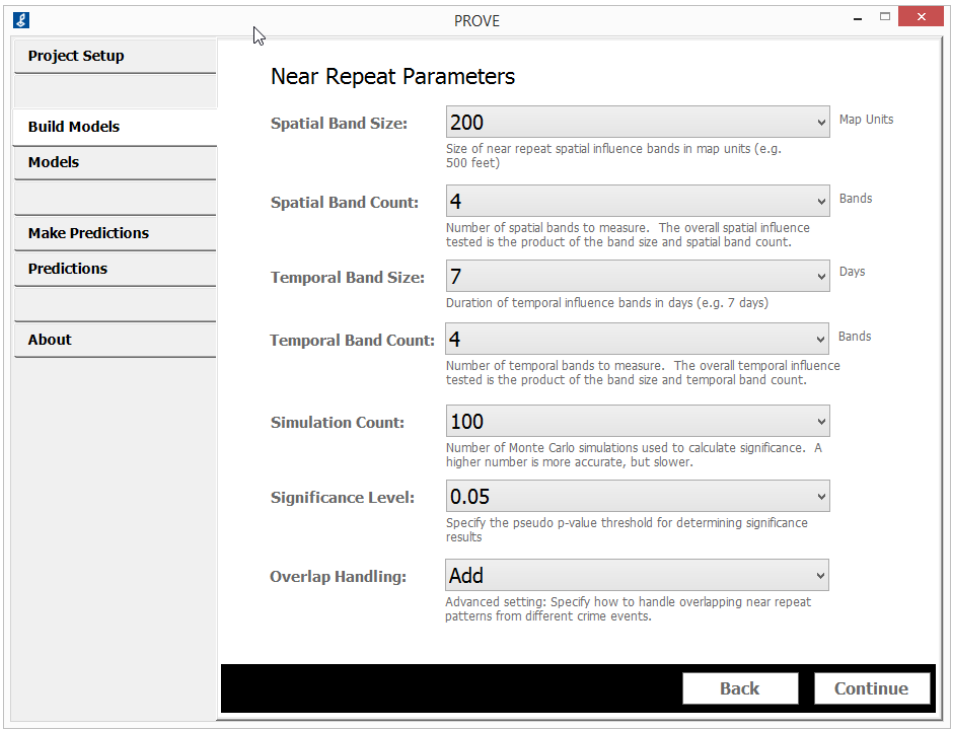
Near-repeat patterns are identified with the use of Monte Carlo simulations. With each simulation, the crime dates are randomized while the locations of the crime incidents remain stable. The default setting is to randomize the dates 100 times.

## Significance level

The significance level for the near-repeat pattern can also be adjusted. The significance value determines if an odds ratio is statistically meaningful or if it is likely due to chance. The default value is 0.05. Only statistically significant odds ratios will print to the output .csv file. Using the default option, statistically significant is defined as p values that are .05 or smaller.

## Overlap handling

Sometimes, the odds ratios that are generated in the near-repeat analysis will overlap. This will happen if two crime incidents are close together. By default, the program will *add* the odds ratios from these two crime incidents when assigning the cell values. This can be changed so the largest odds ratio value is selected (max) or it can be changed so both values are multiplied together.



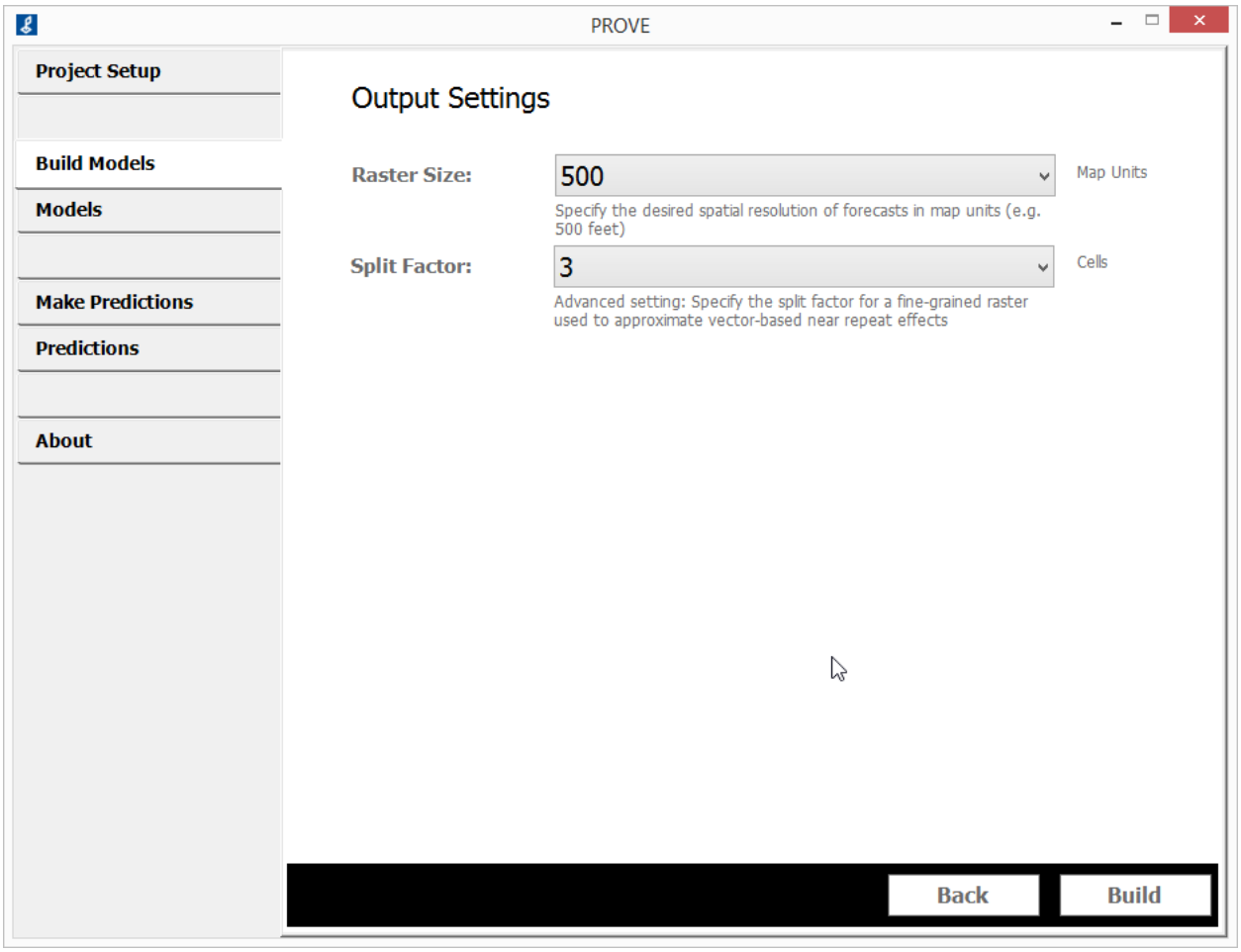

Once the settings for the near-repeat analysis have been set, the user can change the resolution of the output and the raster split factor.

## Raster size

By default, the program will generate crime predictions in 500 unit by 500 unit cells. The unit will match the jurisdiction shapefile. So if the jurisdiction file is in feet, the output file will also be in feet.

## Split factor

The PROVE software conducts the analysis at a fine grained unit of analysis that is smaller than the specified raster cell size. The split factor option allows the user to change size of the fine grained units that are used to conduct the analysis. The default option is to divide each raster cell into 9 smaller cells. In other words, the default is to split the large cells by a factor of 3 (3x3=9). A split factor of 2 would divide each large cell into 4 smaller cells (2x2=4). A split factor of 4 would divide each large cell into 16 smaller cells (4x4=16). Increasing the split factor will increase the precision of the analysis, but it will also increase processing time.

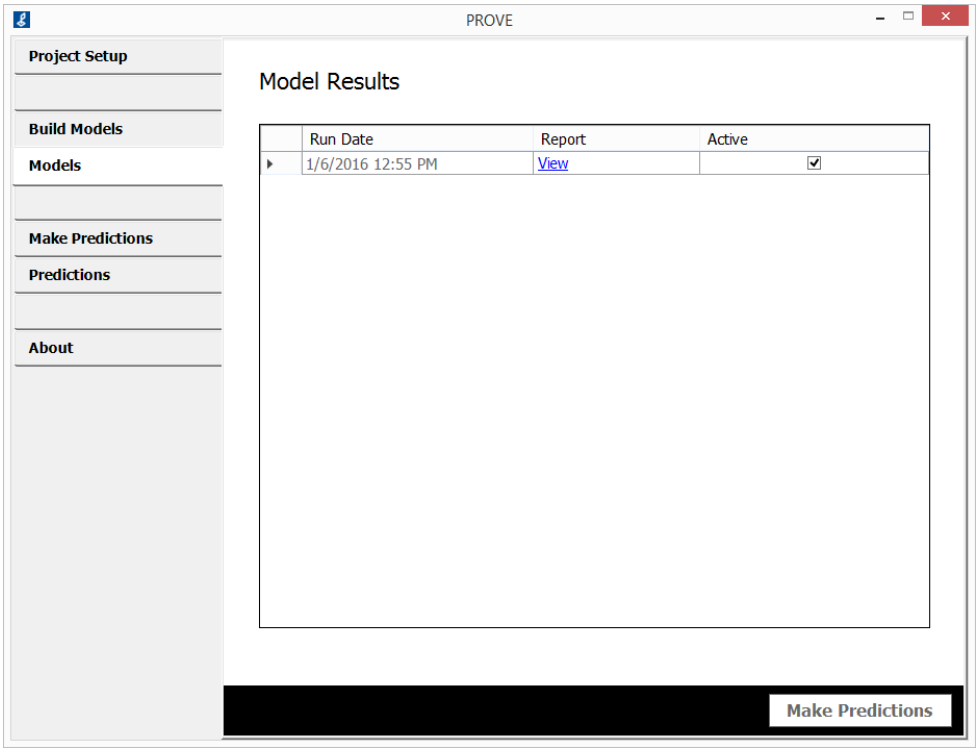



## Build models

After specifying the parameters for the long-term and short-term crime indicators the model calibration stage will begin. The PROVE utility will download all of the census data, generate the demographic variables, create the long-term crime indicator and identify the near-repeat pattern in the data. This process is computationally intensive and can take up to 120 minutes to run if the crime data in the .csv file contain a large number of crime incidents. This stage of the analysis, however, will only need to be run once a year when new census data are released and new crime data are available.

Once the model building stage of the analysis is complete, a results window will open. In this window there is a link where you can view the results of the analyses. If you run multiple analyses in the same project the results from each analysis will be listed in the order they were run.

Using the results from the model building analysis crime predictions can be generated using the most recent crime data available. Clicking on Make Predictions will begin this process.

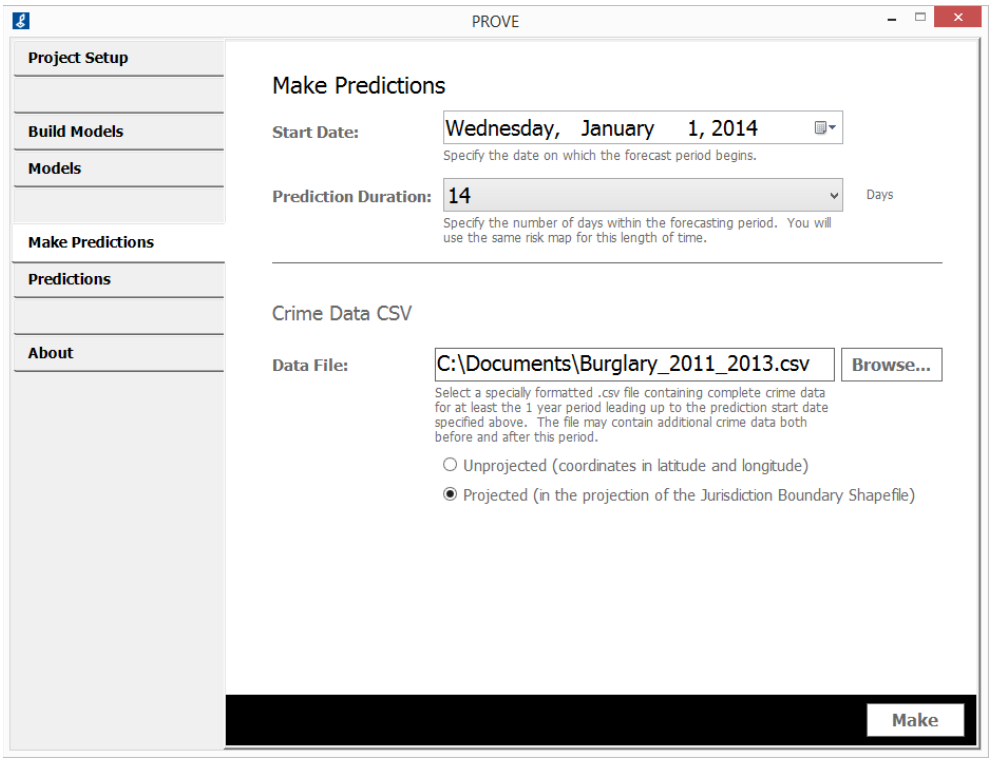



## Generating a prediction

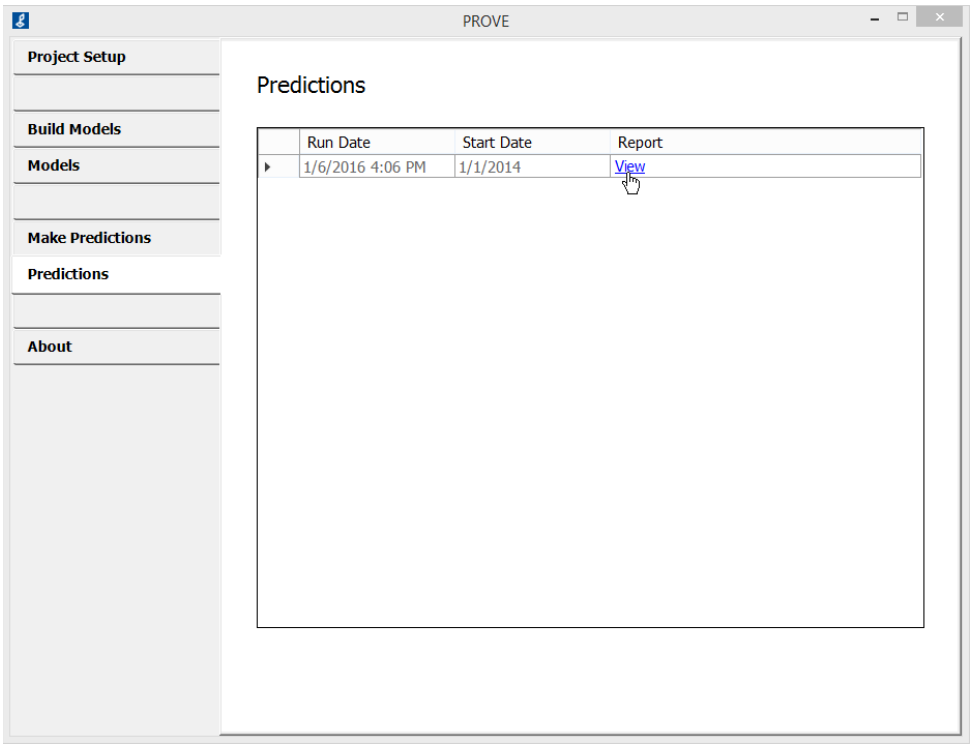
The PROVE software will use the results from the model calibration stage to generate predicted crime hotspot locations. To make these predictions, the user must specify the prediction window. Select the start date and duration of the prediction window. By default, the utility will generate a prediction for the first 14 days of January. The start date can be changed to any time period in the prediction year. The duration of the prediction can also be changed to 7, 14, 21 or 28 days.

Next, the user must upload crime data in .csv format for 1 full year prior to the start date. For example, if the prediction start date is May 17, 2016, crime data for May 15, 2015 - May 16, 2016 must be available.

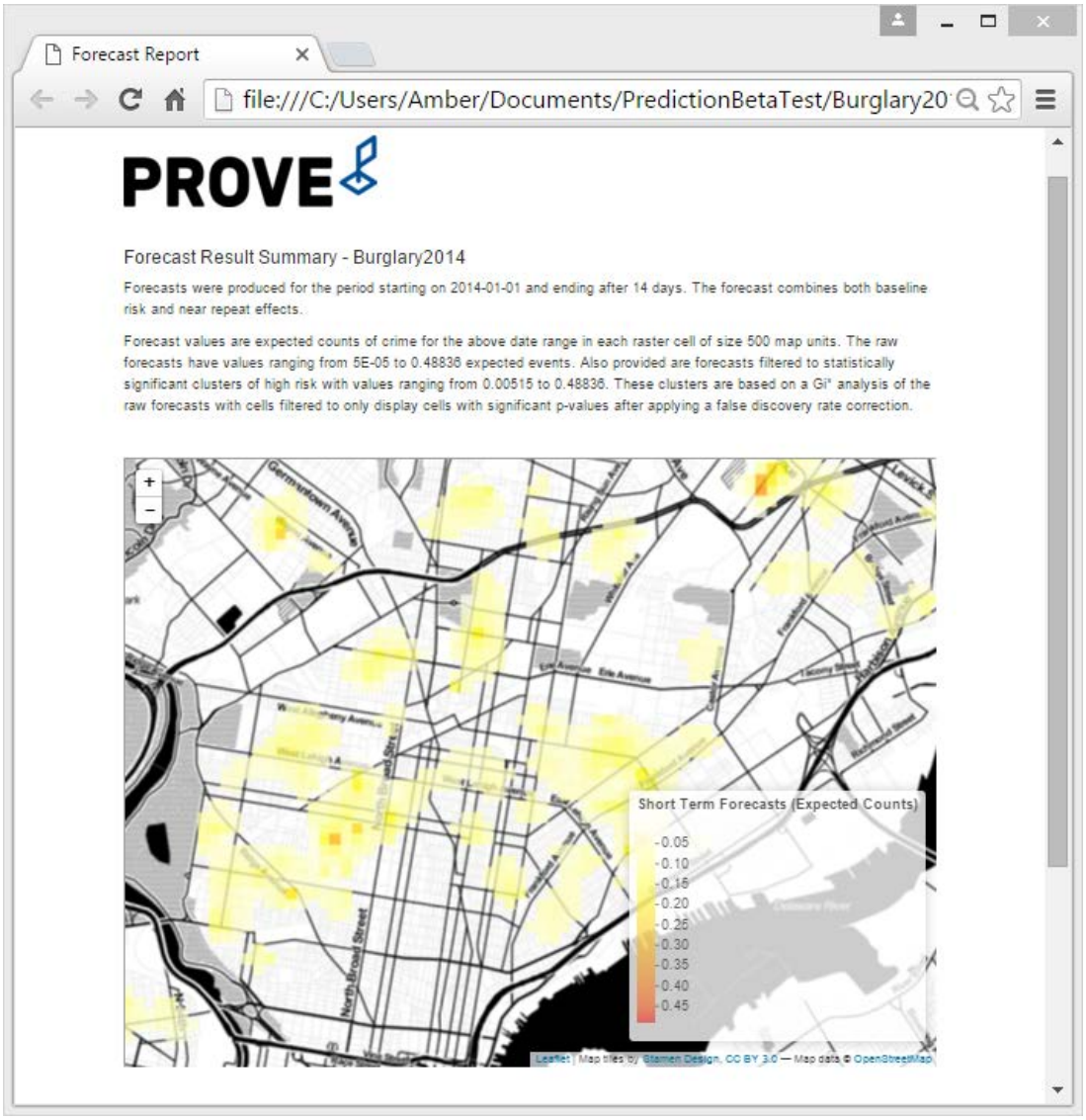
The last step is to specify if the crime data in the .csv file are projected or unprojected. If the data are projected, they must match the projection of the shapefile that was uploaded in the model building stage. If the data coordinates are in latitude and longitude, select the unprojected option.




Once the analysis is complete, a Predictions window will open. This window contains a hyperlink that will allow you to view the results. Click view to view the results of the analysis.



The results of the analysis will be depicted on a map. Locations in red reflect locations that have the highest probability of experiencing a crime for the prediction window that was specified. You can zoom in or out of this map to change the resolution of the image.



# Appendix A

Census data

The PROVE software uses census data and crime data to generate a long-term crime indicator. Census data are pulled from the American Community Survey's (ACS) online database. The ACS is administered every year by the U.S. Census Bureau. The data represent estimates based on a sample of the population. These surveys collect a variety of information including the age, race, income, education and careers of survey respondents. Data are published every year for counties with populations of 65,000 people or more, every three years for populations of 20,000 people or more and every five years at the census block group level. The data that are used in the crime prediction scripts are from the five year data release and are at the census block group unit of analysis.

The variables that are pulled from the ACS to generate demographic variables are listed below:

B19001002—Households income less than $10,000
B19001003—Household income $10,000-$15,000
B1900100—Household income $15,000-$20,000
B19001011—Household income $20,000-$50,000
B19001012—Household income $50,000-$60,000
B19001013—Household income $60,000-$75,000
B19001014—Household income $75,000-$100,000
B19001015—Household income $100,000-$125,000
B19001016—Household income $125,000-$150,000
B19001017—Household income greater than $200,000
B19001001—Total population of households
B25077001—Median home value
B19013001—Median income
B03002003—White non-Hispanic population
B03002001—Total population

# Appendix B

Long-term crime indicator construction

The long-term crime indicator is calculated using two years of prior crime data and two years of census data. Two years of data are needed so a model can be built with one year and applied to the next year.

Two census variables, a logarithmic transformed version of the total population and an index measuring socio-economic status, are used as independent variables in the regression model. The utility uses a negative binomial regression model to generate predicted counts for the next year. Two regression models are run in this process. First, the program uses crime and demographic data from two years prior to build a test model. Then it uses the B weights from that model and applies the most recent crime and demographic data into the equation.

So for example, if you want to generate a long-term crime indicator for 2013, you need the following data sources.

crime data for 2011 and 2012
census data for 2011 and 2012[1]

First, the program will use 2011 crime and 2011 demographic variables in a regression model to predict 2012 crime. This equation can be seen below:

$$crime_{2012} = \alpha + \beta_1(SES_{2011}) + \beta_2(logpopulation_{2011}) + \beta_3(crime_{2011}) + \varepsilon$$

The B weights generated by this model are then rolled forward and applied to the next year of available crime and demographic data. So if the output from the first regression model looks like this:

$$\alpha = 1.2$$
$$\beta_1 = -2.3$$
$$\beta_2 = 8.2$$
$$\beta_3 = 3.9$$

then the predicted crime counts for the target year (in this case 2013) would be calculated using the most recent crime and census data.

$$PredictedCounts_{2013} = 1.2 - 2.3(SES_{2012}) + 8.2(population_{2012}) + 3.9(crime_{2012}) + \varepsilon$$

The default option is to use a negative binomial GLM regression model (no spatial smoothing for the outcome variable) and all three predictor variables (2 demographic variables plus prior crime counts).

[1] Note-there is a 1 year lag between when the census data is collected and when it is publically available. The census data used in this program are ACS 5 year estimates. The data that are available at the end of 2011 were collected from 2006-2010. The data that are available at the end of 2012 were collected from 2007-2011.

# Appendix C

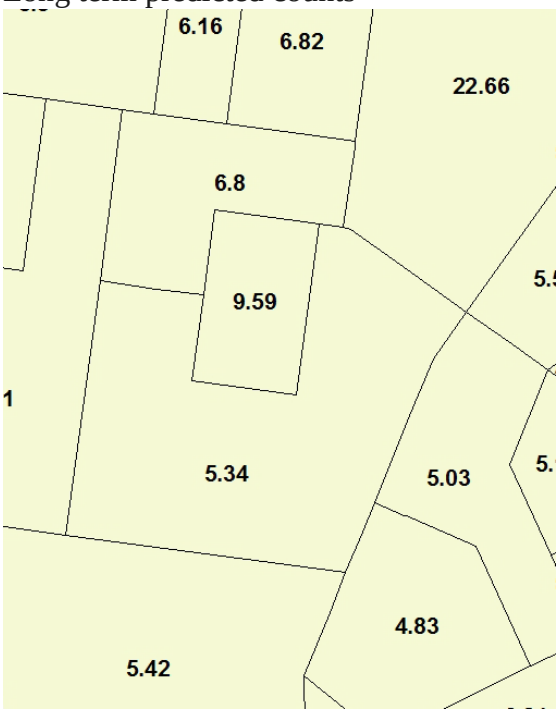
Combining long-term and short-term crime indicators

Long-term values

After the output for the long-term and short-term crime indicators have been separately generated, the software combines them into one file. To do this, both indicators are converted into the same unit of analysis. Before this step, the long-term crime indicator is at the census block group level and the short-term crime indicator is at the grid cell level. To combine the long-term and short-term output, software converts the long-term crime indicator output to grid cells that match the short-term crime indicator. The program will calculate long-term values for a fishnet with a cell size that is defined by the user (this parameter can be changed using the raster split factor option).

To convert long-term predicted counts to cell values, a fine-grained fishnet is created (see Figure 1). The size of the fine-grained fishnet is determined by the raster split factor.

Long term predicted counts




Fishnet overlay

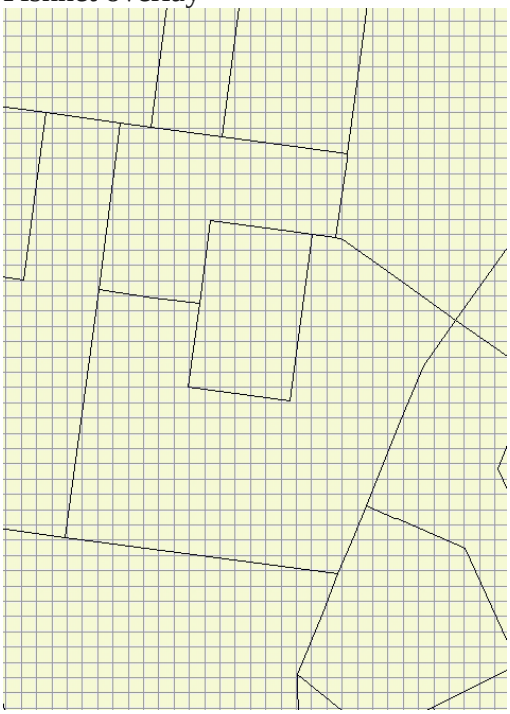


**Figure 1: Census block group long-term predicted counts and fishnet overlay**

The fishnet cell values are calculated with the area of the census block group and the predicted count. The first step is to calculate how many fishnet cells it would take to *perfectly* cover the entire area of the census block group. In Figure 2, the census block group highlighted in green is 742,670.78 sq ft. If the fishnet cells are 100ft x 100ft (1,000 sq ft) it would take 74.267078 cells to *perfectly* cover the area of the census block group.

$$\frac{742{,}670.78\ sq\ ft}{1{,}000\ sq\ ft} = 74.267078\ cells$$

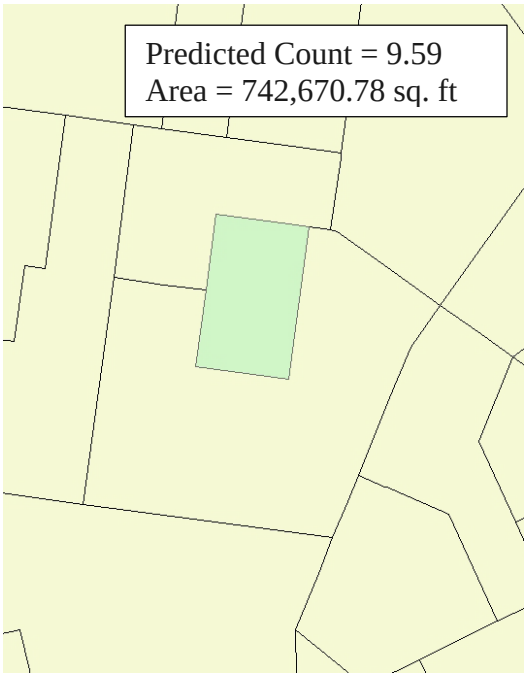



**Figure 2: Census block group**

Next, the predicted count is divided by the number of cells.

$$\frac{9.59}{74.267078} = 0.129$$

A predicted year-long crime count of 0.129 is assigned to each cell whose *centroid* is contained by the census block group boundary. In this example, all of the cells with a green dot will be assigned a value of 0.129 (see Figure 3).

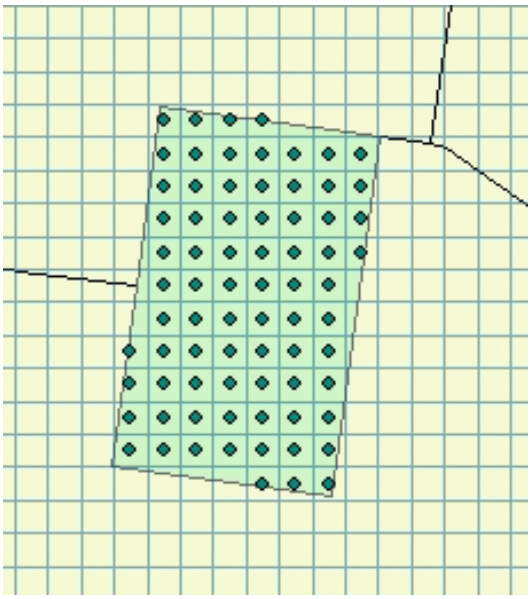


**Figure 3: Cells assigned a value of 1.129**

The last step is to aggregate these fishnet values, whose size is defined by the split factor option, up to the grid cell level. In this example, a split factor of 3 was used, so nine fishnet cells make up one grid cell (3x3=9). For the long-term predicted counts, fishnet values are *added together* to generate the large cell value. For example, in Figure 4, the large cells that contain 9 green dots (green dot = value of 0.129) will all have a long term predicted count of 1.161 crimes.

$$0.129 * 9\ cells = 1.161 \text{ crimes in this grid cell } \textbf{over the course of the year}$$

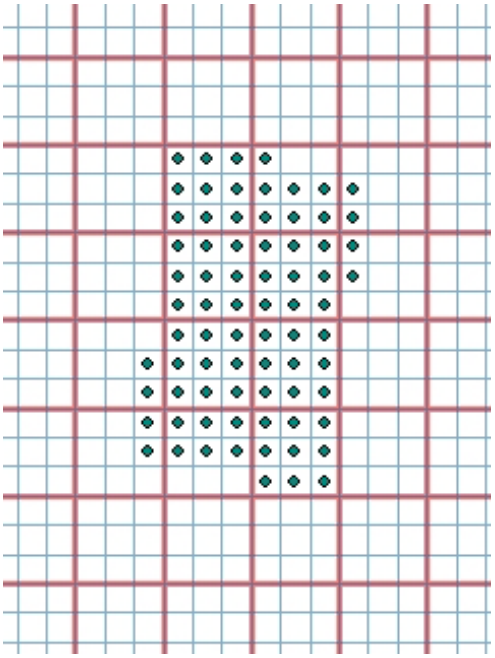


**Figure 4: Aggregation to coarse cells**

The last step in the rasterization of the long-term layer is to divide the long term predicted count by the number of short-term time periods over the course of the year. The default prediction time period is 14 days; since there are 26 bi-weeks over the course of a year, the long term predicted count is divided by 26. If the user chose to change the short-term time period to 7 days, the predicted count would be divided by 52 since there are 52 weeks in a year.

# Appendix D

Generating the short-term crime indicator

Output generated from the near-repeat analysis is also converted to fishnet cells. Odds ratio values are stamped onto the fishnet cells according to the position of the cell centroids. In Figure 5, the red x represents a crime event. The odds ratios that were generated as a result of the near-repeat analysis will be 1.66 for the cells that are closest to the crime event and drop to 1.55, 1.17 or 1.13 as the distance from the event increases.

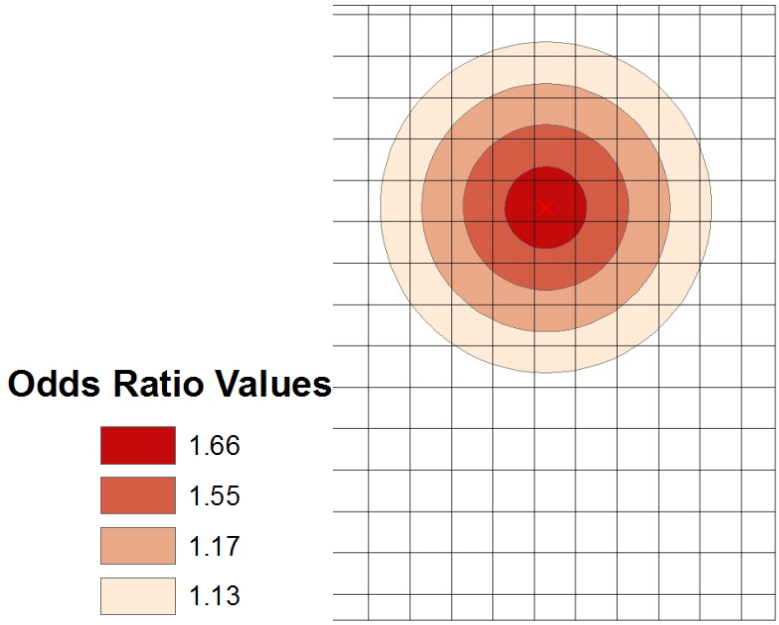



**Figure 5: Risk of near repeat crime decreasing with distance**

Sometimes, the odds ratios from the near repeat analysis will overlap (see Figure 6). By default, the program will *add* the odds ratios from these two crime incidents together when assigning the cell values. A value of 1 is subtracted from each odds ratio before they are added together, then is added back in at the end. In Figure 6, this means if a cell has an odds ratio of 1.55 from crime event #1 and an odds ratio of 1.17 from crime event #2, the cell will be given a value of 1.72.

$$(1.55 - 1) + (1.17 - 1) = 0.72$$

$$0.72 + 1 = 1.72$$

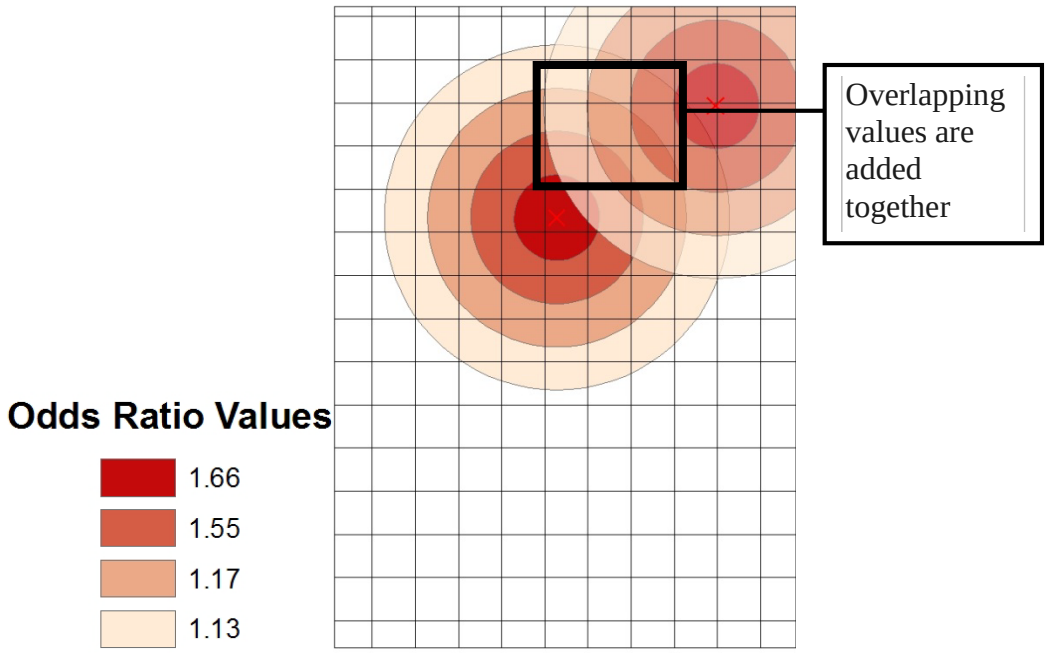



**Figure 6: Assigning risk to overlapping areas**

After the fishnet cell values are assigned, the cells are aggregated up to larger grid cells. This is done by *averaging* all of the fishnet cell odds ratio values. In Figure 7, 9 smaller cells are averaged together to get the large cell values.

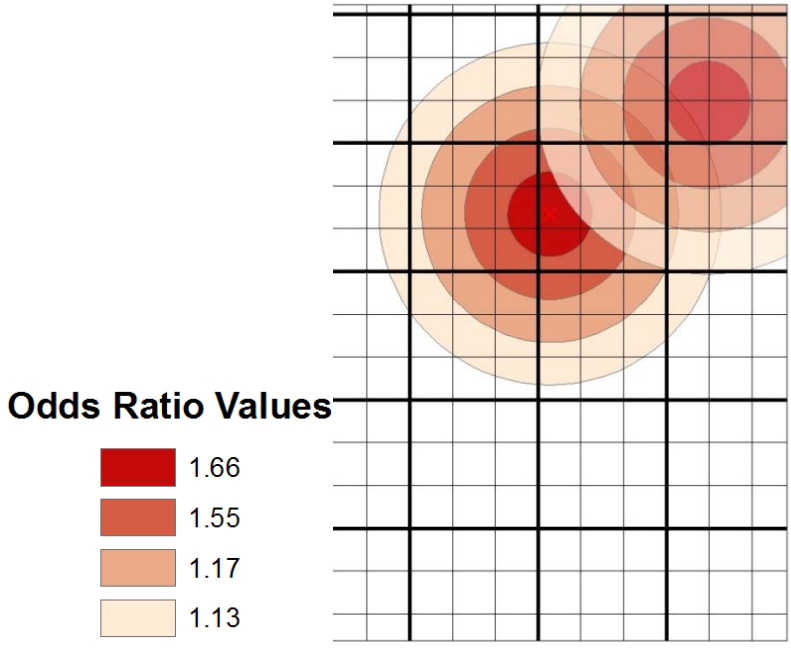



**Figure 7: Aggregating to coarse cells**

Once the data from the long-term prediction and the near-repeat analysis have been converted to grid cell values, the script will output a .csv file that identifies what these values are for each cell

in the dataset. These values will repeat in the .csv file for each short-term time period. By default, each time period is 14 days. In other words, each cell value will update every 14 days based on crime events during the previous time period. Sample output from this process can be seen in Table 1.

**Table 1: Sample csv combining long-term and short-term output**

| startdate | longterm | oddsratio | actualcount |
|---|---|---|---|
| 6/4/2013 | 0.022018 | 1.888889 | 3 |
| 8/27/2013 | 0.020881 | 1 | 3 |
| 1/1/2013 | 0.00212 | 1 | 0 |
| 1/15/2013 | 0.00212 | 1.027778 | 0 |
| 12/3/2013 | 0.013974 | 2.833333 | 0 |
| 11/19/2013 | 0.000679 | 1 | 0 |
| 7/2/2013 | 0.041034 | 1.166667 | 2 |
| 12/3/2013 | 0.015326 | 1.25 | 2 |
| 10/22/2013 | 8.46E-05 | 1 | 0 |

# PROVE Quick start guide

**DRAFT – Please check the PROVE website for the latest version.**

**www.bit.ly/PROVEsoftware**

## Instructions for Setting Up the PROVE Utility

1. Download the PROVE installer. The installer can be found at the following link

   http://s3.azavea.com.s3.amazonaws.com/temp/deleteafter/2016-01/PROVE-Utility-install.exe

2. Run the installer once downloaded
   a. The installer should download and install any necessary packages to use the utility (R and GDAL) if they are not on the computer*
3. When the installation is complete, run the PROVE utility

**Depending on your organization's firewall settings, the utility may not be able to communicate with the necessary sites to download R and GDAL. Please see Appendix A for URLs to whitelist.*

## Preparing Data for the PROVE Utility:

**The PROVE utility requires 2 files to run successfully:**

1. A shapefile of a jurisdiction boundary
2. A CSV file containing crime data that falls within the jurisdiction boundary

**Guidelines for preparing data for use with PROVE:**

- For the jurisdiction shapefile:
  - The jurisdiction shapefile should be one polygon.
  - The polygon may be multi-part and those parts do not necessarily have to be contiguous.
  - Multiple single-part polygons should be dissolved into a single, multi-part jurisdiction polygon.
- For the crime data CSV:
  - PROVE needs at least 3 years of prior crime data to build predictions.
  - Those 3 years of data should be the 3 years immediately prior to the selected prediction year. (If predicting for 2015, data for 2012 through 2014 is required)
  - Prove requires the CSV to have the following fields:
    - id – (number/string) – Unique identifier for the event.
    - class – (string) – Classification for event type. Classifications should be unique and consistent

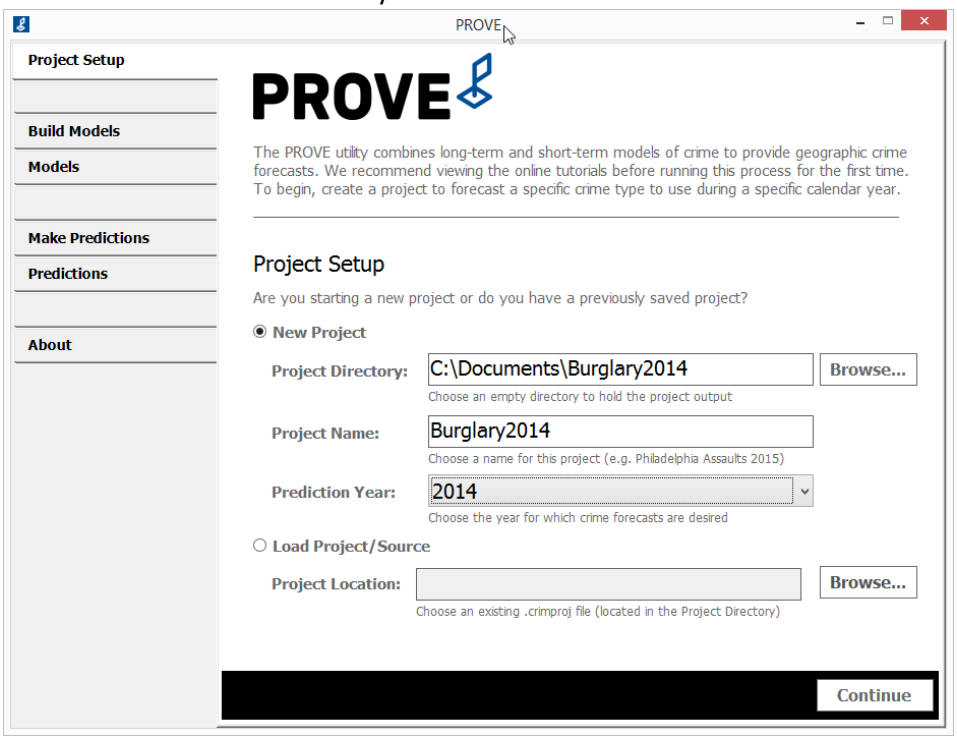
- eventtime – (ISO Datetime) – The date/time value for when the event occurred. This field requires the following format: yyyy-mm-ddThh:mm:ss, where the character 'T' is used to separate date from time. The time should be local time.
- pointx – (number) – The 'x' coordinate for the location of the event.
- pointy – (number) – The 'y' coordinate for the location of the event.

o Any other fields included in the fields are treated as auxiliary. They are preserved through the model building but are ignored.

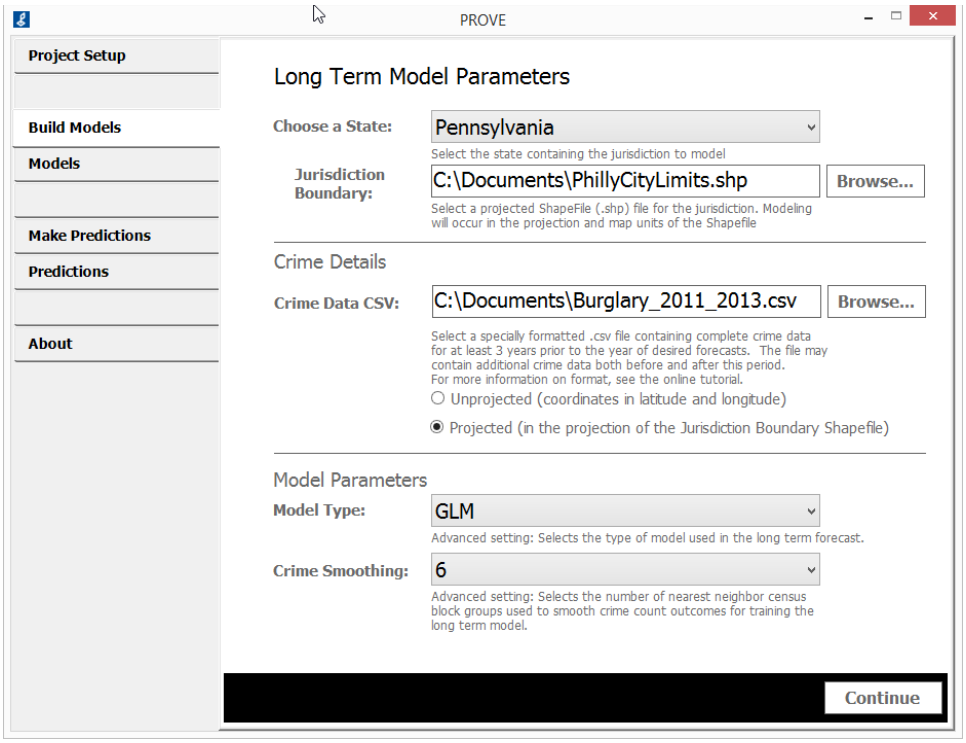
## Running the PROVE utility

The PROVE utility generates crime predictions by combining short-term and long-term crime indicators. First, a model is calibrated by combining short-term and long-term crime indicators for a previous year during a model building stage. Using those calibrated models, metrics are rolled forward one year to create predictions for the desired timeframe.

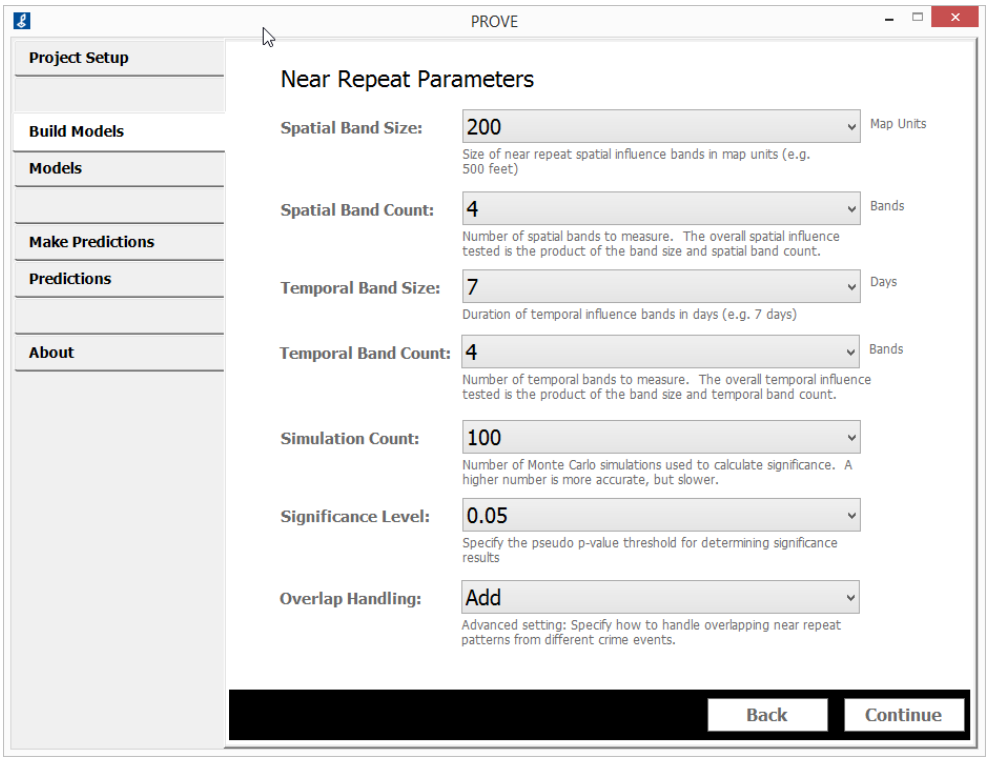
1. When the utility starts, the Project Setup page is displayed. For a new project, select that radio button and fill out the necessary fields.

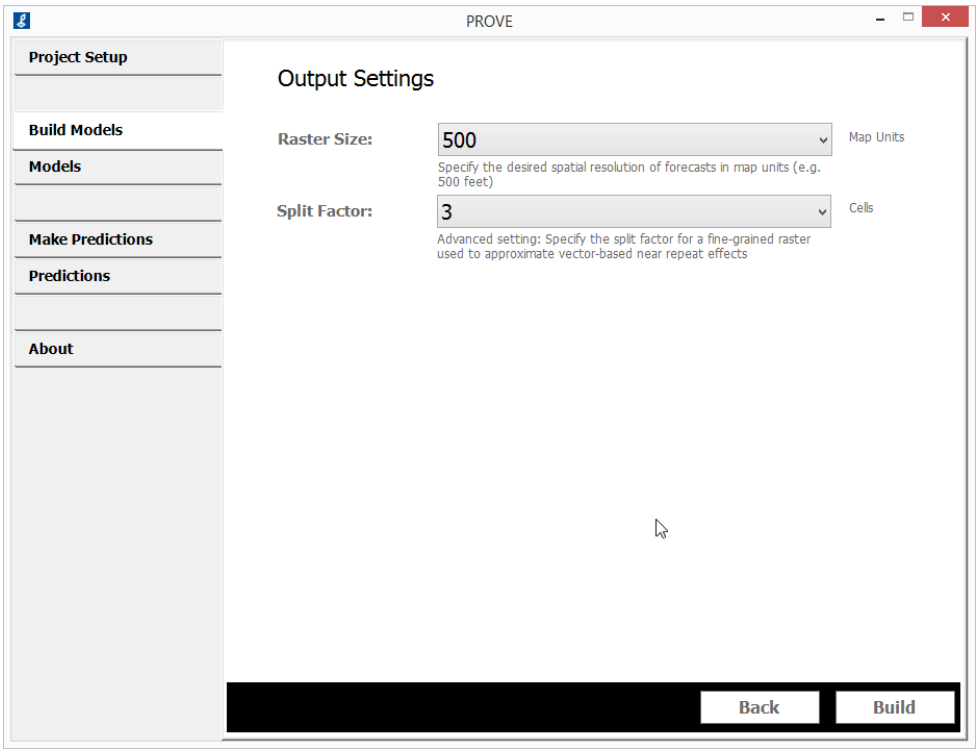



2. Click Continue. Next, the jurisdiction boundary file and crime data will need to be uploaded to the utility. If crime x and y coordinates of the crime data are in latitude and longitude, select the unprojected option. If the data match the projection of the jurisdiction shapefile, select the projected option.

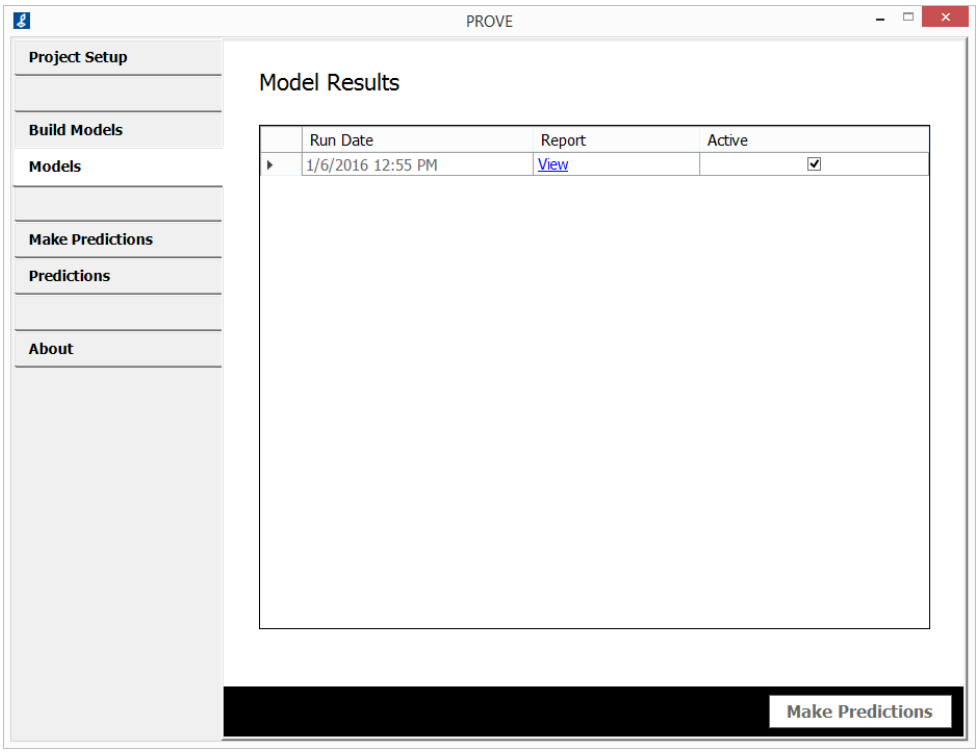



3. The next 2 pages present a number of parameters to be set for the model building. The default options are a good choice for those unfamiliar with the parameters. Advanced users may change these to tweak their models.

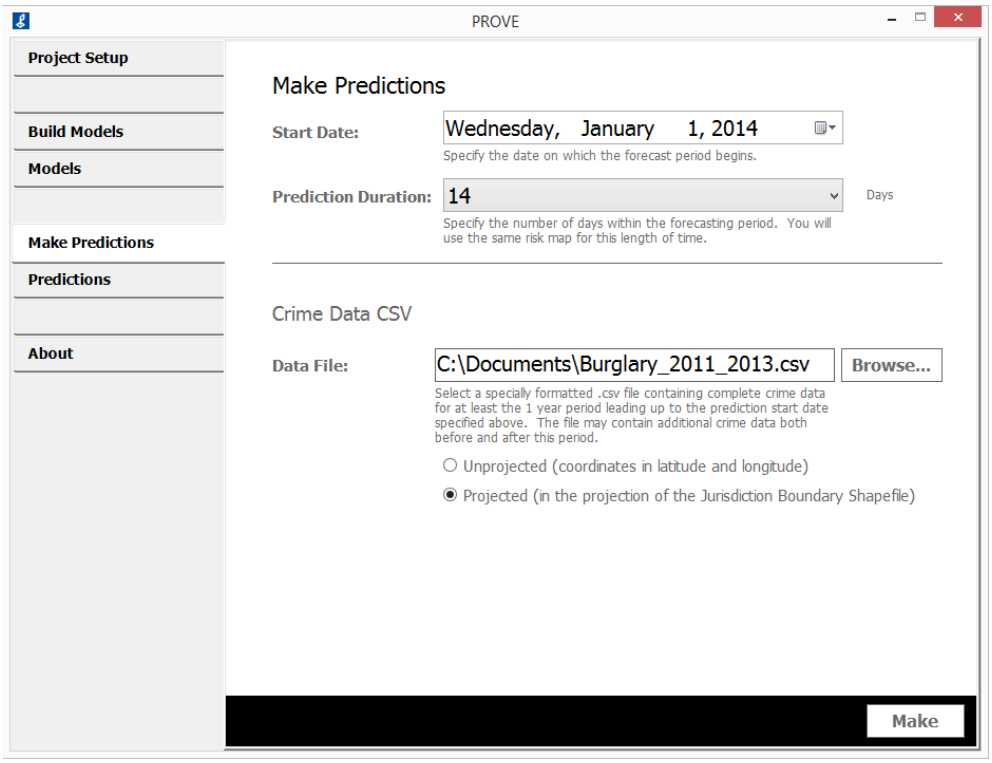




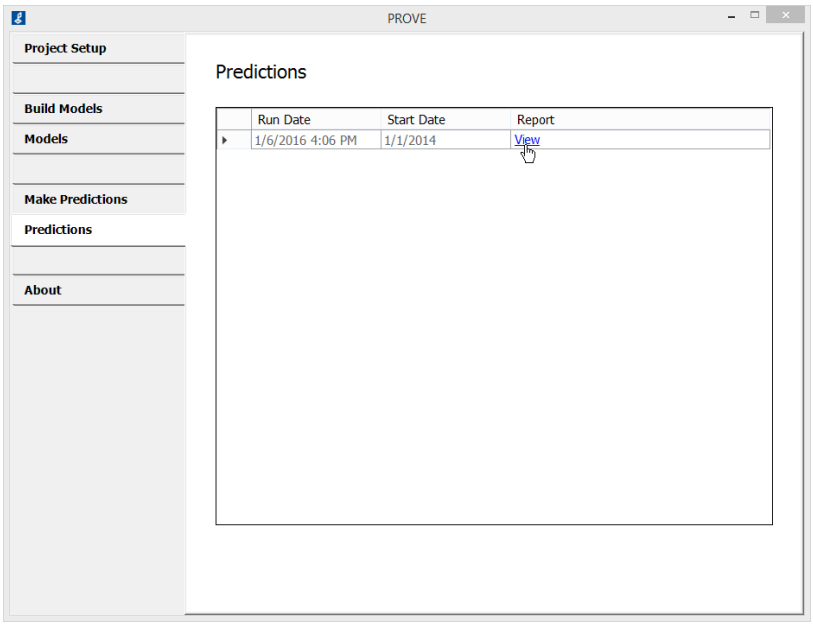

4. At the end of the project setup, click Build Models to start the utility. This part of the program may take up to 120 min to run the first time you use the program, especially if there are a lot of records in your crime data file. Census data are downloaded and a baseline model is calibrated using the crime data provided.
    a. Windows may ask you for permission to run certain packages that are required for the utility to run (such as R). Allow this to allow PROVE to continue running.
    b. If the model building does not run to success, it will indicate `Execution halted` in the log and you will not be able to move on to predictions. Please contact Azavea if this occurs during testing.
    c. The diagnostics for the models that were built during this phase can be viewed by clicking on the "Models" tab.
5. If the model building runs to "Success" then the program was able to successfully calibrate a baseline model. To use this baseline model to generate crime predictions, click "Make Predictions".



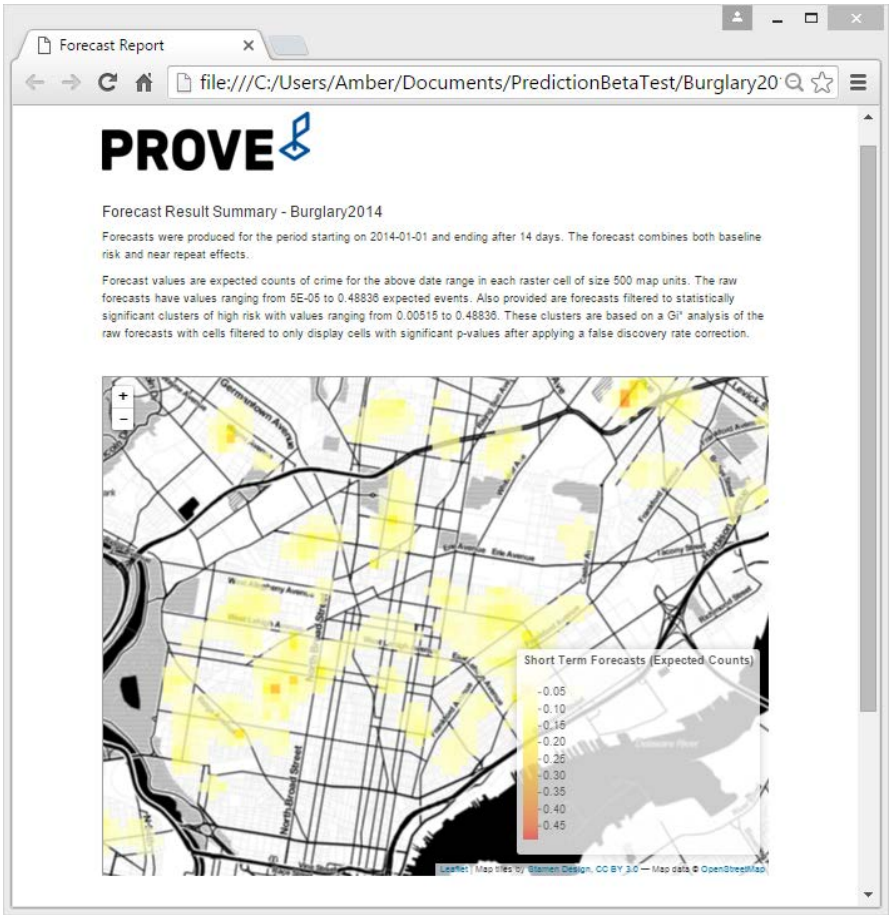

6. Select the date range which you would like a prediction for. Enter another crime csv file that contains a full year of crime data prior to the start date. Again, specify if these data are projected or if they are in latitude and longitude.




7. If the model runs successfully, click “View” to view the prediction.




8. Once the predictions have been made, PROVE will produce a file that can be loaded in any web browser and viewed.



## Appendix A – White-Listed Domains to Enable PROVE Functionality

Many organizations have firewalls set up to limit traffic to sites outside of a secured network. The PROVE utility, however, needs to communicate with certain sites in order to operate properly. For this purpose, some exceptions to the firewall must be made. The purpose of this document is to provide a list of and describe the exceptions necessary to allow PROVE to operate when behind a firewall.

On Port 80 (the port used for http:// requests), the following domains should be white-listed for access through the firewall:

| URL | Port | Purpose |
|---|---|---|
| http://cran.r-project.org/ | 80 (HTTP) | Download R environment and packages |
| http://download.gisinternals.com/ | 80 (HTTP) | Download GDAL toolkit |
| http://a.tile.stamen.com/ | 80 (HTTP) | Base-map tiles for the Black and White Stamen layer |
| http://b.tile.stamen.com/ | | |
| http://c.tile.stamen.com/ | | |
| http://d.tile.stamen.com/ | | |