Received: 31 October 2019 | Revised: 14 June 2020 | Accepted: 16 June 2020

DOI: 10.1111/1745-9133.12514

CRIMINOLOGY & Public Policy

CUTTING-EDGE RESEARCH IN POLICE POLICY AND PRACTICE

Was the pope to blame? Statistical powerlessness and the predictive policing of micro-scale randomized control trials

Ralph B. Taylor | Jerry H. Ratcliffe

Temple University

Correspondence
Ralph B. Taylor, Department of Criminal Justice, Gladfelter Hall, Temple University, 1115 Pollett Walk, Philadelphia, PA 19122.
Email: ralph.taylor@temple.edu

Data analyzed here were collected under Grant 2014-R2-CX-0002 from the National Institute of Justice, Jerry H. Ratcliffe, Principal Investigator, and Ralph B. Taylor, co-Principal Investigator. Any views or opinions expressed herein do not necessarily reflect the official policies of the Department of Justice, the National Institute of Justice, the Philadelphia Police Department, the city of Philadelphia, or Temple University. Authors appreciate helpful tips on pc_simulate provided by Louis Freonas, and insightful comments provided by the editors, reviewers, and Cory Haberman on an earlier version of this article. Portions of this research were presented by the second author at the annual meetings of the American Society of Criminology, November 2019, in San Francisco.

Funding information
National Institute of Justice, Grant/Award Number: 2014-R2-CX-0002

Research Summary: Hinkle et al. (2013) highlighted a statistical powerlessness problem in hot-spots policing experiments in midsized cities with moderate property crime rates. The current work demonstrates that this problem is less readily resolved than previously suspected. It reviews results from a predictive policing randomized control trial in a large city with property crime rates higher than Chicago or Los Angeles. It reports, for the first time, a graphical analysis indicating the marked car patrol intervention, practically effective at the 500′ by 500′ (mission) grid level, three grids per shift, likely had a district-wide impact on reducing reported property crime. In addition, it reviews results of a series of thought experiments exploring statistical power impacts of four modified experimental designs. Only one alternative design, with spatially up-scaled predictive policing mission areas and concomitantly higher property crime prevalence rates, produced acceptable statistical power levels. Implications follow for current theoretical confusion in community criminology about concentration effects and units of analysis, and how models organize impacts across those different units. Implications follow for practice amid ongoing concerns about whether predictive policing works and, if it did, how to gauge its impacts and social justice costs.

Policy Implications: The current work brings to the fore important questions beyond "does predictive policing work?” Can we design predictive policing randomized experiments capable of showing statistical effectiveness? Furthermore, if we can, and if those studies include larger mission areas than the micro-scaled geographic grids used so far, how do we integrate social justice concerns into effectiveness metrics given the broader segments of communities likely affected?

KEYWORDS
community criminology, effectiveness metrics and social justice concerns, hot-spots policing, predictive policing, spatial and temporal scaling

> “There are at present insufficient rigorous empirical studies to draw any firm conclusions about either the efficacy of crime prediction software or the effectiveness of associated police operational tactics. It also remains difficult to distinguish a predictive policing approach from hot spots policing” (National Academies of Sciences Engineering and Medicine, 2018, p. S-4)

This work builds on the results from a micro-scaled randomized control Predictive Policing Experiment that worked practically, but not statistically, with graphically demonstrable district-wide crime reduction impacts. Four thought experiments pose the following question: Can statistical powerlessness due to the rarity of Part I property crimes in micro-scaled locations during specific time windows, in locations predicted to be the most property crime prone in their respective districts, in a city with Part I property crime rates exceeding those of Chicago and Los Angeles, when those rates are addressed with a demonstrably effective treatment, be surmounted with alternative experimental research designs? Stated differently, what alternative hypothetical experimental designs could have coped with the impaired statistical power associated with the extremely low-property-crime prevalence rates at the micro-time-and-place-scale of a predictive policing intervention?

Implications follow for theory, policy, and practice. In brief, they are as follows. For theory, the current focus on extremely small crime intervention sites represents the culmination of half a century of drilling down below the neighborhood level to examine, predict, and ultimately understand variations in crime patterns and levels at micro-spatiotemporal scales. The current work on near-repeat effects (Bernasco, 2008; Bowers & Johnson, 2004; Johnson & Bowers, 2004; Ratcliffe & Rengert, 2008; Townsley, Homel, & Chaseling, 2003) and spatial concentration effects (Eck, Lee, O, & Martinez, 2017; Lee, Eck, O, & Martinez, 2017; Weisburd, 2015) attests to some threads of that drilling down. As yet, however, nobody has made a convincing case that, for a particular crime type, a *specific* geo-scale or even a specific narrow range of geo-scales matches up to the relevant micro-spatiotemporal dynamics driving crime occurrences. Therefore, from a theoretical perspective, predictive policing scholars, many of whom focus their research on small grids, may wish to consider shifting to larger spatial frames. Results from specific hypothetical scenarios examined here support such a widening.

For policy, broadening the spatial frame means two things. It means thinking harder about the dynamics behind localized crime escalation patterns, perhaps through collaborating more closely with agencies and key stakeholders who can shed light on features driving these localized intensification patterns. For example, this may mean integrating concepts from third-party policing (Mazerolle & Ransley, 2005) with predictive policing.[1] Second, if larger areas are subsumed, since larger resident populations or volumes of pedestrian traffic could come under greater scrutiny, this requires even more closely considering potential adverse side effects. These could manifest in terms of racial, ethnic, or class inequities, net widening, or deepening concerns about criminal justice agencies acting in institutionally de-legitimizing ways in these more generalized and aggregate environments (Berk, Heidari, Jabbari, Kearns, & Roth, 2018; Richardson, Schultz, & Crawford, 2019; Shapiro, 2017; Tyler, Fagan, & Geller, 2014).

For practice, it means police analysts and leaders, as well as involved community leaders, need to carefully consider the costs, benefits, and cost effectiveness of two different practices (Rummwens & Hardyn, 2020): a focus on small individual spots, in the form of street corners (Lawton, Taylor, & Luongo, 2005) or individual problematic addresses (Frisbie et al., 1978; Mazerolle, Kadleck, & Roehl, 1998; Sherman, 1989); weighed against spatially expanding the microscopic scale of predictive policing analytics. We return to this practice question in the discussion.

At the outset we acknowledge other scholars have already highlighted statistical powerlessness problems in micro-scaled studies of hot-spots policing (Hinkle, Weisburd, Famega, & Ready, 2013). Here, however, this discussion is extended in three ways. First, points raised earlier apply as well to predictive policing where hot spots get updated on a daily basis and therefore can move around. Second, the scope of the problem is broader than suspected. We demonstrate that the stated concern applies not just in smaller cities with lower crime rates but also in the country's biggest cities with the highest big-city property crime rates. Furthermore, the current work in effect follows up on one of Hinkle et al.'s (2013) proposed solutions, expanding study sites. Based on the results shown here, that proposal seems less effective than anticipated at solving the statistical powerlessness problem.

The remainder of the article is organized as follows. The progression down the cone of resolution traveled by community criminology in the last four plus decades is noted. This shift, and accompanying transformations in policing patrol prevention research, has led to an unresolved state of affairs. Turning to recent concentration of crime at places work, we ask the following: Does it provide resolution to questions of spatial units and crime concentration? The answer is "yes and no." Yes, the units that most spatially concentrate crime can be identified, but no, predictive policing experiments cannot be organized around those units. Switching to hot-spots policing, and its tech-savvy younger sibling, predictive policing, it seems that the unresolved theoretical state of affairs regarding spatial scaling in community criminology has similarly afflicted this realm.

Bringing the lens closer, reported property crime levels in Philadelphia are noted, and key results from a recent randomized control trial in predictive policing that took place there are sketched. The practical impacts in targeted mission areas, as well as the graphical analysis of district-wide impacts for the most successful treatment, the latter being reported here for the first time, are both reviewed. Four thought experiments exploring impacts of research design on statistical power are outlined. The results then turn to whether, and if so how, each thought experiment successfully resolved the statistical powerlessness problem plaguing the Philadelphia Predictive Policing Experiment. Discussion returns to implications for theory, policy, practice, and social justice concerns.

For the time-stressed reader, here are the main takeaway ideas. Predictive policing experiment researchers may have to spatially scale up the size of the mission areas examined if they want to demonstrate statistical as well as practical crime prevention effectiveness. Theoretical models are needed linking address crime concentration dynamics and crime concentration patterns in scaled-up mission areas, thereby explaining emergent properties at the latter level. Furthermore, if larger zones get more predictive policing attention, policy makers crafting comprehensive predictive policing effectiveness metrics must figure out not only costs and benefits (Rummens & Hardyn, 2020) but also trade-offs with social justice, and how to weigh all these concerns simultaneously. Of course all of these concerns are subsidiary to the broader goal of police working with crime data and, as importantly, collaborating with community stakeholders to identify key problem crimes and key problem locations, and then co-crafting with those stakeholders effective, cost-effective, socially just and acceptable solutions.

1 | GET TINY, BUT WHICH TINY?

More than 40 years ago, economist E.F. Schumacher (1975) considered the role of spatial scale in human affairs, ultimately arguing for bigness in vision and smallness in action. Just a year later, Brantingham, Dyreson, and Brantinghm (1976) moved down a "cone of resolution," demonstrating a particular effect of scale as one moves from "bigness" to "smallness," to use Schumacher's (1975) terminology. As one moves from states through progressively smaller spatial scales to census blocks, an area that appears high on a crime indicator resolves, as one moves to the next smaller scale, into geographies with nonuniform high- and low-rate areas. They observed that at *every* level "crime occurrence is not uniformly distributed across space but rather clusters into clear regions of high and low rates of occurrence" (Brantingham et al., 1976, p. 265). Crime clusters, regardless of the scale of resolution.

Numerous community criminology studies, focusing toward the lower, more micro-level range of spatial scales—hot spots, streetblocks, or corners—have described these patterns and sometimes have explained the dynamics behind them, relying on a broad range of theoretical frames including crime pattern theory, situational crime prevention, gang set space, facilities/land use, routine activities theory in different variants, and rational offender perspectives (Block & Block, 1995; Braga, 2001; Braga, Hureau, & Papachristos, 2011; Brantingham & Brantingham, 1995, 1999; Clarke & Eck, 2007; Eck & Weisburd, 1995; Eck, Chainey, Cameron, Leitner, & Wilson, 2005; Groff & McCord, 2011; Loukaitou-Sideris, 1999; Maltz, 1995; Ratcliffe, 2012; Roberts, Taylor, Garcia, & Perenzin, 2014; Roncek & Maier, 1991; Sherman, Gartin, & Buerger, 1989; Spelman, 1995; Taylor, 1997; Taylor, Gottfredson, & Brower, 1984; Van Patten, McKeldin-Coner, & Cox, 2009; Weisburd & Mazerolle, 2000; Weisburd, Groff, & Yang, 2012; Weisburd, Morris, & Groff, 2009; Yang, 2010). Micro-spatial units at only three levels—streetblocks (aka "street segments"), framed using Hawleyesque micro-ecological principles or broader Durkheimian ideas (Roberts et al., 2014; Taylor, 1997; Weisburd et al., 2012); individual streetcorners, framed using marketing and/or competitive group (gang) dynamics including ordered segmentation (Lawton et al., 2005; Simon & Burns, 1997; St. Jean, 2007; Suttles, 1968; Taniguchi, Ratcliffe, & Taylor, 2011; Thrasher, 1926, 1927; Tita & Ridgeway, 2007; Whyte, 1943); and individual facilities/addresses, framed using crime pattern theory's crime attractor/generator constructs (Brantingham & Brantingham, 1995, 1999, 2008; Jennings et al., 2013; Kinney, Brantingham, Wuschke, Kirk, & Brantingham, 2008)—have clearly identified theoretical dynamics operating at the same spatial scale as the observed spatial units.

That said, in all three instances the relevant theoretical dynamics specifically incorporate not only the observed spatial units but also their surrounding environment, allowing for both contextual and emergent dynamics (Beavon, Brantingham, & Brantingham, 1994; Brantingham &

Brantingham, 1993; Brantingham, Glässer, Jackson, & Vajihollahi, 2009; Kinney et al., 2008; Suttles, 1968; Taylor, 1997, 2015, pp. 136–137). So in these three instances, even though the key spatial units are delineated, dynamics at more than one spatial scale are operating and impacting different levels. Consequently, the search for one foundational spatial unit for crime analysis and prevention at the micro-scale founders or at least gets blurred.[2]

Accordingly, to help narrow the range of spatial scaling options for investigation, researchers have turned in a different, radically empirical direction, comparing and contrasting spatial inequalities arising from concentration effects associated with different-sized spatial units.[3] Other work has shown that crime is concentrated at the levels of streetblocks (Braga et al., 2011; Weisburd, 2015), corners (Braga et al., 2011), and individual addresses (Eck et al., 2017; Frisbie et al., 1978; Sherman et al., 1989). At each level, because small fractions of these spatial units account for reported crime volumes far in excess of the percentage of space involved, one can say that the reported crime in question is concentrated within a small fraction of spatial units at that scale. Concentrations of a large fraction of offenses within a much smaller fraction of offenders, a large fraction of delinquent acts in a small fraction of delinquents (Wolfgang, 1983, p. 84), or a large fraction of victimizations within a much smaller fraction of victims have been similarly observed (Eck et al., 2017). Such concentration effects are not particular to crime or victimization dynamics. They appear similarly in nature (Eck et al., 2017; Monmonier, 2008), demands for emergency services (Gawande, 2011), and many other areas. As a result, theoretical, empirical, and practice questions surface.

Theoretically, for each crime in question, is there one micro-scale spatial unit that we could recommend because it has a real-world counterpart and its crime dynamics have been clarified? Three potential candidates, commendable in no small part because they exist as physical entities, are streetblocks (aka "street segments"), street corners, and individual addresses or facilities.

Much is understood about streetblock disorder- and crime-linked dynamics (Roberts et al., 2014; Taylor, 1997, 2015; Weisburd et al., 2012). But, streetblocks are disqualified as the *only* foundational spatial unit for understanding these dynamics for multiple reasons (Taylor, 2015, pp. 134–135). Significant crime happens between them, on street corners (Braga et al., 2011), rather than in them. Furthermore, "streetblocks can demonstrate internal [spatial] differentiation in their crime patterns" (Taylor, 2015, p. 134; see also Roberts et al., 2014).

Turning to corners, on the theoretical plus side much is understood about spatially congruent socio-spatial dynamics linked to corners (Liebow, 1967; Suttles, 1968; Thrasher, 1927; Whyte, 1943), especially those dynamics linked to drug sales and related crimes (Bourgois, 1996; Lawton et al., 2005; Simon & Burns, 1997). But a focus on corners leaves out crime happening away from them. This concern can be addressed by spatially expanding corners using, for example, Thiessen polygons (Taniguchi et al., 2011). Doing so, however, obscures how different groups claim different bits of neighborhood spaces (Suttles, 1968). In addition, the spatial expansion creates abstract spatial units that now combine disparate social groups from the different streetblocks leading away from the corner.

Individual addresses/parcels/facilities, concepts from crime pattern theory including nodes, crime attractors, and crime generators, do at least clarify spatially congruent dynamics (Taylor, 2015, pp. 135–136). Theoretical concerns, however, persist. Most notably, crime events can develop at a location but come to fruition a distance away. Distance decay functions of violent crime and bars, empirically documented more recently (Ratcliffe, 2012), and going back to the 1970s and Moby Dick's bar in Minneapolis (Frisbie et al., 1978) where you could "Get a whale of a drink," underscore the spread problem that can originate with a single problematic address.

Put aside for a moment questions of effective and cost-effective crime prevention and control practices, as well as associated questions of predictability. Empirically, unit selection is simply a matter of (a) are there different amounts of crime concentration at various spatial scales? and (b) if those spatial concentration differentials are observed, which spatial unit provides the strongest degree of spatial crime concentration? In other words, what is the most efficient scale with maximum crime concentration so enforcement would be effective but also yield minimal collateral effects? Although this is an empirical question, the theoretical stake are high. Different levels of concentration at different spatial scales "implies that there are different processes at each level, or that there is some form of hierarchical arrangement where higher level contexts help shape the outcomes of lower level processes (e.g., street segments provide a context that moderates the address level dynamics of crime)" (Eck et al., 2017, p. 3). Stated differently, there are two possible situations. Observing different concentrations at different levels supports the assumption of theoretical discontinuity (different crime-generating processes at different levels). Or, theoretical homology (same crime-generating processes) which leads to observing similar concentrations of crime across different spatial scales (Taylor, 2015, pp. 94–96), If one takes a strict meso-level focus considering just one spatial scale like a streetblock where there is some degree of spatial crime concentration, one necessarily overlooks spatially dependent contextual impacts (Taylor, 2015, pp. 106–118).

Furthermore, not only does it matter *whether* crime is more concentrated at some spatial scales than others, but also it matters whether, if relative concentration varies, the *direction* of the relative differences. "If crime is more concentrated as one examines smaller units, this implies that one *should build explanations from the bottom up*. The value of the larger units is that they can provide contexts for processes occurring in smaller units" (Eck et al., 2017, p. 3, emphasis added). Such spatial contextual dependencies, for streetblocks, were formalized in streetblock microecological principle 3: "Block life is conditioned by features of adjoining blocks" (Taylor, 1997, p. 134). Wilcox and colleagues (2003) stated this point more broadly.

So, what do empirical patterns show? Eck et al. (2017), using Cincinnati data and considering crime-involved as well as crime noninvolved places, found higher levels of spatial concentration at the address level than at the streetblock level and higher levels of spatial concentration at the streetblock level than at the 2500′ × 2500′ grid level. They concluded: "[B]ecause crime is not equally concentrated at different spatial units, this implies that scale matters;" that "it seems unlikely there is a single explanation for crime concentration that covers all scales;" and "we should build explanations from the smallest units—address-level places—upward … understanding the most micro-level processes is fundamental for understanding crime processes in larger area[s]" (Eck et al., 2017, p. 6).

2 | THE PREDICTION TURN

Practical considerations come to the fore when attention turns to prediction and predictive policing because deploying personnel based on predictions costs money, and that resource allocation will prove well spent only if crime in targeted sites at targeted times, or nearby in space and time, consequently declines. Hot-spots policing has been defined as follows:

> Hot spots policing covers a range of police responses, but they all focus resources on locations where crime incidents have been highly concentrated. By focusing on micro-geographic locations with high concentrations of crime hot spots policing aims

> to increase the general deterrence of police actions, in this case by increasing perceptions of the certainty of enforcement action … there may also be a specific deterrent impact of hot spots policing … police can also alter the situational opportunities that exist at hot spots[.] (National Academies of Sciences Engineering and Medicine, 2018, pp. 46–47)

Predictive policing has been variously defined by different scholars. The National Academies (2018, p. 49) report defined it as “a strategy for proactive policing that uses predictive algorithms based on combining different types of data to anticipate where and when crime might occur and to identify patterns among past criminal incidents.” Other definitions are broader; for example, Fitzpatrick, Gorr, and Neill (2019), p. 473) opined that “predictive policing comprises a broad variety of approaches for crime forecasting and prevention.” European scholars Hardyns and Rummens (2018, abstract, p. 201) placed it in the context of intelligence-led policing (Ratcliffe, 2008):

> In the context of crime analysis, the large amount of crime data available can be considered an example of big data, which could inform us about current and upcoming crime trends and patterns. A recent development in the analysis of this kind of data is predictive policing, which uses advanced statistical methods to make the most of these data to gain useable new insights and information, allowing police services to predict and anticipate future crime events.

The relationship between hot-spots policing and predictive policing similarly varies. Some scholars have viewed hot-spots policing, predictive analytics, and predictive policing as clustered under the same umbrella.[4] The National Academies (2018, p. 50) proactive policing report also saw overlap, as well as some distinction: “Predictive policing overlaps with hot spots policing but is generally distinguished by its reliance on sophisticated analytics that are used to predict likelihood of crime incidence within very specific parameters of space and time and for very specific types of crime.” In other words, with predictive policing, hot spots at small spatial scales are updated and potentially relocated using near-term crime inputs. Hardyns and Rummens (2018, p. 204), along similar lines, saw it as a next step beyond hot-spots identification and policing: “Predictive policing can thus be considered a step forward in the crime mapping evolution because of its specific focus on spatiotemporal predictions of crime, thus enabling a more accurate estimation of future crime patterns.” Gorr and Lee’s (2015) early warning system for temporary hot spots, and their distinctions between chronic hot spots, temporary hot spots, and flare-ups, marked a point between hot-spots analysis and policing, and predictive analytics and policing.

In short, although hot-spots prediction and predictive policing link up in multiple ways to hot-spots empirical and conceptual foundations, and crime control centered on such locations, as well as to related strands in intelligence-led policing, big data, and computational criminology, different scholars have characterized the links between hot-spots work and predictive work in different ways. Furthermore, from a practical perspective, and as noted in Footnote 1, with predictive policing of rapidly shifting locations, in-depth problem-oriented policing and community input and coalition building proves much more challenging.

TABLE 1 Property crime prevalence rates and MDEs for different sized hypothetical mission areas

Level of spatial scaling relative to original	Control + Awareness base rate	Proportional reduction (60 %)	MDE (minimal detectable effect)
Original	.04	× .40	– .024
× 5	.201	× .40	–.121
× 10	.397	× .40	– .238
× 15	.592	× .40	–.355

3 | GAUGING EFFECTIVENESS

Gauging the effectiveness of predictive policing requires multiple evaluation metrics. Scholars have differed, however, on how many are needed. The National Academies (National Academies of Sciences Engineering and Medicine, 2018, p. 51) report suggested two: "The effectiveness of predictive policing is difficult to establish because, to be a bona fide new policing strategy, it may require combining two components. The first is a software algorithm or prediction regime that is able to better predict future criminality than any existing alternative mechanisms … second, predicted grids should incur an operational response that is identified specifically with predictive policing". Both of these merit empirical assessment. Hardyns and Rummens (2018, p. 213) suggested, more expansively, gauging effectiveness "using three criteria: (1) effectiveness of the predictive analysis (how many correct predictions were made or how many crimes were missed by the predictions?); (2) crime rates before predictive policing was introduced versus after it was introduced (an indirect and likely delayed effect because of more efficient policing); and (3) costs relative to current methods being replaced by predictive policing." Furthermore, no one has yet figured out how to jointly consider these indices of effectiveness alongside important associated "ethical" and "juridical" considerations (Hardyns & Rummens, 2018, p. 214; Richardson et al., 2019). The last would be a fourth metric meriting attention in the benchmarking discussion.

In short, it is not yet clear exactly what the required set of effectiveness metrics is; whether to combine them into one overall effectiveness metric; if one did amalgamate them, how to weight the specific contributions to an overall effectiveness indicator; or, finally, how to balance effectiveness against ethical concerns about potential adverse social justice impacts.

4 | SPATIAL UNITS DEPLOYED

A range of units has been used for hot spots including corners, grids, streets, and more (Haberman, 2017, Table 1; see also Braga, Turchan, Papachristos, & Hureau, 2019). Hot-spots or predictive policing analyses and police implementation are often simplified by using fixed grid cells (Gorr & Lee, 2015, p. 35). Size as well as shape is crucial: "The key question of grid design is the size or scale of hot spot to be considered" (Gorr & Lee, 2015, p. 35). Some of the most widely used predictive policing analytics have used gridded areas ranging from 100 m × 100 m (320′ × 320′) to 250 m × 250 m (820′ × 820′) (Hardyns & Rummens, 2018, Table 1). Well-known algorithm PredPol (see below) uses 150 m × 150 m (approximately 500′ × 500′) grids. Other work has used "micro areas (typically composed of one or more block-long street segments)" (Fitzpatrick et al., 2019, p. 474). Only one head-to-head match-up contrasts the relative predictive power of grid cells versus streetblocks. Noting that "much urban crime and policing activity happens on (and along) streets, so that they represent a more meaningful representation of location than arbitrarily-defined grid squares," Rosser, Davies, Bowers, Johnson, and Cheng (2017, p. 575) observed that street network predictions substantially outperformed grid-cell predictions. Gorr and Lee's (2015, p. 42) assessment of different grid sizes for chronic hot-spots crime-capturing led them to conclude that "hot-spot size potentially could have dramatic impacts on performance in hot-spot field experiments." So, here too, agreement proves elusive.

In the case of *deployment*, variations in spatial scale are similarly noted. These differences could prove important not only practically (Rummens & Hardyn, 2020) but theoretically as well, given what we know about crime concentration differentials across space and time (Gorr & Lee, 2015).

5 | EFFECTIVENESS OF PREDICTIVE POLICING EXPERIMENTS

Fewer than a handful of predictive policing randomized control trials have gauged the impacts of predictive policing. Best known is Mohler et al.'s (2015) Los Angeles and Kent (U.K.) randomized control trial using 150-m × 150-m grids and algorithms based on software using the ETAS or epidemic-type aftershock model. This algorithm considers both nearby recent crimes and longer term dynamics (Mohler, Short, Brantingham, Schoenberg, & Tita, 2011). The focus was on three types of property crime (combined): burglary, car theft, and larceny from a motor vehicle. "The ETAS model estimates both long-term and short-term hotspots" (Mohler et al., 2015, p. 1402). In contrast to Gorr and Lee's (2015) algorithms, it cannot distinguish between chronic hot spots, more temporary hot spots, and even shorter term flare-ups. One key outcome metric for the algorithm itself was differences in the predictive accuracy index (PAI; Chainey, Tompson, & Uhlig, 2008) when the predictions were deployed. Algorithm predictions were benchmarked against crime analysts' predictions (Mohler et al., 2015, Table 2). The PAI associated with ETAS was 6.8 compared with 3.5 for analysts' predictions.[5] Turning to crime reductions, the authors considered division-wide crime reductions in LA because they "randomly assigned days to treatment and control and, more importantly, allow[ed] prediction locations to change twice daily" (Mohler et al., 2015, p. 1407). They found that "increasing patrol dosage under experimental conditions" linked significantly to lower daily property crime volume at the district level (p. 1407).

More disappointing results emerged from one other predictive policing study. The Shreveport randomized control trial PILOT program to reduce property crime observed treatment reductions, but these failed to prove statistically significant. A low number of involved districts, a program run only for a short period, relatively low property crime counts, and a treatment that had demonstrable but not overwhelming effects created a low statistical power problem (Hunt, Saunders, & Hollywood, 2014, p. 37). Rand researchers noted that "crime would have needed to fall by 30 percent … in the treatment groups to statistically identify the effect of PILOT" (Hunt et al., 2014, p. 38). As will be shown below, even in situations with proportional treatment-linked crime drops *twice* what Rand researchers said they would have needed, a predictive policing experiment can be statistically underpowered.

One experimental study, albeit about hot spots rather than about predictive policing, contained valuable lessons and suggestions for the current topic. A hot-spots randomized controlled trial focusing on broken windows policing tactics geared to reducing residents' fear, and increasing collective efficacy and perceived police legitimacy, was repurposed by Hinkle and colleagues (2013). They considered streetblock (aka "street segment") reductions in calls for service regarding a broad array of "completed and attempted offenses" (Hinkle et al., 2013, p. 219). Combining data across three cities in a single-level regression count model with streetblocks as the unit of analysis, and controlling for 6-month pre-intervention streetblock crime counts, yielded a nonsignificant

TABLE 2 Statistical power estimates for property crime prevalence rates across shift days: Typical power estimation for difference in two proportions

	AR-1	AR-2	AR-3	AR-4 5 ×	AR-4 10 ×	AR-4 15 ×	
AR-1	.1388						No Papal visit
AR-2	.2404	.2372					All eggs in one basket
AR-3	.3294	.345	.3221				Philadelphia as London
AR-4-5 ×	.3702	.5399	**.8744**	.3618			Expanding mission areas - 5 ×
AR-4-10 ×	.644	**.8074**	**.996**		.631		Expanding mission areas - 10 ×
AR-4-15 ×						**.8478**	Expanding mission areas - 15 ×
.1367	Obtained during the study, control and marked car treatment only						

Notes. Estimated levels of statistical power shown. In bold if ≥.80. Power estimates for $p < .05$, one sided.

AR-1: No Papal visit = 180 days rather than 90.

AR-2: All eggs in one basket = 15 treatment districts with just the effective marked car treatment, and 5 control districts.

AR-3: Philadelphia as London increases the number of districts 4.6 times, so k1 and k2 become 23 and 23 districts rather than 5 and 5 districts.

AR-4: Expanding mission areas: 5, 10, or 15 times the original area. Spatial scaling assumes (1) the property crime prevalence rates increase linearly with the size of the mission areas, and (2) treatment effectiveness remains the same as a proportional difference. Initial control/marked car treatment property crime prevalence rates = .033/.013.

Power for each single alternative reality shown on diagonal. Power for combination of alternative realities appear on the off diagonals.

Stata (v. 15) *power_twoproportions* for cluster randomized design = estimation software.

treatment impact of $b = .023$ (their Table 2), corresponding to a (exp(b)) 2.33 percent *increase* in crime calls during the study. The authors noted that, "[I]t was suggested that it is difficult to draw any conclusions from such analyses of these data as the statistical power of such street-segment level tests is lacking due to the low base rates in the study sites" (Hinkle et al., 2013, p. 219).

Hinkle et al. (2013) went on to comment on concerns of low statistical power, noting that "high variability in the outcome … can reduce power," but also suggesting in their case low power was "likely due to the fact that the baseline level of crime … was quite low" (p. 222). Hinkle et al. (2013) provided a most thoughtful and insightful discussion about how statistical power matters play out in studies like these. Here we extended their inquiry in four specific ways. First, we situated statistical power considerations not only in the context of a specific experiment as they did but also in the context of specific alternative hypothetical experimental designs. This clarified which particular alterations helped with statistical powerlessness. Second, we tested their idea of doing the experiment in more places, but here we kept the additional places within one hypothetical jurisdiction rather than across multiple ones. We examined empirically how much this reduced the powerlessness problem. Third, the power considerations were situated, not in the context of a modest crime change, but in a study with a substantial decline. And, finally, the concerns Hinkle et al. (2013) expressed apply not just to small- or medium-size cities with modest crime rates in comparison with those of big cities; they apply to higher crime rate big cities as well. These last three points show that the statistical powerlessness problem Hinkle et al. (2013) highlighted is far more pernicious than previously suspected. Even when the crime decline observed was substantial rather than modest, and even when more sites participated (albeit manufactured and in the same jurisdiction), statistical powerlessness still won. One study modification Hinkle et al. (2013) did not mention, running a study for longer, also received attention here. In summary, we explore whether increasing the size of the intervention areas, increasing the length of the study, or increasing the number of areas improves the statistical power of an experiment.

6 | FOCUS

As shown by the opening quote to this article and the above reviewed work, predictive policing has inherited much of the spatial ambiguity of hot-spots policing. In light of mixed results to date either from various analyses or from experiments (Gorr & Lee, 2015; Hinkle et al., 2013; Hunt et al., 2014; Mohler et al., 2015), we need to know more about the effectiveness of predictive policing. "There is a need for thorough empirical tests and evaluations for predictive policing to be considered an effective tool" (Hardyns & Rummens, 2018, p. 215). The study here helps address this need using four thought experiments and statistical power calculations of two types. It considers whether more sites help. Two additional considerations are lengthening study duration, and increasing from micro-scale to meso-scaled intervention units, with corresponding increases in property crime prevalence rates. As will be shown, only the latter successfully overcomes the statistical powerlessness problem. The next subsections situate study concerns in the context of a specific jurisdiction and experiment.

6.1 | Philadelphia property crime

Philadelphia's reported Part I property crime rate in 2015 was 3,147/100,000 residents, a rate that was 33% *higher* than Los Angeles's property crime rate (2,380/100,000) and 7% *higher* than Chicago's property crime rate (2,946/100,000) in the same period.[6] Given these figures, one could make the case that Philadelphia might provide an even better test site for a predictive policing experiment focused on property crime reduction than Los Angeles, the best known test site for PredPol.

6.2 | Philadelphia Predictive Policing Experiment

The Philadelphia Predictive Policing Experiment (3PE) used a modified version of the HunchLab (now ShotSpotter® Missions™) predictive policing software program for crime predictions, and the operational strategy randomized 20 of the 22 Philadelphia Police Districts into one of four experimental conditions. In five "awareness" districts, office staff had access to three HunchLab predicted crime areas and patrol officers were asked to pay attention to these areas when able. Each area was $500' \times 500'$. In five "marked car" districts, the awareness model was enhanced with the addition of a dedicated marked police car that concentrated on the three predicted areas. Five "unmarked car" districts also had a vehicle assigned to the three predicted areas except it was an unmarked car. Finally, in five control districts, local officers did not have access to the software. HunchLab's software (and thus the operational patrols) focused on property crime for 3 months, and then after a 2-month break, on violent crime for 3 months. The reason for the break forms one of our alternative reality scenarios. Districts were re-randomized prior to the violent crime phase. A full suite of experimental results (Ratcliffe et al., 2020) and operational implementation challenges and lessons (Ratcliffe, Taylor, & Fisher, 2019) were reported elsewhere. The full final study report also is available (Ratcliffe, Taylor, Askey, Fisher, & Koehnlein, 2019). But for now, it is useful to know that the only experimental treatment with meaningful results from the 3PE was the use of marked police cars dedicated to property crime hot spots. "Property crime comprised residential and commercial burglary, motor vehicle theft, and theft from vehicles" (Ratcliffe et al., 2019, p. 2). We describe this finding as meaningful because it demonstrated a substantial reduction in the property crime prevalence rate (60%), even though that reduction was not statistically significant. This treatment effect forms the basis for the simulation that follows, in that it sets a minimum detectable effect size sought in the simulations.

Two other results merit mention.

6.2.1 | Treatment delivery to sites (dosage)

Ride-along observers recorded police activities in 15-minute blocks and whether patrolling took place inside or outside the assigned mission grids. The property crime experiment was conducted during the 8 AM to 4 PM shift. As described in Ratcliffe et al. (2020, Figures 2 and 3, 17/27):

> [O]fficers patrolled the treatment areas to varying levels throughout the shift, with officers getting to the treatment areas earlier for the 8 a.m. to 4 p.m. property crime phases, but officers patrolling treatment areas more extensively later in the shift for the violent crime phase (6 p.m. to 2 a.m.). At least 50% saturation of treatment areas was achieved for 3.5 and 3.75 hours, respectively, for property and violent crime phases.

More specifically for property crime (Ratcliffe et al., 2020, Figure 2), for marked and unmarked cars considered together, using 15-minute block coding by ride-along observers in 79 observations, this means that in at least 50% of these observations observers coded activity occurring within the mission areas for 15 specific 15-minute segments, or 3.75 hours.

6.2.2 | Potential district-level property crime reductions in the marked car treatment districts

What was happening district-wide before, during, and after the property crime experiment? How were district-wide weekly property crime counts, for *all* Part I property crimes save arson, shifting before, during, and after the experiment?

To get a clue, the global temporal relationship between time and property crime was contrasted with the local temporal relationship. If the local relationship diverged significantly during the period the property experiment was active, that would provide an initial suggestion something might have been going on. These graphical results are being reported here for the first time.

We constructed scatterplots for all weekly district-level property crime counts from the first week in 2014 to midway through 2016 for all Philadelphia police districts. Similar scatterplots were scanned for each set of five districts, in each treatment assignment (control, awareness, unmarked patrol, marked patrol). These suggested generally increasing crime counts from the opening of 2014 until about midway through the year, followed by gently declining weekly crime counts thereafter. A quadratic smoothed curve captured this overall temporal relationship, and the curve was bracketed with a 95% confidence band. The relationship took the same form for the entire city, as well as for each set of five districts. The curvilinear smoothing captured the overall relationship and counted each week in the series equally.

A superimposed locally weighted regression line using LOWESS (locally weighted scatterplot smoothing) generated fitted property crime values that prioritized crime counts close in time to each fitted value (Cleveland, 1979; Cleveland & Devlin, 1988; Cleveland & McGill, 1984; Schmidt, Ittermann, Schulz, Grabe, & Baumeister, 2013). This latter fitting procedure provided robust estimation because it "guards against deviant points distorting the smoothed points" while simultaneously adapting "local fitting of polynomials … used for many decades to smooth time series plots" (Cleveland, 1979, p. 829) to capture the ongoing relationship at specific points in the series. This "local fitting methodology … provide[s] an exploratory graphical tool; graphing smooth surfaces that are fitted to the data can give us insight into the behavior of the data" (Cleveland & Devlin, 1988, p. 596).[7]

This process, showing the overall smoothed relationship between time and property crime counts, uncertainty around the overall smoothed relationship, and the temporally locally weighted smoothed relationship, was generated for each set of treatment and control conditions using different recommended bandwidth settings (Cleveland, 1979, p. 834). At *all* recommended bandwidths (.2, .4, .5, .6, .8), the marked car treatment districts were the *only* condition where the locally smoothed property counts deviated below the 95% lower confidence level (LCL) of the overall smoothed relationship for all or part of the treatment period.

For an example, Figure 1 shows the global and local smoothed relationships, the latter using a bandwidth of .6. For most of the treatment weeks, the local smoothed values fall below the LCL based on the overall relationship based on 2.5 years of data. The dip begins before the start of the treatment period probably in part as a result of the smoothing approach.

To be clear, this is a graphical exploration and not definitive. Nonetheless, because the marked car districts were the only condition where locally smoothed predicted property crime counts fell below the expected range (95% LCL), there is a suggestion of district-wide property crime reduction in the marked car districts. In all the smoothed plots, this was the *only* condition showing below-expected locally smoothed predicted property crime counts during the treatment phase.

The next section describes the simulation methods, power levels, and parameter choices that essentially comprise the rules of the thought experiments.

7 | METHOD

The 3PE used block randomization (Gill & Weisburd, 2013). Districts were grouped into five groups of four, based on recent crime harm and demographics.[8] In the property crime experimental phase, for the three specific mission grids for each shift, the district-day shift crime statistics (Table 6, final report) reported a property crime prevalence rate, expressed as a proportion, of .0333 in the control condition and .0133 in the marked car condition. The prevalence rate in the marked car condition was .4 times the prevalence rate in the control condition. The property crime prevalence rates refer to the proportion of days during which one or more property crimes occurred in one or more of the mission areas. The mission areas are *not* spatially buffered (cf. Ratcliffe et al., 2020). This difference (−.02) is then set as the *minimum detectable effect* or MDE (Ellis, 2010, p. 63). Variations from this minimum detectable effect are explained below.

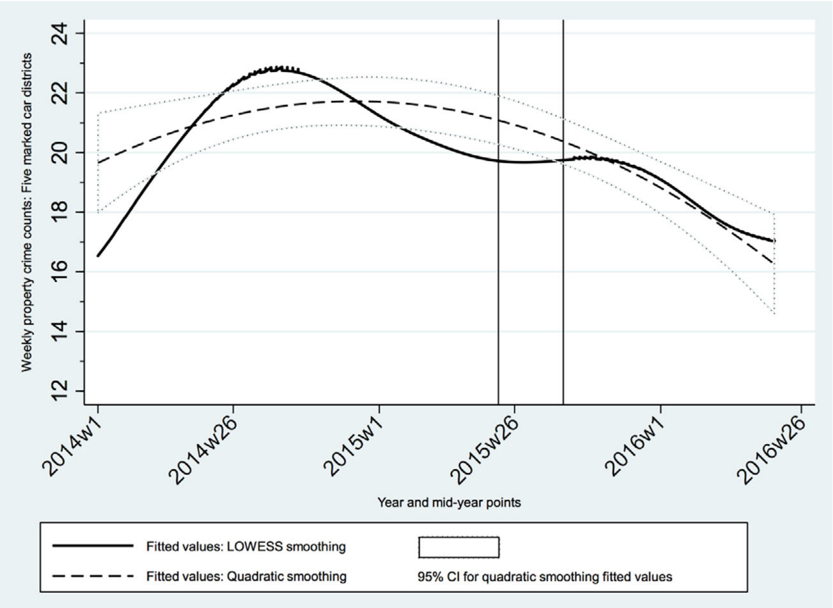

FIGURE 1 Smoothed weekly property crime counts in five districts randomized to marked car treatment. [Color figure can be viewed at wileyonlinelibrary.com]

Notes. Dashed line shows fitted values using quadratic smoothing. Dotted lines show the corresponding upper and lower 95% confidence intervals (Stata graph command qfitci). Solid line shows fitted regression values using robust locally weighted regression (LOWESS) smoothing with bandwidth set to 60%. Vertical lines indicate beginning week and ending week of the property crime experiment. During the experiment, locally smoothed fitted values dip below the 95% lower confidence interval for the overall quadratic fitted values.

7.1 | Different approaches to statistical power estimation yield different results

More background on statistical power purposes and calculations can be found in Appendix A in the online supporting information.[9] The standard approach with *pc_twoproportions* considers only mean proportions in the different conditions, and otherwise, it pays no attention to the data. By contrast, the *pc_simulate* add-on Stata program, based on the work of Burlig and colleagues (2017; Burlig, Preonas, & Woerman, 2020), conducts Monte Carlo simulations based on reading available record-level data files.[10] There are numerous reasons why these power estimates might differ from the estimates from *pc_twoproportions*. First, *pc_simulate* conducts simulations for each scenario requested. Second, it is reading actual data. Third, it is specifically geared to panel data and panel issues such as serial autocorrelation. Fourth, it starts with only nontreatment data, that is, control condition data, although we expand that here to include control + awareness conditions. As the help file explains:

> This program performs power calculations by simulating a randomized experiment using an existing dataset. For each iteration [simulation], it randomly assigns units to treated and control groups, imposes an average treatment effect on treated units, estimates this treatment effect using a regression model, and records whether the null hypothesis of zero treatment effects is rejected at the chosen significance level.

The fraction of times the null hypothesis is rejected represents the statistical power level. All simulation-based power estimates, unless otherwise specifically noted, are based on 1,000 simulations.

The data-driven simulation approach might provide different power estimates from the standard approach that relies on descriptive statistics and some other data features. The key question will be whether the two power estimation approaches agree on what type of design would generate acceptable levels of statistical power.

7.2 | A note on source data and power curves

The data-based simulations of *pc_simulate* required more than 90 days of property crime prevalence rates from the five control districts. So for estimating power of the actual design, we used 90 days of data from both the control and the awareness condition districts (total = 10 districts) as the baseline. The results from the awareness condition districts (property crime prevalence rate = .047) were somewhat close to and higher than the property prevalence crime rate in control districts (rate = .033). These ten districts combined generated an average property crime prevalence rate of .04. Assuming the marked car program would have generated the same proportional reduction of 60% meant that the minimum detectable effect (MDE) was −.024.

One last note: Statistical power analyses often report power curves, which show the power levels expected under a range of conditions. Instead, here, power is reported for the specific study conditions, as well as for the specific alternative scenarios. In essence, the current work investigates power levels for a series of specific thought experiments. The four alternate scenarios are described below. These four alternate realities have implications for the data set used for each thought experiment.

7.3 | The alternate realities

7.3.1 | No Papal visit

Alternate reality 1 (AR-1) envisioned John Lennon's "Imagine" scenario of no religion, in that Pope Francis did not visit Philadelphia September 22, 2015 through September 27, 2015 and did not lift the spirits of many millions during and after his visit there. Absent Pope Francis's visit, the PPD would not have had to devote considerable staffing to planning for the Pope's visit and ensuring his safety, and that of millions of visitors, while he was in Philadelphia. Subsequently, in this alternate reality scenario, the property crime experiment ran for double its time, for 180 days rather than 90. Let us imagine that, assuming faster-than-light data processing, the sharing of initial promising results with department leadership might have convinced them that results would look even stronger if prevention benefits could be obtained for twice the experimental time frame. With the standard power estimation approach, this scenario sticks with the actual cluster randomized design, but it focuses just on the ten districts assigned to either the control or the marked car treatment, and adjusts the parameters of cluster size, m1 and m2, to 180 each rather than to 90 each.

For the *pc_simulate* exploration, AR-1 required 180 days of data. The 90-day data set for (control + awareness) districts was simply copied, and the second 90-day period was set to begin right after the first 90-day period. MDE was still −.024.[11] Dates for the "second" 90 days needed adjusting.

7.3.2 | All eggs in one basket

Alternate reality 2 (AR-2) imagined that, benefiting from Philip K. Dick-like precognition, it was known in advance that the marked car treatment would be the most effective. Furthermore, in the *power_twoproportion* estimation, the scenario further envisioned researchers, armed with this knowledge, had convinced both the funder and the PPD to devote 15 rather than 5 districts to that intervention, scrapping the other two treatments. With 20 randomized clusters (districts), this is modeled to result in three quarters of the districts (15) being assigned to the marked car treatment and one quarter (5) being assigned to the control condition.

For AR-2, the data-tied simulations required a 20-district data set. The 90-day data set for (control and awareness) conditions was duplicated, the districts were renumbered in the second data set, and then the two data sets were joined together. The program thought it had 20 districts of data for a 90-day period and could assign 15 districts to the marked patrol treatment and 5 to the control condition. MDE was still –.024.

7.3.3 | Philadelphia becomes London

Alternate reality 3 (AR-3) imagined Philadelphia had become a mega-city (Butcher, 1995). While police districts retained their geographic size, Philadelphia's 142.7 square miles suddenly expanded to the size of London, 659 square miles. This increased the number of districts available for assignment 4.62 times. In both estimation protocols, this assumed that 46 districts, rather than 10, were available for assignment either to the control condition or to the marked car condition.

In AR-3, the data-tied simulation approach used the 10 control-plus-awareness districts 90-day data set but built samples of 46 districts per sample, rather than 10, using bootstrapping procedures. Each bootstrap sampled simulation randomly assigned 23 control and 23 treatment districts. MDE was still .024.

7.3.4 | Expanding mission areas

Alternate reality 4 (AR-4) imagined that each district had become a version of the Philadelphia badlands, a moniker given to a small area in North Philadelphia during the 1980s crack epidemic.[12] In the citywide badlands scenario, the intervention sites expanded to several times their original size. Furthermore, this geographic expansion increased the property crime prevalence rates correspondingly. Three expansions of mission-areas-and-property-prevalence-rates multiplied up by either 5 times, 10 times, or 15 times. In this alternate reality, high-crime areas were so widespread that when the size of the mission areas increased, the proportion of shift days with at least one property crime scaled up accordingly as well. For the standard power estimation, this simply required multiplying up the prevalence rates for the marked and control districts, and adjusting MDEs accordingly.

The simulation-based estimates started with 90 days of data from 10 (control-plus-awareness) districts. New prevalence rates for each level of spatial scaling were created by cloning a new outcome variable identical to the original, then randomly selecting a preset fraction of observations, and recoding those observations from "no property crime occurred" to "property crime occurred."

Given the changed prevalence rate, it was necessary to reset MDE so that the same proportional reduction in property crime prevalence was associated with the treatment. Details appear in Table 1. As the property crime prevalence rate grew larger, so too did the minimum detectable effect size. Whatever the control condition prevalence rate, the treatment prevalence rate was expected to be 60% lower.

Not surprisingly, given how large these MDEs became with hypothetically larger mission areas, statistical power levels soon turned extremely high. For each degree of spatial scaling up, the MDE associated with a more than acceptable level of statistical power, that is, a level exceeding 80%, was noted.

Of course, this scaling-up thought experiment was problematic in two respects. First, it assumed that the day-shifts-with-at-least-one-property-crime prevalence rates scaled up exactly as did the size of the mission areas. This was a liberal interpretation since HunchLab was programmed to identify the highest crime grid areas in each district. Furthermore, it assumed the marked car intervention's crime prevention effectiveness remained constant, regardless of the size of the mission areas. Both of these points get revisited in the discussion. But, just a gentle reminder: These are thought experiments, and these two assumptions are just part of the experiment as it applies to this alternate reality.

8 | RESULTS

Results using standard power estimation (Stata's *power_twoproportions*), focusing on the control districts and the marked patrol car districts, are presented first. These estimates consider only broad features of the data (number of districts, prevalence rates in treatment and control conditions, minimum detectable effect size, ICC). These are followed by estimates based on the actual daily (control-plus-awareness districts) data obtained and then used in 1,000 randomization simulations in each scenario. Finally, additional details appear on spatial scaling, MDEs, and acceptable power .

8.1 | Standard power estimation with pc_twoproportions

Results appear in Table 2. These estimates recognized cluster randomization with randomization occurring at the district level. These power estimates assume a control condition property crime prevalence rate of .0333 and a marked car treatment crime prevalence rate of .013. These two rates were each multiplied by 5 or 10 or 15 times for the different spatial scaling scenarios under AR-4.

8.1.1 | As conducted

As conducted, statistical power was estimated to be .137. Clearly, this was abysmally low. Yet, it was comparable to the post hoc power analysis of reductions in calls for service in Hinkle et al.'s (2015, Figure 3) repurposed hot-spots policing experiment. Their experiment with over 100 streetblocks had post hoc power < .10 to detect an average of one fewer calls for service on treatment street blocks, and power < .25 for detecting an average drop of two calls for service on treatment street blocks. In short, similarly powerless policing crime reduction experiments have surfaced in the literature.

8.1.2 | AR-1: No Papal visit

Under AR-1, the Papal 2015 visit to Philadelphia did not occur, and the property experiment continued for an additional 90 days. Under this scenario, the size of each cluster, m1 and m2, each went from 90 to 180. Would this have helped statistical power? Negligibly. Under this scenario, estimated power was .139. If Pope Francis had not visited, and PPD leadership had permitted extending the study, it would not have helped with the statistical powerlessness situation.

8.1.3 | AR-2: All eggs in one basket

This scenario imagined that all 15 treatment districts were assigned to the marked patrol treatment, therefore, creating 15 clusters (districts) receiving the treatment. Under this scenario, estimated statistically power almost doubled to .24 but still proved woefully inadequate.

8.1.4 | AR-3: Philadelphia as London

Under this scenario, Philadelphia expanded in size to match London. Police districts remained the same size, but the number of control districts expanded from 5 to 23, as did the number of treatment districts. Under this scenario, estimated statistical power was greater than under the previous scenarios but still paltry: .32.

8.1.5 | AR-4: Expanding mission areas

Making mission areas 5, 10, or 15 times larger, and scaling up property crime prevalence rates correspondingly, produced estimated power levels, respectively, of .362, .631, and .848. The latter represented the first acceptable level of statistical power seen so far with the standard power estimation approach: This power estimate crossed the .8 threshold.

8.1.6 | Combinations

Since scaling up mission areas and prevalence rates 15 times resulted in an acceptable level of statistical power (.85), it was not combined with other alternatives. However, three pairwise combinations of alternate realities resulted in acceptable (≥.80) levels of statistical power. Two combinations involved scaling up mission areas and prevalence rates ten times. When combined with many more districts for marked patrol and control, with 23 each if Philadelphia were as big as London, power was estimated at .996. When combined with putting all the eggs in one basket, and assigning 15 districts rather than 5 to the marked patrol condition, estimated power just exceeded the desired threshold (.81).

One of these three acceptable power combination involved scaling up the mission area sizes and prevalence rates five times. If mission areas were this many times larger and prevalence rates this many times higher, and Philadelphia were the size of London, so 23 rather than 5

TABLE 3 Statistical power estimates for property crime prevalence rates across shift days: Data-based randomization simulations using control and awareness districts (n = 10) as baseline

			Power	
		MDE	**estimate**	
Design				
Operationalized design (cluster randomized design with block randomization)				
		−.024	.253	
Operationalized design minus stratification with blocking				
		−.024	.236	
Alternate realities				
	AR-1	−.024	.242	No Papal visit
	AR-2	−.024	.364	All eggs in one basket
	AR-3	−.024	.631	Philadelphia as London
	AR-4 (5 ×)	−.12	> **.99**	Expanding mission areas - 5 ×
	AR-4 (10 ×)	−.24	> **.99**	Expanding mission areas - 10 ×
	AR-4 (15 ×)	−.36	>**.99**	Expanding mission areas -15 ×

Notes. Estimated levels of statistical power shown. In bold if ≥.80. Power estimates for $p < .05$, one sided. 1,000 simulations for each estimate. Unless stated otherwise, estimates based on ten districts with two conditions treated as baseline: control and awareness conditions. Property crime prevalence rate = .04 across ten districts. AR-1: No Papal visit = 180 days rather than 90. 10 district data set. Data were doubled, time stamp extended in second data set, then the two sets added together. AR-2: All eggs in one basket = 15 treatment districts and 5 control districts. Used data from ten control and awareness districts, but then doubled the data set and the total number of districts. AR-3: Philadelphia as London increases the number of districts 4.6 times. Uses 10 district (control and awareness) data set, but builds bootstrap samples of 46. AR-4: Expanding mission areas = all spatial scaling options use ten district (control plus awareness) data set. Spatial scaling up on size of mission areas: 5, 10, or 15 times the original area. Spatial scaling assumes the prevalence rates increase linearly with the size of the mission areas, and treatment effectiveness remains the same as a proportional difference. Initial (control and awareness)/marked car treatment property crime prevalence rates = .04/.016, MDE = −.024. Simulation, data-tied power estimates from *pc_simulate*.

Power estimate results repeated with the original data and the same commands *may not exactly match those shown here*. See Online Appendix C for sample code. This is because the program is running simulations, and the program does not seem to allow the user to set a random number seed. Differences observed, when replicating, were typically 1 digit difference in the second decimal place or smaller.

districts could be assigned to each of the two conditions, estimated power was a more than acceptable .87.

8.2 | Data-based simulation power estimates with pc_simulate

8.2.1 | General agreement with standard power estimation approach

Results appear in Table 3.[13] The data-tied simulation results agreed in several respects with the results from the standard power estimation approach. First, the design as implemented, with (.25) or without (.24) blocked randomization, produced woefully inadequate statistical power estimates just like the standard approach. Second, the statistical power estimates produced by the first three alternate realities (AR-1: no Papal visit, AR-2: all eggs in one basket, and AR-3: Philadelphia as London) were each far below minimally acceptable power levels (respectively, .24, .36, .63). Furthermore, as with the standard approach, scaling up the mission areas and concomitant property crime prevalence rates 15-fold created designs with more than acceptable statistical power (> .99).

TABLE 4 Details on MDEs generating acceptable power under different spatial scaling scenarios

Minimal detectable effect (MDE) generating minimally acceptable power	Proportional property crime prevalence rate reduction for corresponding MDE	Estimated power	Degree of spatial scaling
−.07	.35	.84	Expanding mission areas − 5 ×
−.11	.275	.86	Expanding mission areas − 10 ×
−.07	.12	.85	Expanding mission areas − 15 ×

Notes. The original miminal detectable effect (MDE) associated with each level of spatial scaling corresponded to a 60% reduction, caused by the treatment, of the (control+awareness) condition property crime prevalence rate . MDEs in decrements of .01 were run to find the MDE that would generate just acceptable levels of statistical power. Those appear in the first column. The second column translates each MDE into a percentage reduction in the property crime prevalence rate for that degree of spatial scaling. For example, under the spatial scaling x 5 scenario the (control+awareness) condition property crime prevalence rate of (.04 × 5) = .2. And, if this rate was reduced just a third (.35; MDE = −.07) by the treatment condition, this pattern would generate a design with acceptable statistical power (.84).

Power estimate results repeated with the original data and the same commands *may not exactly match those shown here*. This is because the program is running simulations, and the program does not seem to allow the user to set a random number seed. Differences observed, when replicating, were typically 1 digit difference in the second decimal place, e.g., .236 vs. .223. See example code in Online Appendix C.

So in all these respects, the simulation-based results agreed with the standard power estimation results about which designs would produce acceptable power and which would not.

8.2.2 | Points of disagreement with standard power estimation approach

Concentrating on which designs created acceptable power levels and which did not, the two approaches disagreed when mission areas and property crime prevalence rates were scaled up either fivefold or tenfold. The standard power estimation approach suggested those scaled up designs would possess insufficient levels of statistical power (.36 and .63, respectively); in contrast, the data-tied simulation approach suggested each would generate more than adequate levels of statistical power (both > .99).

8.2.3 | More details on minimum detectable effect sizes (MDEs) for spatially scaled up mission areas

To learn more about the boundary conditions for which the spatially scaled up mission areas generated minimally acceptable levels of statistical power, power was examined for each version of this alternative reality, for MDEs that were smaller than the one specified, going in .01 decrements. The MDE and the corresponding proportional reduction in the property crime prevalence rate associated with minimally acceptable statistical power, for each spatial scaling scenario, appear in Table 4. Recall that in the actual experiment, the observed property crime prevalence rate reduction was 60% for the marked patrol condition compared with the control condition. Proportional property crime prevalence rate reductions much smaller than this would have a good chance of being detected, that is, power ≥ .80, if mission areas were larger and prevalence rates were correspondingly higher.

Under the expanded mission areas scenario (AR-4) where areas are scaled up five times, a treatment-caused proportional reduction of just 35% (MDE = −.07) had an acceptable chance (power = .84) of being detected. Stated differently, a marked car treatment that was only about half (.35/.6) as effective in reducing crime as the one observed in the study would still have a good chance of proving effective, statistically speaking, with treatment areas five times larger, and property crime day shift prevalence rates correspondingly higher. The same holds, roughly (.275/.6), for a scenario with mission areas 10 × larger and prevalence rates 10 × higher. Under the 15 × larger/prevalence rate 15 × higher scenario, an intervention only a fifth as effective as the original one, expecting a proportional prevalence rate reduction of .12 rather than .60, had a good chance (power = .84) of proving statistically effective. These patterns of MDEs associated with minimally acceptable statistical power under the various spatial-and-property-crime-prevalence scaling scenarios helped address the question of treatment effectiveness. Spatially expanding mission areas will inevitably dilute the effectiveness of one assigned marked patrol car to those areas. Nonetheless, even *with* diluted effectiveness, the treatment still had good chances, statistically speaking, of demonstrating effectiveness.

9 | DISCUSSION

Researchers summarizing progress in predictive policing indicate more work is needed to determine the effectiveness of predictive policing (Hardyns & Rummens, 2018). “There have been relatively few rigorous and controlled evaluations of predictive policing programs to date” (Fitzpatrick et al., 2019, p. 485). Questions persist about effectiveness metrics (Hardyns & Rummens, 2018), how to combine appropriate metrics, and how to weave into the metrics discussion of counterbalancing social justice concerns. The latter touch not only on worries about potential overpolicing but also about the algorithms themselves (Berk et al., 2018; Fitzpatrick et al., 2019; Richardson et al., 2019). In the wake of George Floyd’s murder in Minneapolis on May 25, 2020, at the hands of a Minneapolis Police Department officer, and the widespread public concern expressed in the United States and other countries about police reform, the importance of these counterbalancing social justice concerns looms even larger.

The Philadelphia Predictive Policing Experiment proved practically successful on two counts. In the marked car treatment condition, micro-scale mission areas and their immediate surroundings experienced a 60% reduction in property crime prevalence (Ratcliffe et al., 2020). Furthermore, the graphical analysis, reported here for the first time, suggested district-wide property crime count reductions in the marked car condition during the study period. In this condition, and *only* in this condition, the locally weighted smoothing function for predicted crime counts fell below the expected range for predicted property crime counts where the latter was based on the overall property crime trend over 2 years. The departure surfaced regardless of the bandwidth used for local smoothing.

Yet, the experiment proved unsuccessful statistically due to abysmally low levels of statistical power. Here, focusing just on the micro-scale mission areas themselves, three 500′ × 500′ grids in each treatment district for 8 hours per day during the experiment, the low statistical power estimates proved comparable, using two different power estimation procedures, to those reported by Hinkle et al. (2013) in their repurposed hot-spots policing experiment.

The current work extended Hinkle et al. (2013) as follows. First, it showed that the statistical powerlessness concern they identified for midsize cities with moderate crime levels applied as well to a big city with property crime rates at the time higher than those of Chicago or Los Angeles. The scope of the powerlessness affliction is broader than earlier researchers anticipated. Second, it tested Hinkle et al.'s (2013) idea of expanding test sites. Rather than adding additional cities, the expansion was accomplished here using an alternative reality simulation (AR-3) that expanded Philadelphia to the size of London and multiplied the number of police districts accordingly (× 4.6). Results showed this did not help much with the statistical powerlessness problem. One final caution about expanding the number of test sites requires closer examination of the scientific benchmark of external validity.

The current results imply that adding cities, and within these adding control and treatment within-city sites, may do little to enhance statistical power in situations like those described here. Nonetheless, researchers, while simultaneously recognizing the challenges of multisite studies, might argue in favor of multicity experiments using an enhanced external validity rationale (Hinkle et al., 2013, p. 231), along with an improved statistical power rationale. The former rationale merits scrutiny.

Models for enhancing external validity include "random sampling [of additional sites] for representativeness;" a "model of deliberate sampling for heterogeneity;" or "an impressionistic modal instance model" (Cook & Campbell, 1979, pp. 75–77). "Where targets are specified the model of random sampling for representativeness is the most powerful model for generalizing" (Cook & Campbell, 1979, p. 78). So, following an enhanced potential generalizability argument, multicity studies would be most strongly recommended if those cities were randomly sampled. Such a random sampling approach also "permit[s] examining the data for differential effects on a variety of subpopulations" (Cook & Campbell, 1979, p. 78). Nevertheless, the extraordinary complexity of mounting a randomized control trial with one police department, let alone a random sample of police departments in different sampled jurisdictions; the limited benefits to statistical power levels of adding treatment and control locations; and finally, the nature of external validity, call into question the wisdom of pursuing such multisite predictive policing experiments with micro-scaled mission grids.

Expanding just briefly on the nature of external validity, it is necessarily *an empirical question* and cannot be determined a priori. "A study has external validity if its results *hold up* across people, across settings, and across different times" (Taylor, 1994, p. 158, emphasis added). "In the last analysis, external validity – like construct validity – is a matter of replication" (Cook & Campbell, 1979, p. 78).

In short, it is not certain that multisite experiments are the way out of the powerlessness problem. Perhaps design details, like block randomization (Gill & Weisburd, 2013), or alternative analytics like mixed models (Browne, Lahi, & Parker, 2009; Finch, Bolin, & Kelley, 2019), might sufficiently enhance the statistical power of multicity studies focusing on micro-scaled predictive policing to make them minimally viable. But this, too, is an empirical question. Hopefully, researchers will address it.

Another remedy tested here was running the experiment for longer (AR-1), doubling the number of weeks in the treatment period. Again, statistical powerlessness remained unremedied.

Only dramatic scaling up of the size of the mission areas, along with corresponding increases in property crime prevalence rates, generated adequate levels of statistical power. Using mission areas several times the original generated acceptable levels of statistical power, *even if a marked* *car intervention was only half or a quarter as effective in property crime reduction as it was in the actual experiment*. With a fivefold or tenfold spatial scaling up, an intervention half as effective generated acceptable power. With a 15-fold spatial scaling up, an intervention a quarter as effective generated acceptable power.

In short, the main takeaway of the current analysis is that predictive policing experiments need to move away from micro-scaled intervention sites on the order of 500′ × 500′ grids, considering instead larger spatial units. This appears to be the only route to conducting predictive policing experiments with acceptable levels of statistical power for analytically documenting treatment-caused Part I property crime reductions.

But scaling up also raises efficiency questions (Rummens & Hardyn, 2020). And, moving to larger spatial units intensifies already raised questions (Hardyns & Rummens, 2018) about incorporating social justice concerns into the discussion of effectiveness metrics.

Is there anything sacrosanct about micro-scale grids? From an observational perspective, the crime concentration work suggests not (Eck et al., 2017). That work indicates the greatest spatial concentration occurs at the address level, suggesting that local crime patterns involve address-level dynamics and build, either through aggregation dynamics or emergent properties, from there. But addresses pose numerous practical challenges, including figuring out which addresses merit frequent attention, and, most crucially, when. From a prediction perspective there is nothing special about 500′ × 500′ grids. One predictive algorithm did better when it used streetblocks than when it used micro-scaled grids similar to those used here and in PredPol (Rosser et al., 2017). From a theoretical perspective, there is nothing special either. No one has yet made a convincing case that these spatial units align with well-understood, theoretically grounded dynamics at a corresponding scale. As described above, the theoretical justifications at this level seem only loosely formulated and lacking in empirical validity.

Given such a recommendation for spatially scaling up predictive policing mission areas, numerous concerns follow.

From a practice perspective, three merit mention. First, if wider zones are targeted, potentially adverse social justice consequences must be considered alongside effectiveness criteria. Expanding intervention sites elevates the potential for net widening, as well as for concomitant deepening concerns about police legitimacy or police procedural justice. Given the "cone of resolution" issues discussed earlier, larger sites will increase the number of low-crime areas among the areas of crime concentration. Figuring out which effectiveness metrics to use (Hardyns & Rummens, 2018), how to weight them relative to each other, and how to simultaneously fold in to the metric discussion not only departmental costs in personnel and other arenas (Gorr & Lee, 2015) but also potential adverse social justice impacts tied to worries about unfocused policing, overpolicing, and algorithmic fairness (Berk et al., 2018; Richardson et al., 2019) is a daunting task. Daunting, but essential.

Second, what shape should the expanded mission areas be? What is sought is a "compromise between general prediction performance and spatiotemporal resolution" (Rummens & Hardyn, 2020, p. 6/9). This may depend on the crime type in question, the city type (Ariel, Weinborn, & Sherman, 2016), and associated crime patterns. For example, to learn more, crime analysts can pick apart near-repeat patterns. For shootings, the near-repeat patterns might suggest a cluster of corners for predictive policing targeting shootings. For burglaries, the near-repeat pattern might suggest a string of streetblocks for the predictive policing intervention. In all of these cases, consideration should include the ease of patrolling the activity area (Rosser et al., 2017).

Third, it is crucial that the intervention zones be sizable enough. Local researchers assisting local police departments can use the simulation approach to statistical power described here to provide specific guidance, with specific pretest crime data, about what would be sizable enough.

Of course, these considerations come *after* what should be the primary concern: how the police, alongside community stakeholders, co-determine which crime problems, where and when, most merit crime control and prevention resources, and how those resources can be delivered fairly and cost effectively.[14]

From a policy perspective, the most crucial concern may be how to arrive at and then mandate standardized effectiveness and social justice metrics. Process is as important as product. Through a public policy process, stakeholders can figure out what these metrics should be and how social justice and prevention/effectiveness get weighed against one another. In the same way that individual probationers can find it deeply disturbing that their behavior is predicted by a computer (Metz & Satariano, 2020), community stakeholders may feel similarly. Getting *both* the process and product right is more important than previously, now that larger patches of the community may be affected, and now that citizens across the United States and around the world have protested for police reform following the death of George Floyd. It is hard to overstate the policy relevance of these questions about metrics.

From a theory perspective, here is the key challenge. The latest work on spatial concentration patterns point to address-level dynamics as foundational. Can scholars construct models that start with address-level dynamics, perhaps preceded by broader contextual dynamics, and go on to explain how zones around these addresses acquire emergent properties making these zones into areas of concentrated crime? This is about figuring out the "grammar," or the meta-modeling, of how all this works (Taylor, 2015).

Save for perhaps a few threads from crime pattern theory, the needed models are unspecified. Long-term crime control, or prevention that is more than just tertiary, requires understanding these dynamics.

Also on the theory front: readers may wonder why predictive policing study impacts, like this one, are found to be weak when there are so many hot-spots policing studies with robust, statistically significant findings. Reviewing these studies suggests the following: (1) Some of these hot-spots studies have used calls for service, which occur more frequently than reported Part I serious crime incidents. The latter are the preferred outcome in predictive policing studies. (2) Even when using reported Part I serious crime incidents, not all studies have looked at outcomes by crime type. At least one study looked at just total Part I incidents. (3) Not all hot-spots studies looking at specific Part I serious crime category counts for outcomes have used randomized control trial experimental designs. More rigorous designs may make it less likely to observe significant treatment effects. (4) Furthermore, not all rigorously designed hot-spots studies looking at serious crime outcomes have observed statistically significant treatment impacts in the anticipated direction. And finally, (5) some rigorously designed hot-spots studies with serious Part I reported crime outcomes and significant treatment impacts have used treatment areas that were larger, when comparisons could be made, than typically used in predictive policing studies. If the outcome is crime prevalence rates, crime occurrence/nonoccurrence, rates go up with larger areas. The details behind this short answer appear in online Appendix B.

The current work has limitations and strengths. Limitations include assessing only a limited number of alternate scenarios and how they might affect statistical power. There are additional scenarios, as well as additional gradations of the current scenarios, and additional elaborated power scenarios combined with research design questions like randomization with blocking or alternative analytics like mixed models, that all could be explored. Certainly, further alternate realities are plausible and worthy of consideration. An additional limitation is that some of the alternate realities were less plausible than others. That said, the point was to explore hypothetical scenarios and think about the implications of those for actual analyses. Strengths include basing power estimates on actual obtained experimental data, employing two different approaches for gauging levels of statistical power, and finding that the two different approaches often agreed on when statistical power was likely to be at or above an acceptable threshold.

Building on this last point, because predictive policing studies usually move micro-grids every day or every shift, it can be hard to estimate, before a study starts, whether the planned study would be "doomed from the beginning" due to low base rates.[15] The simulation approach to power used here, which relies on record-level data from a control or baseline condition, could be applied to pretest data in prespecified control locations to estimate just how doomed the study would be, before getting underway.

In closing, the challenges for police practitioners are stark. Predictive policing studies probably need to move away from micro-scaled mission areas in light of low reported crime levels. Some might suggest shifting the goal posts and concentrating on more frequent outcomes like calls for service or Part II crimes, and keeping intervention areas small. Yet the public's demand that police address serious crime continues unabated. The focus should remain on serious crime. Furthermore, the public's simultaneous demand for socially just policing practices not only reinforces the importance of not focusing on low-level enforcement but also makes the need for evidence-based practices stronger than ever. The alternative, adopting front-line technologies and tactics that are not as evidence-based as we might hope, has known dangers. Policing pracademics are well versed in the evaluations of DARE and how that strategy provided a cautionary tale (Rosenbaum & Hanson, 1998; Rosenbaum, Flewelling, Bailey, Ringwalt, & Wilkinson, 1994). The statistical restrictions highlighted in this article could spur a much-needed discussion among evidence-based policing proponents about what constitutes "evidence" and "effectiveness" for the purposes of guiding police agencies, and how to balance effectiveness with social justice concerns. For statistical effectiveness to remain a cornerstone of policies based on evidence, police leaders may need to re-evaluate and increase their commitment to evaluation science. Otherwise, the "craft" aspect of policing will continue to dominate the "science" (Willis & Mastrofski, 2018).

CONFLICT OF INTEREST STATEMENT

The authors confirm that they have no conflict of interest to declare.

ORCID

Ralph B. Taylor https://orcid.org/0000-0002-7116-4796

ENDNOTES

[1] Predictive grids, by definition, can shift on a daily basis. If predicted grids shift spatially over substantial distances, dedicated problem-oriented policing (Goldstein, 1990) or third-party policing (Mazerolle & Ransley, 2005), both of which can be applied to hot spots because these New Yo endure more or less in one place over time, would not prove feasible. We thank the editors for this insightful point.

[2] Hot spots cannot be that foundational unit, for numerous reasons. Unless we are talking about volcanoes, hot spots do not exist in the natural or manmade world. "*Mantle* plumes are areas of hot, upwelling mantle. A hot spot develops above the plume. *Magma* generated by the *hot spot* rises through the rigid plates of the *lithosphere* and produces active volcanoes at the Earth's surface" (Oregon State University, n.d., para. 1). "To conclude that hot spots are free standing entities existing in the real world is to commit the logical fallacy of reification" (Taylor, 2015, p. 126). Consequently, "there is no coherent unity *intrinsic* to each hot spot itself. Its definition is fundamentally relativistic" (Taylor, 2015, p. 126). Thresholds of higher-than-surrounding-crime can be assessed not only

to discern hot spots but also to classify into different types such as (Gorr & Lee, 2015) chronic, spikes, or panics (Ratcliffe, 2019). The lack of a real-world referent allows researchers and practitioners to define hot spots across a wide range of spatial scales. “Police departments have focused on high-crime areas ranging from large police beats to ‘grid areas’ (collections of street blocks approximately the size of a few football fields) to microplaces (single street blocks and intersections)” (Haberman, 2017, p. 635). Gorr and Lee (2015) similarly pointed to the range question: “The key question of grid design is the size or scale of hot spot to be considered,” going on to note that a hot spot could be as small as “one block long street segments” (p. 35) or as big as “several contiguous blocks of major commercial areas such as the central business district” (p. 36). Of course there are practical advantages to such flexibility in police departments implementing hot-spots policing efforts, allowing them to tailor locations to current purposes and accommodate resource and logistical constraints (Gorr & Lee, 2015). But this still leaves unanswered the unit of analysis question.

[3] This approach runs the risk of repeating the mistakes of operationism (Feigl, 1945), a misadventure for which criminology was scolded in the 1930s (Laub, 2006; Michael & Adler, 1933; Taylor, 2015, p. 28).

[4] Here is a current example. “Predictive analytics in policing is the practice of forecasting crime patterns across time and space to inform decision-making for crime prevention. A major example is the identification of crime hot spots … as such, crime hot spots are excellent targets for crime prevention through directed patrol or problem-solving, making it desirable to build models and construct analytic methods for *predicting* their occurrence … the application of predictive analytics to crime prevention falls under the broad category of proactive policing” (Fitzpatrick, Gorr, & Neill, 2019, p. 474, emphasis added).

[5] Note that higher PAI values are better than lower values. Although the PAI is widely employed as a crime reduction benchmark, its results can be misleading, depending on how the denominator is handled (Drawve & Wooditch, 2019).

[6] Calculation from Table 8 “Offenses known to law enforcement” associated “by city” data files provided by the FBI for the 2015 UCR reports: https://ucr.fbi.gov/crime-in-the-u.s/2015/crime-in-the-u.s.-2015/resource-pages/downloads/download-printable-files

[7] The procedure calculates $\mathrm{y}_{\mathrm{predicted}}$ values by smoothing across nearby values, with closer-in-time (closer on the *x* axis) data points weighted more heavily. It simultaneously iteratively reestimates based on $(y_{\mathrm{observed}} - y_{\mathrm{predicted}})$ residuals, so that “large residuals result in small weights and small residuals result in large weights” (Cleveland, 1979, p. 830) as in iteratively weighted least-squares estimation. “There are four tuning parameters but only one, *f*, merits exploring for a range of different values” (Cleveland, 1979, p. 833). “[W]here the sole purpose of the smooth is just to enhance the visual perception … choosing *f* [fraction of data points considered in defining nearby] in the range of .2 [bandwidth = 20 percent] to .8 [bandwidth = 80 percent] should serve most purposes” (Cleveland, 1979, p. 834). Higher values of *f* produce an overall smoother set of fitted values.

[8] It was possible to gauge the improvements in power associated with block randomization but only with the power option relying on data-based simulations. In the software employed, pc_simulate, the stratify option allows one to incorporate a blocking variable into the randomization. This option, however, changes the meaning of the *n* of cases option so that it now reflects the *n* in “each stratified randomization cell,” which is two. This precludes calculating power in some of the alternative scenarios.

[9] Online appendices can be found at www.rbtaylor.net/supplemental/pub_cpp_2020_app.pdf.

[10] This approach appears infrequently in the literature because it is so recent. The module is obtained by issuing the Stata command *ssc install pcpanel*.

[11] Example do files for each type of power estimation can be found in the online appendices at www.rbtaylor.net/supplemental/pub_cpp_2020_app.pdf.

[12] While not comprising a formal area, the badlands is traditionally an area in and around Kensington Avenue and the neighborhoods to the west. The name has been frequently used in local news media and was featured in Lopez’s (Lopez, 1995) well-known novel, *Third and Indiana*.

[13] All estimates shown did not “absorb” fixed effects associated with time. Runs that did “absorb” were conducted as well, and generated estimates very close, usually within .01, of those described here. Details not shown.

[14] The authors appreciate the astute reviewer pointing out the primary concern always should be not how do we design an effective experiment but how do police and community identify, strategize about, and successfully address specific crime problems in specific locations at specific times. We agree wholeheartedly.

[15] We thank an anonymous reviewer for this turn of phrase.

REFERENCES

Ariel, B., Weinborn, C., & Sherman, L. W. (2016). "Soft" policing at hot spots—do police community support officers work? A randomized controlled trial. *Journal of Experimental Criminology*, *12*(3), 277–317. https://doi.org/10.1007/s11292-016-9260-4

Baldassare, M. (2002). *California in the next milennium*. Berkeley, Calif.: University of California Press.

Beavon, D. J. K., Brantingham, P. L., & Brantingham, P. J. (1994). The influence of street networks on the patterning of property offenses. In R. V. Clarke (Ed.), *Crime prevention studies* (Vol. 2, pp. 115–148). Willow Tree Press.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments:THe state of the art. *Sociological Methods & Research*, *0*(0), 0049124118782533. https://doi.org/10.1177/0049124118782533

Bernasco, W. (2008). Them again?:SAme-offender involvement in repeat and near repeat burglaries. *European Journal of Criminology*, *5*(4), 411–431. https://doi.org/10.1177/1477370808095124

Block, R. L., & Block, C. R. (1995). Space, place and crime: Hot spot areas and hot places of liquor-related crime. In J. E. Eck & D. Weisburd (Eds.), *Crime and place* (pp. 145–183). Criminal Justice Press.

Bourgois, P. (1996). *In search of respect*. New York: Cambridge University Press.

Bowers, K. J., & Johnson, S. D. (2004). Who commits near repeats? A test of the boost explanation. *Western Criminology Review*, *5*(3), 12–24.

Braga, A. A. (2001). The effects of hot spots policing on crime. *The ANNALS of the American Academy of Political and Social Science*, *578*, 104–125.

Braga, A. A., Hureau, D. M., & Papachristos, A. V. (2011). The relevance of micro places to citywide robbery trends: A longitudinal analysis of robbery incidents at street corners and block faces in boston. *Journal of Research in Crime and Delinquency*, *48*(1), 7–32. https://doi.org/10.1177/0022427810384137

Braga, A. A., Turchan, B., Papachristos, A. V., & Hureau, D. M. (2019). Hot spots policing of small geographic areas effects of crime. *Campbell Systematic Reviews*, *15*, e1046. https://doi.org/10.1002/cl2.1046

Brantingham, P., Glässer, U., Jackson, P., & Vajihollahi, M. (2009). Modeling criminal activity in urban landscapes. In N. Memon, J. David Farley, D. Hicks, & T. Rosenorn (Eds.), *Mathematical methods in counterterrorism* (pp. 9–31). Springer.

Brantingham, P. J., Dyreson, D. A., & Brantinghm, P. L. (1976). Crime seen through a cone of resolution. *American Behavioral Scientist*, *20*, 261–274.

Brantingham, P. L., & Brantingham, P. J. (1993). Environment, routine, and situation: Toward a pattern theory of crime. In R. V. Clarke & M. Felson (Eds.), *Routine activity and rational choice* (Vol. 5, pp. 259–294). Transaction.

Brantingham, P. L., & Brantingham, P. J. (1995). Criminality of place: Crime generators and crime attractors. *European Journal on Crime Policy and Research*, *3*(3), 5–26.

Brantingham, P. L., & Brantingham, P. J. (1999). Theoretical model of crime hot spot generation. *Studies on Crime and Crime Prevention*, *8*(1), 7–26.

Brantingham, P. L., & Brantingham, P. J. (2008). The rules of crime pattern theory. In R. Wortley & L. G. Mazerolle (Eds.), *Environmental criminology and crime analysis*. Willan.

Browne, W. J., Lahi, M. G., & Parker, R. M. A. (2009). A guide to sample size calculations for random effects models via simulation ahd the MLPowSim software package. University of Bristol, *Centre for Multilevel Modeling*. Retrieved from http://www.cmm.bristol.ac.uk/MLwiN/MLPowSim/index.shtml

Burlig, F., Freonas, L., & Woerman, M. (2017). Panel data and experimental design. *Energy Institute at Haas. Working Paper 277*. Retrieved from https://ei.haas.berkeley.edu/research/papers/WP%20277.pdf

Burlig, F., Preonas, L., & Woerman, M. (2020). Panel data and experimental design. *Journal of Development Economics*, *144*(May), 102458. https://doi.org/10.1016/j-jdeveco.2020.102458

Butcher, M. (1995). *The A-Z of Judge Dredd: The complete encyclopedia from Aaron Aardvark to Zachary Zziiz*. St. Martin's Press. New York.

Chainey, S., Tompson, L., & Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, *21*(1-2), 4–28. https://doi.org/10.1057/palgrave.sj.8350066

Clarke, R. V., & Eck, J. E. (2007). *Understanding risky facilities: Problem-oriented guides for police problem-solving tools series No. 6*. U.S. Department of Justice. Washington, D.C.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, *74*, 829–836.

Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, *83*(403), 596–610. https://doi.org/10.2307/2289282

Cleveland, W. S., & McGill, R. (1984). The many faces of a scatterplot. *Journal of the American Statistical Association*, *79*(388), 807–822.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation*. New York: Rand-McNally.

Drawve, G., & Wooditch, A. (2019). A research note on the methodological and theoretical considerations for assessing crime forecasting accuracy with the predictive accuracy index. *Journal of Criminal Justice*, *64*, 101625. https://doi.org/10.1016/j.jcrimjus.2019.101625

Eck, J. E., Chainey, S., Cameron, J. G., Leitner, M., & Wilson, R. E. (2005). *Mapping crime: Understanding hot spots* (NIJ Special Report). Office of Justice Programs, National Institute of Justice. Washington, D.C.

Eck, J. E., Lee, Y., O, S., & Martinez, N. (2017). Compared to what? Estimating the relative concentration of crime at places using systematic and other reviews. *Crime Science*, *6*(8). https://doi.org/10.1186/s40163-017-0070-4

Eck, J. E., & Weisburd, D. (1995). Crime places in crime theory. In J. E. Eck & D. Weisburd (Eds.), *Crime and place Crime prevention studies* (Vol. 4, pp. 1–34). Criminal Justice Press.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis and the interpretation of research results*. New York: Cambridge University Press.

Feigl, H. (1945). Operationism and scientific method. *Psychological Review*, *52*(5), 250–259.

Finch, W. H., Bolin, J. E., & Kelley, K. (2019). *Multilevel modeling using R* (2nd ed). CRC Press.

Fitzpatrick, D. J., Gorr, W. L., & Neill, D. B. (2019). Keeping score: Predictive analytics in policing. *Annual Review of Criminology*, *2*, 473–491.

Frisbie, D. etal. (1978). *Crime in Minneapolis*. Minneapolis, Minn.: Minnesota Crime Prevention Center.

Gawande, A. (2011, Jaunary 24). The hot spotters. *The New Yorker*, 41–51.

Gill, C. E., & Weisburd, D. (2013). Increasing equivalence in small sample place-based experiments: Taking advantage of block randomization methods. In B. C. Welsh, A. A. Braga, & G. J. N. Bruinsma (Eds.), *Experimental criminology: Prospects for advancing science and public policy* (pp. 141–162). Cambridge University Press.

Goldstein, H. (1990). *Problem-oriented policing*. Philadelphia: Temple University Press.

Gorr, W. L., & Lee, Y. (2015). Early warning system for temporary crime hot spots. *Journal of Quantitative Criminology*, *31*(1), 25–47. https://doi.org/10.1007/s10940-014-9223-8

Groff, E. R., & McCord, E. S. (2011). The role of neighborhood parks as crime generators. *Security Journal*, *25*(1), 1–24. https://doi.org/10.1057/sj.2011.1

Haberman, C. P. (2017). Overlapping hot spots? Examination of the spatial heterogeneity of hot spots of different crime types. *Criminology & Public Policy*, *16*(2), 633–660. https://doi.org/10.1111/1745-9133.12303

Hardyns, W., & Rummens, A. (2018). Predictive policing as a new tool for law enforcement? Recent developments and challenges. *European Journal on Criminal Policy and Research*, *24*(3), 201–218. https://doi.org/10.1007/s10610-017-9361-2

Hinkle, J. C., Weisburd, D., Famega, C., & Ready, J. (2013). The problem is not just sample size: The consequences of low base rates in policing experiments in smaller cities. *Evaluation Review*, *37*(3-4), 213–238. https://doi.org/10.1177/0193841x13519799

Hunt, P., Saunders, J., & Hollywood, J. S. (2014). Evaluation of the Shreveport Predictive Policing Experiment. *RAND Corporation*. Retrieved from https://www.rand.org/pubs/research_reports/RR531.html

Jennings, J. M., Milam, A. J., Greiner, A., Furr-Holden, D. M., Curriero, F. C., & Thornton, R. J. (2013). Neighborhood alcohol outlets and the association with violent crime in one Mid-Atlantic City. *Journal of Urban Health*, *91*(1), 62–71.

Johnson, S. D., & Bowers, K. J. (2004). The stability of space-time clusters of burglary. *British Journal of Criminology*, *44*(1), 55–65.

Kinney, J. B., Brantingham, P. L., Wuschke, K., Kirk, M. G., & Brantingham, P. J. (2008). Crime attractors, generators and detractors: Land use and urban crime opportunities. *Built Environment*, *34*(1), 62–74.

Laub, J. H. (2006). Edwin H. Sutherland and the Michael-Adler report: Searching for the soul of criminology seventy years later. *Criminology*, *44*(2), 235–257. https://doi.org/10.1111/j.1745-9125.2006.00048.x

Lawton, B. A., Taylor, R. B., & Luongo, A. J. (2005). Police officers on drug corners in Philadelphia, drug crime, and violent crime: Intended, diffusion, and displacement impacts. *Justice Quarterly*, *22*(4), 427–451. https://doi.org/10.1080/07418820500364619

Lee, Y., Eck, J. E., O, S., & Martinez, N. N. (2017). How concentrated is crime at places" A systematic review from 1970 to 2015. *Crime Science*, *6*(6).

Liebow, E. (1967). *Tally's corner*. New York: Little, Brown.

Lopez, S. (1995). *Third and Indiana*. New York: Penguin Books.

Loukaitou-Sideris, A. (1999). Hot spots of bus stop crime: The importance of environmental attributes. *Journal of the American Planning Association*, *65*(4), 395–411.

Maltz, M. D. (1995). Criminality in space and time. In J. E. Eck & D. Weisburd (Eds.), *Crime and place* (pp. 315–347). Criminal Justice Press.

Mazerolle, L. G., Kadleck, C., & Roehl, J. (1998). Controlling drug and disorder problems: The role of place managers. *Criminology*, *36*, 371–404.

Mazerolle, L. G., & Ransley, J. (2005). *Third-party policing*. New York: Cambridge University Press.

Metz, C., & Satariano, A. (2020, February 9). An algorithm that grants freedom, or takes it away. *New York Times*. Print edition, Page BU1. Retrieved from https://www.nytimes.com/2020/2002/2006/technology/predictive-algorithms-crime.html?searchResultPosition=2021

Michael, J., & Adler, M. J. (1933). *Crime, law and social science*. New York: Harcourt, Brace.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., & Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, *106*(493), 100–108. https://doi.org/10.1198/jasa.2011.ap09546

Mohler, G. O., Short, M. B., Malinowski, S., Johnson, M., Tita, G. E., Bertozzi, A. L., & Brantingham, P. J. (2015). Randomized controlled field trials of predictive policing. *Journal of the American Statistical Association*, *110*(512), 1399–1411. https://doi.org/10.1080/01621459.2015.1077710

Monmonier, M. (2008). *Cartographies of danger: Mapping hazards in America*. Chicago: University of Chicago Press.

National Academies of Sciences Engineering and Medicine. (2018). *Proactive policing: Effects on crime and communities*. Washington, D.C.: The National Academies Press. https://doi.org/10.17226/24928

Oregon State University. (n.d.). What is a hot spot? *Volcano World*. Retrieved from http://volcano.oregonstate.edu/what-is-a-hot-spot

Ratcliffe, J. H. (2008). *Intelligence-led policing*. Cullompton, Devon: Willan.

Ratcliffe, J. H. (2012). The spatial extent of criminogenic places: A changepoint regression of vilence around bars. *Geographical Analysis*, *44*, 302–320.

Ratcliffe, J. H. (2019). *Reducing crime: A companion for police leaders*. New York: Routledge.

Ratcliffe, J. H., & Rengert, G. F. (2008). Near repeat patterns in Philadelphia shootings. *Security Journal*, *21*(1-2), 58–76.

Ratcliffe, J. H., Taylor, R. B., Askey, A. P., Fisher, R., & Koehnlein, J. M. (2019). *The Philadelphia Predictive Policing Experiment: Final report submitted to the National Institute of Justice* [Grant 2014-R2-CX-0002]. Retrieved from https://bit.ly/376RuYf

Ratcliffe, J. H., Taylor, R. B., Askey, A. P., Thomas, K., Grasso, J., Bethel, K. J., … Koehnlein, J. (2020). The Philadelphia predictive policing experiment. *Journal of Experimental Criminology*, Advance online publication. https://doi.org/10.1007/s11292-019-09400-2

Ratcliffe, J. H., Taylor, R. B., & Fisher, R. (2019). Conflicts and congruencies between predictive policing and the patrol officer's craft. *Policing & Society*, 1–17. https://doi.org/10.1080/10439463.2019.1577844

Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYU Law Review*, *94*(May), 192–233.

Roberts, D., Taylor, R. B., Garcia, R. M., & Perenzin, A. (2014). Intra-streetblock ordered segmentation in a high crime urban neighborhood. *Journal of Architectural & Planning Research*, *31*(2), 143–162.

Roncek, D., & Maier, P. (1991). Bars, blocks, and crime revisited: Linking the theory of routine activities to the empiricism of "hot spots.". *Criminology*, *29*, 725–753.

Rosenbaum, D. P., Flewelling, R. L., Bailey, S. L., Ringwalt, C. L., & Wilkinson, D. L. (1994). Cops in the classroom: A longitudinal evaluation of Drug Abuse Resistance Education (DARE). *Journal of Research in Crime and Delinquency*, *31*(1), 3–31.

Rosenbaum, D. P., & Hanson, G. S. (1998). Assessing the effects of school-based drug education: A six-year multilevel analysis of Project D.A.R.E. *Journal of Research in Crime and Delinquency*, *35*(4), 381–412.

Rosser, G., Davies, T., Bowers, K. J., Johnson, S. D., & Cheng, T. (2017). Predictive crime mapping: Arbitrary grids or street networks? *Journal of Quantitative Criminology*, *33*(3), 569–594. https://doi.org/10.1007/s10940-016-9321-x

Rummens, Anneleen & Hardyns, Wim (2020) The effect of spatiotemporal resolution on predictive policing model performance. *International Journal of Forecasting*, online first, https://doi.org/10.1016/j.ijforecast.2020.03.006.

Schmidt, C. O., Ittermann, T., Schulz, A., Grabe, H. J., & Baumeister, S. E. (2013). Linear, nonlinear or categorical: How to treat complex associations? Splines and nonparametric approaches. *International Journal of Public Health*, *58*(1), 161–165. https://doi.org/10.1007/s00038-012-0363-z

Schumacher, E. F. (1975). *Small is beautiful: Economics as if people mattered*. New York: Harper & Row.

Shapiro, A. (2017). Comment: Reform predictive policing. *Nature*, *541*(January 26), 458–460.

Sherman, L. W. (1989). Repeat calls for service: Policing the "hot spots". In D. J. Kenney (Ed.), *Police and policing: Contemporary issues* (pp. 150–165). Praeger.

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

Simon, D., & Burns, E. (1997). *The corner: A year in the life of an inner-city neighborhood*. New York: Broadway Books.

Spelman, W. (1995). Criminal careers of public places. In J. E. Eck & D. Weisburd (Eds.), *Crime and place* (pp. 115–144). Criminal Justice Press.

St. Jean, P. K. B. (2007). *Pockets of crime: Broken windows, collective efficacy, and the criminal point of view*. Chicago: University of Chicago Press.

Suttles, G. D. (1968). *The social order of the slum*. Chicago: University of Chicago Press.

Taniguchi, T. A., Ratcliffe, J. H., & Taylor, R. B. (2011). Gang set space, drug markets, and crime around drug corners in Camden. *Journal of Research in Crime and Delinquency*, *48*(3), 327–363. https://doi.org/10.1177/0022427810393016

Taylor, R. B. (1994). *Research methods in criminal justice*. New York: McGraw Hill.

Taylor, R. B. (1997). Social order and disorder of streetblocks and neighborhoods: Ecology, microecology and the systemic model of social disorganization. *Journal of Research in Crime and Delinquency*, *33*, 113–155.

Taylor, R. B. (2015). *Community criminology: Fundamentals of spatial and temporal scaling, ecological indicators, and selectivity bias*. New York: New York University Press.

Taylor, R. B., Gottfredson, S. D., & Brower, S. (1984). Understanding block crime and fear. *Journal of Research in Crime and Delinquency*, *21*, 303–331.

Thrasher, F. (1926). The gang as a symptom of community disorganization. *Journal of Applied Sociology*, *1*(1), 3–27.

Thrasher, F. M. (1927). *The gang: A study of 1,313 gangs in Chicago*. Chicago: University of Chicago Press.

Tita, G., & Ridgeway, G. (2007). The impact of gang formation on local patterns of crime. *Journal of Research in Crime and Delinquency*, *44*(2), 208–237. https://doi.org/10.1177/0022427806298356

Townsley, M., Homel, R., & Chaseling, J. (2003). Infectious burglaries: A test of the near repeat hypothesis. *British Journal of Criminology*, *43*(3), 615–633. https://doi.org/10.1093/bjc/43.3.615

Tyler, T., Fagan, J., & Geller, A. (2014). Street stops and police legitimacy: Teachable moments in young urban men's legal socialization. *Journal of Empirical Legal Studies*, *11*(4), 751–785. https://doi.org/10.1111/jels.12055

Van Patten, I. T., McKeldin-Coner, J., & Cox, D. (2009). A microspatial analysis of robbery: Prospective hot spotting in a small city. *Crime Mapping: A journal of research and practice*, *1*(1), 7–32.

Weisburd, D. (2015). The law of crime concentration and the criminology of place. *Criminology*, *53*(2), 133–157. https://doi.org/10.1111/1745-9125.12070

Weisburd, D., Groff, E. R., & Yang, S.-M. (2012). *The criminology of place: Street segments and our understanding of the crime problem*. New York: Oxford University Press.

Weisburd, D., & Mazerolle, L. G. (2000). Crime and disorder in drug hot spots: Implications for theory and practice in policing. *Police Quarterly*, *3*(3), 331–349.

Weisburd, D., Morris, N., & Groff, E. R. (2009). Hot spots of juvenile crime: A longitudinal study of arrest incidents at street segments in Seattle, Washington. *Journal of Quantitative Criminology*, *25*(4), 443–467. https://doi.org/10.1007/s10940-009-9075-9

Whyte, W. F. (1943). *Street corner society*. Chicago: University of Chicago Press.

Wilcox, P., Land, K. C., & Hunt, S. A. (2003). *Criminal circumstance: A dynamic multicontextual criminal opportunity theory*. New York: Aldine deGruyter.

Willis, J. J., & Mastrofski, S. D. (2018). Improving policing by integrating craft and science: What can patrol officers teach us about good police work? *Policing & Society*, *28*(1), 27–44.

Wolfgang, M. E. (1983). Delinquency in two birth cohorts. *American Behavioral Scientist*, *27*(1), 75–86.

Yang, S.-M. (2010). Assessing the spatial–temporal relationship between disorder and violence. *Journal of Quantitative Criminology*, *26*(1), 139–163. https://doi.org/10.1007/s10940-009-9085-7

AUTHOR BIOGRAPHIES

Ralph B. Taylor is a community criminologist and the author of Human *Territorial Functioning* (Cambridge University Press, 1988), *Research Methods in Criminal Justice* (McGraw Hill, 1994), *Breaking Away from Broken Windows* (Westview, 2001) and *Community Criminology* (New York University Press, 2015). He has authored or co-authored over 80 articles in criminology, criminal justice, social psychology, urban affairs and sociology. Current interests center on prediction and meta-modeling of community crime patterns and levels; how crime and reactions to crime link to physical, social, and cultural elements of communities of different sizes; the interface between communities and police; and evaluation.

Jerry Ratcliffe is a former British police officer, college professor, and host of the Reducing Crime podcast. His research focuses on evidence-based policing and crime analysis, and he works with police agencies around the world on crime reduction and criminal intelligence strategy. After an ice-climbing accident ended a decade-long career with London's Metropolitan Police, he earned a first class honors degree and a PhD from the University of Nottingham. He has published over 90 research articles and nine books, including most recently Reducing Crime: A Companion for Police Leaders.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Taylor RB, Ratcliffe JH. Was the pope to blame? Statistical powerlessness and the predictive policing of micro-scale randomized control trials. *Criminology Public Policy*. 2020;1–31. https://doi.org/10.1111/1745-9133.12514