To catch a terrorist: can ethnic profiling work?

In a world threatened by terrorists from a small number of countries, it is tempting to think that racial profiling for security purposes, even if morally objectionable, might save lives. But is it mathematically sound? **William Press** shows that even with unrealistically perfect data it is surprisingly difficult to gain any benefit from such profiling.

It is widely accepted that, in support of the common good, a government may inconvenience its citizens, some more than others. Government officials were surprised, therefore, at the strength of opposition to the so-called Patrick Plan for reducing traffic fatalities. In essence, the plan provided that vehicles driven by anyone named Patrick should bear a prominent identifying "P" sticker and be liable to be stopped by the police, without any other cause, for a check of the vehicle's documents and the driver's sobriety. Patrick vehicles were also excluded from certain streets at certain times, notably late at night in the vicinity of bars and pubs. In support of the plan, the government adduced evidence that Patricks were in fact responsible for more than their share of serious traffic accidents, especially ones involving alcohol. To counter any bad feelings among the Patricks, the government put up posters showing happy drinkers raising their glasses, with the caption, "Thank you, Pat, for making us a little bit safer!"

People objecting to the plan could generally be divided into two groups. Some argued on grounds of fairness, that it seemed fundamentally unjust that any one Patrick, more than likely a safe driver himself, should be penalised for the actions of unrelated and unknown other Patricks. Others argued against the plan on grounds of efficacy, challenging the government's evidence on various statistical grounds – though in the end never entirely impeaching it. It was pointed out that a Patrick could escape the penalty by legally changing his name, an action that seemed unlikely to affect his drinking or driving habits. Newspapers reported cases of law-abiding, pedestrian Patricks coming out of bars and being run over – often by totally drunken drivers named Bruce.

Only a few statisticians and economists positioned themselves in the middle as would-be utilitarians. If the societal gain were large enough, they said, and the inconvenience to the Patricks were small enough, then the plan would be tolerable. However, no two utilitarians could ever agree on how to implement such a calculation, much less on the empirical data that would support it.

***

© iStockphoto.com/Tom Tomczyk

Police strategies that allocate enforcement resources on the basis of prior probabilities that individuals will commit crimes are called "actuarial methods". One can argue that all police work is at some level actuarial, since the officer on the beat or in the patrol car surely relies on past experience in deciding whether to investigate some small suspicious sign that is not itself criminal. However, in a modern context, an actuarial method is a more formal method for imputing criminal risk on the basis of an individual's match to some list of characteristics – a profile. An example, not particularly controversial, is the screening of income tax returns for signs of tax evasion, and the allocation of auditing resources thereby. More controversial are the uses of actuarial methods in sentencing and parole decisions, and at transportation checkpoints – particularly at airports.

If we feel some discomfort at the report of the Patrick Plan it may be because we detect in it something other than a benign, if misguided, use of actuarial methods. We recognise that the name Patrick is not uniformly distributed across all ethnicities. Some of these ethnicities, more than others, are stereotypically connected to drunkenness. Is the government's singling out of Patricks in fact a racial profile, if only somewhat obfuscated and denatured? And does it matter if it is?

Racial profiling, as commonly defined[1], is any actuarial method that conditions an individual's prior probability of criminal behaviour explicitly on his or her race, ethnicity, nationality or religion. Mature democratic societies recognise racial profiling as not merely another type of actuarial policing, but as something deeply corrosive of democratic values. Racial profiling violates the democratic covenant that individuals are to be judged by a universal rule of law, not by shifting standards that vary with their being assigned to stereotypical racial or other categories. To reject racial profiling is both a fundamental moral decision that we make as individuals, and also a political decision that we make as a society, recognising that stable democracies require in practice not just majority rule, but also minority rights and the rule of law.

This rejection of racial profiling is independent of the question of whether racial profiling actually "works", that is, whether race and national origin do in fact have some statistical predictive power in screening for some types of criminal behaviour. In today's world of course the question particularly arises with terrorism. To engage in the discussion of whether racial profiling could in principle be effective in law enforcement is not to condone it. Rather it is to recognise that, as in the story of the Patrick Plan, we should be prepared to respond to those utilitarians who may honourably believe that a small degree of quasi-racial profiling might be justified if it could, for example, yield a large increase in terrorists who get caught.

Harcourt's *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*[2] discusses at length a number of reasons why actuarial methods in general, and racial profiling in particular, may simply not work. Some arguments rely on economic reasoning. For example, by concentrating law-enforcement resources on a profiled group, one is implicitly diluting those resources on the complement group, the rest of the population. Concentration of resources makes sense in the first place only if there is some elasticity in criminal behaviour, that is, if, on the margin, criminality decreases in the face of more enforcement. But then surely there is also some elasticity for the complement population, so that their rate of criminality will increase with less enforcement. And there are many more individuals in the complement. Whether net crime decreases or increases thus depends on the marginal propensities of the two groups to commit crimes, and also on the ratio of their numbers. It can come out either way[3,4].

Another possible unintended outcome is Harcourt's "ratchet effect": as law-enforcement resources are concentrated on a target profiled group, so that more arrests and convictions occur there, that group's prominence in crime statistics rises. Noting this disturbing trend, policy-makers respond with an even greater concentration of resources on the profiled group. In the limiting case, all resources are concentrated on this group, and so also are all arrests and convictions. This may be a stable mathematical fixed point, but its optimality is entirely spurious.

When all law-enforcement resources concentrate on one group, all arrests are in that group. Other groups appear statistically entirely law abiding

While serving to point out the complexity of real-life situations, arguments like the two just given seem somehow to duck the central statistical question: how exactly should one use actuarial information if one has it? And, even ignoring second-order effects such as the above economic argument, does such information actually reduce crime? Suppose that we have perfect actuarial information on each individual. For example, suppose that for an individual i, we know exactly the probability p_i that he is a terrorist. How shall we allocate law-enforcement resources? Proportional to his probability p_i? Or should we concentrate even more heavily on those individuals with the largest values of p_i – the ones we think most likely to be terrorists? For example, if we know that over the past few years nine terrorists out of ten have been male, should we at airports put nine-tenths of our staff to body-search men and one-tenth to women, or should we split it 99 to one, or even more heavily?

To make the example more explicit, suppose that N individuals pass through a country's transportation network and that one is a terrorist. At an airport security checkpoint, most will pass through normally, but we want to pull out a certain number of passengers for a rigorous secondary security screening. We select them based on their profile probabilities p_i. The most general strategy is that passenger i gets pulled out with some selection probability q_i that we, the authorities, get to choose. We choose q_i based on his probability, p_i, of being the terrorist. Our probability of pulling him out and searching him depends in some way on how likely a terrorist we think him. Our strategy is thus defined by a function or algorithm $q_i = q(p_i)$.

Suppose that passenger j is actually the terrorist. Then, on average, he can get through $1/q_j$ checkpoints in the transportation network before he happens to be selected for screening (and, we presume, arrested). We want to minimise this number, in expectation value, over the whole population of travellers i, any one of whom might be the terrorist j. Mathematically, we want to minimise

$$\mu = E\left(\frac{1}{q_j}\right) = \sum_{i=1}^{N} \frac{p_i}{q_j} \qquad (1)$$

If there were no other constraints, the minimum is clearly achieved by selecting *every single passenger* for secondary screening. By selecting all passengers for secondary screening we catch the terrorist the first time he attempts to go through a checkpoint. (In equation (1), we will have made all of the q_is as big as possible, namely 1.) Unfortunately, in real life, we are resource-limited and cannot do secondary screening on all passengers. (The queues at most big airports are quite bad enough already.) The fraction of passengers we select at any checkpoint is constrained to some constant resource-limited value. It is also proportional to the sum of our selection probabilities q_i. This sum $M = \sum_{i=1}^{N} q_i$ is the number of passengers out of the total N that we can afford to select for secondary screening, if each passenger were to go through one checkpoint on average.

But, given equation (1) and that we cannot screen everyone, what are the possible strategies now? One strategy is to sample randomly and uniformly, without using the profile probabilities p_i at all. Then all the selection probabilities are the same for everyone: for every passenger $q_i = M/N$. This gives for μ, the average number of checkpoints the terrorist can get through,

$$\mu = \sum_{i=1}^{N} \frac{p_i}{M/N} = \frac{N}{M} \qquad (2)$$

Thus if our checkpoint personnel pull out one passenger in every five for screening, the terrorist will on average pass through four checkpoints unscreened and be pulled out at the fifth.

Another strategy might be to sample likely terrorists more heavily. This seems like a natural thing to do. If we think men with curly hair are twice as likely to be the terrorist, we pull them out of the queue twice as often as we pull straight-haired or bald men. We are sampling in proportion to their p_i. This would be called *importance sampling* in the context of Monte Carlo integration[5]. In this case $q_i = Mp_i$, with the constant M needed to enforce our resource limit. Then we can again calculate the average number of checkpoints our terrorist would pass through before being caught:

$$\mu = \sum_{i=1}^{N} \frac{p_i}{Mp_i} = \frac{N}{M} \qquad (3)$$

Surprisingly, that answer N/M is *exactly the same as it was for uniform random sampling*. So importance sampling did not help us at all! Our terrorist, with or without curly hair, would still, on average, pass through four checkpoints unchecked, to be caught (again on average) only at the fifth. Our knowledge about most terrorists having curly hair has not helped us at all at the checkpoint.

You might think that this might be because our importance sampling failed to concentrate *heavily enough* on the likely offenders, those with the largest values of p_i (those with the curliest hair). In fact, just the opposite is true. It is a straightforward calculation using Lagrange multipliers to find the values of q_i (our "heaviness weighting") that for fixed p_is minimise our number-of-checkpoints answer μ, subject to the constraint that M, the proportion of people overall that we check, is held constant. The answer turns out to be

$$q_i = \frac{M\sqrt{p_i}}{\sum_{k=1}^{N}\sqrt{p_k}} \qquad (4)$$

In words, equation (4) says that individuals should be selected for screening only in proportion to the *square root* of their prior probability. This does use the priors, but only weakly. It results in secondary screening being distributed over a much larger segment of the population than would be the case with importance sampling or any stronger use of profiling. Crudely, if a curly-head is nine times as likely to be the terrorist, we pull out only three times as many of them for special checks. Surprisingly, and bizarrely, this turns out to be the most efficient way of catching the terrorist. (His hair may actually of course be of any type.)

The figure this gives for the number of checkpoints we shall expect to need to catch him is the minimum number possible, by any system. It turns out to be:

$$\mu = \frac{1}{M}\left(\sum_{i=1}^{N}\sqrt{p_i}\right)^2 \qquad (5)$$

Equation (5) can readily be shown to be always smaller (i.e. better) than equation (3). Square root sampling will let us identify our terrorist with the maximum efficiency possible given the resources that we have. It gives us the smallest possible number of checkpoints that he can expect to pass through unscathed. Put simply, you underplay your prior information rather than overplay it. And it results in the terrorist being likely to be caught at an earlier checkpoint than by any other method.

The idea of sampling by square root probabilities is quite general and can have many other applications. It applies whenever a "singular" event is hidden among many "ordinary" events that must be sampled with replacement, as long as the singular event can be recognised if it happens to be picked. For example, one can thus sample paths through a trellis or hidden Markov model when their number is too large to enumerate explicitly, but one path can be recognised (e.g. by secondary testing) as the desired singular one. It seems peculiar that the method of square root sampling is not better known; indeed it has been independently discovered at least several times[6–9].

The reason why our airport security checkpoint is, in effect, sampling with replacement is that ours is only one of many airport security checkpoints through which terrorists pass. An innocent individual who happens to "look suspicious" and who therefore has a large profile value p_i tends to be selected at these airports over and over again by importance sampling or by any more concentrated algorithm. (We shall call such algorithms "strong profiling".) Normally, and morally, we focus on the unfairness and inconvenience of strong profiling to the affected individual. We now see, however, that the aggregate effect of such innocent, but high profile, individuals is, on average, to draw enforcement resources away from the actual terrorist, so that fewer actual terrorists are caught. It might seem counter-intuitive that we should pass over many higher probability individuals in favour of lower probability ones – until we recognise that the alternative strategy provides quantitatively greater sanctuary for those terrorists whose profile values happens not to be so high.

For the model presented here, and generically over a large class of models[9], strong profiling is no more effective than uniform sampling of the whole population. But strong profiling, because it tends to be or become racial profiling in practical cases, extracts a huge moral toll from a democratic society. It is therefore hard to find any utilitarian argument in favour of its use: there is no positive to balance against the negative.

> Screening by strong racial profiling is no more effective than uniform sampling – and may actually be less effective

The utilitarian might want to make a last-ditch case for the use of something like square root sampling. Entering the airport security queue, you would be somehow accurately profiled as to your probability of being a terrorist. The computer, doing the square root calculation and throwing digital dice, would then signal whether you should be selected for secondary screening. Occasionally a ninety-year-old woman of European extraction would be selected, but not nearly as often as would be a young male of apparently Middle Eastern heritage. The performance, in the idealised case of perfect profile probabilities, could be somewhat better than uniform sampling. If the probabilities are less than perfect, the performance will be poorer, including the possibility of being worse than uniform sampling. Meanwhile, the moral and social cost would not seem to be significantly less than for (mathematically ineffective) strong profiling.

If there is any general advice that we can give to policy-makers, or to our colleagues in law enforcement, it would seem to be this: no strategy of using racial (or any actuarial) profiles is likely, in practice, to be substantially more effective at catching terrorists than uniform random sampling of the population that can be screened. Many such strategies, especially those with strong profiling, will be *less* effective than uniform random sampling. Indeed, uniform sampling, without the use of profiling, is surprisingly good. It is robust against false assumptions, it is a deterrent, it is easy to implement, it is about as effective as any real-life system can be – and it is devoid of moral and political hazard. The choice between a strategy of profiling and one of uniform random sampling should not be viewed as difficult. Leave alone the Patricks of whatever race, ethnicity, nationality, and religion; and worry more, uniformly, about the rest of us.

References

1. American Civil Liberties Union (2008) Campaign against Racial Profiling. `http://www.aclu.org`
2. Harcourt, B. E. (2007) *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. Chicago: University of Chicago Press.
3. Knowles, J., Persico, N. and Todd, P. (2001) Racial bias in motor vehicle searches: theory and evidence. *Journal of Political Economy*, **109**, 203–229.
4. Dominitz, J. and Knowles, J. (2006) Crime minimization and racial bias: what can we learn from police search data? *Economic Journal*, **116**, F368–F384.
5. Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (2007) *Numerical Recipes: The Art of Scientific Computing*, 3rd edition. Cambridge: Cambridge University Press, Section 7.9.1.
6. Shooman, M. L. (1991) A micro software reliability model for prediction and test apportionment. In *Proceedings 1991 IEEE International Symposium on Software Reliability Engineering*, pp. 52–59. Los Alamitos, CA: IEEE Computer Society Press.
7. Abagyan, R.A. and Totrov, M. (1999) Ab initio folding of peptides by the optimal-bias Monte Carlo minimization procedure. *Journal of Computational Physics*, **151**, 402–421.
8. Zhou, Y. and Abagyan, R. (2002) Efficient stochastic global optimization for protein structure prediction. In *Rigidity Theory and Applications* (eds M. F. Thorpe and P. M. Duxbury). New York: Springer.
9. Press, W. H. (2009) Strong profiling is not mathematically optimal for discovering rare malfeasors. *Proceedings of the National Academy of Sciences USA*, **106**, 1716–1719.

William Press is the Warren J. and Viola M. Raymer Professor in Computer Science and Integrative Biology at the University of Texas at Austin.