Who is most likely to offend in my store now? Statistical steps towards retail crime prevention with Auror

B. W. McDonald[1] L. M. Hall[2] X. P. Zhang[3]

(Received 25 January 2016; revised 2 August 2017)

Abstract

Auror is establishing itself both locally and internationally as a leader in retail crime solutions. In mid-2015 a study group of mathematicians and statisticians teamed up with Auror to analyse data from the first two and a half years of their venture to identify and prevent retail theft. The aim was to explore methods for nominating the top ten individuals most likely to offend in a particular store at a particular time. Various methods were employed to explore the relationships between retail crime incidents, including generalised linear models, regression trees, and similarity matrices. The relationships identified were then used to inform predictions on individuals most likely to reoffend. The focus of the current analysis is to model the behaviour

DOI:10.21914/anziamj.v57i0.10507, © Austral. Mathematical Soc. 2017. Published August 14, 2017, as part of the Proceedings of the 2015 Mathematics and Statistics in Industry NZ Study Group. ISSN 1445-8810. (Print two pages per sheet of paper.) Copies of this article must not be made otherwise available on the internet; instead link directly to the DOI for this article. Record comments on this article via
http://journal.austms.org.au/ojs/index.php/ANZIAMJ/comment/add/10507/0

of reoffenders. At the time of the study group the project was still in the early phases of data collection. As data collection proceeds, prediction methods will likely give better and better intelligence to aid crime prevention efforts.

Contents

1 Introduction

The New Zealand start-up company Auror (formerly Eyedentify [1]) is on a mission to address the issue of retail theft. In New Zealand (NZ) and worldwide, retail crime is a massive problem which affects everyone. Based on data from Retail NZ, retail criminals steal up to \$2 million worth of goods in NZ each day, and the total losses to the retail industry are estimated to be \$1.08 billion per year [2, 3]. This equates to an annual bill of \$605 for every New Zealand household. Rates of retail crime in New Zealand are significantly higher than the global average. Nearly a third of crimes reported to police in Australasia are theft incidents, and it is estimated that only 1% of offences are reported, allowing many offences to go unpunished [3]. Research in America indicates that 79% of justice professionals view retail theft as a gateway crime to more serious offences [4]. Given that 55% of offenders start shoplifting in their teens [5], identifying recidivism, especially among youth, could play a very important part in overall crime reduction.

Many retailers have existing channels for sharing information about offenders and reporting incidents to the police. Auror has not only simplified this process, but has turned these reports into actionable intelligence which retailers can use to minimise future incidents. This simplified process has the potential to save retail staff up to an hour per incident [5]. This will hopefully provide an incentive for higher reporting rates, and allow the focus to move from one-off incidents towards long-term theft prevention and a proactive rather than reactive response to the problem. In terms of crime minimization, better intelligence means security staff can refuse entry to stores by known suspects, or watch suspects closely and search them on egress from the store, or if there is a warrant for the suspect's arrest, inform police immediately of the suspect's location.

Auror uses the 80/20 rule as part of their long-term solution [3]. They have identified that 20% of the offenders are responsible for 80% of the incidents. Results from their data collection to date indicate that the top 10% of offenders are responsible for 55% of the total value stolen. The ability to identify and prosecute the most prolific offenders is an important step in the reduction of retail crime.

Crime intelligence which helps police is a positive consequence of Auror's platform, but more important is the benefit for retailers. The business model requires strong buy-in from retailers who are responsible for the data collection. The question Auror aims to answer for individual retailers is: Who is most likely to offend in my store now? This question has three main components. The first of these is 'who'. This corresponds to the group of offenders who are linked to multiple offences in the Auror database. Based on offender history, the aim is to come up with lists of the most likely offenders in an individual store. This list can be created in multiple ways, a few of which are discussed throughout this paper. The second component of this key question is 'in my store'. Offenders often target stores of the same type within a similar location. The key store types in the dataset to date include grocery stores, gas stations and general retail. Offenders are usually strongly linked to one of these store types, and often target stores within a particular geographical region. (Region may be defined at differing spatial levels including suburb and city.) 'Now' is the final component of the question, and corresponds to the possibility of real time adjustments to likely offender lists, based on new incident reports. Combining these three key components will enable retailers to be more prepared to prevent and identify incidents. This actionable intelligence will work best when all staff members are kept up to date with the likely offenders. Many stores have offender lists for their region which are provided by the police and updated annually. The system proposed by Auror has many advantages, including customizing the information for individual stores, and using up-to-date information.

While still in the early stages of development, by the time of the study group, Auror already had success stories to prove they had a model that worked for the industry. In early 2014, Waikato Police were able to apprehend an individual, after data from Auror enabled them to link the individual to seven offences of theft throughout the Auckland and Waikato regions. A Hamilton police inspector described Auror's services as a real force-multiplier for Police in their efforts to prevent crime [6]. The data provided enables linkage between incidents which at first appear unrelated. On the basis of early data Auror suggest that over 30% of offences are being prevented through the use their system. December 2015 saw success stories in Auckland and Christchurch as police and Auror collaborated to locate and arrest numerous recidivist thieves [7, 8]. These successes, and anticipated future impacts on crime, were contributing factors to the Auror leadership team winning the 'Young Innovators of the Year' at the New Zealand Innovators Awards in October 2015 [9].

2 Data summary

Auror provided the Mathematics in Industry New Zealand (MINZ) working group with raw data consisting of 6233 incidents over a 28 month period. This dataset includes 5256 unique individuals, of which 1942 were repeat offenders. Repeat offenders are involved in 72% of the recorded incidents, which included 243 stores in 138 suburbs around the country. The raw data consists of sixteen variables relating to the offender(s) and the incident. Information collected about offenders includes gender, height range, build and age group. When vehicles are involved in an incident it is often possible for retailers to obtain the licence plate number. Vehicles were involved in 33% of incidents, and the dataset contains 1299 unique vehicles. Auror is aware of the potential for profiling by gender, race and other attributes, and has been careful to avoid this in the variables they have chosen to record. Recording personal data also has the potential for privacy breaches, so the team has ensured that the data is consistent with privacy legislation. Stickers are placed at the front of stores using the Auror system so that customers are aware of the security measures in place.

There are a variety of variables which relate to the incident. These include time, date, location, store type and products stolen. Based on the supplied variables, we were able to create a selection of new variables to supplement our analysis. We were curious as to whether there were significant differences between weekend and weekdays, business hours and after hours, school holidays and term time, and weather conditions. The weather conditions variable was created using maximum daily temperature in the relevant city. Some of these created variables were able to offer new insights, but for others, the relationships were not fully explored or the effects were confounded by other factors. There is also potential for including other variables in the analysis. We suggest that, as more data and insight become available, models with new variables should be explored, to maximise intelligence for retailers. If an individual is known by name, and known to be in prison, then the individual's risk of reoffending is (temporarily) reduced. Similarly an individual may have recently been warned by a store, or issued a trespass notice not to enter the store, in which case they may avoid it for a few days, again resulting in reduced risk.

It is particularly important to know the location where the offender was last identified, and how long ago that was. Ideally the information and ranking would be updated continually in real time, so that if an individual is spotted at an unusual place, then nearby stores should be warned. The effect of location may be on several levels: same shopping centre; same suburb; nearby suburbs; same city; same region; elsewhere in the country. Recent incidents should be weighted higher. As well as the most recent event, the model may include some measure of the individual's usual home range in terms of locations of all past incidents.

Some offenders are regular in their habits, for example only stealing while the kids are at school, that is, from 9 am to 3 pm Monday–Friday excluding holidays. Others are seldom seen during the morning, and may be less active during rainy or cold weather. So day, time, and weather information are relevant.

Finally offenders often work in groups, and some are known to steal to order (for example they may be required to steal a quota of iphones, or manuka honey, for on-selling). So it is relevant to know the location of recent incidents at which known associates of the individual were identified, how long ago that was, and what type of item or store was involved.

One of the challenges with the data was the lack of a defined response variable. Many of the methods we initially brainstormed were regression or data mining based and required a response variable. That is, we wanted to describe increased risk of offending, but we only had data on offenders and nothing on non-offenders—leaving a situation analogous to a regression where every response is '1'. There is no data on undetected incidents, or incidents where the offender did not go through with the intended theft. This missing data rules out the possibility of a model predicting the occurrence of an offence, and first time offenders are absent from predictions due to the lack of data on them. Consequently the analyses covered in this report focus on re-offenders, especially those individuals who appear frequently in the dataset.

The remainder of this report covers the methodology and results for a variety of methods explored by the study group. The methods are categorised by whether, or how, a response variable was defined.

The first approach was not to use a response variable as such, but to explore associations between individuals. The idea is that if an individual is *ipso facto* a prolific offender, then if we know others who are similar to that individual, then we should watch out for them as well. In Section 3 a method using similarity matrices is described. Section 4 examines the accomplices that an individual might have.

The second approach was to create response variables based on repeat offending. Two variables were created. The first was a binary variable which was equal to one if the individual offended more than once in the dataset, and zero otherwise. The second response variable assigned a count of the total number of incidents by each offender. This variable enabled easy identification of the most prolific offenders. Section 5 looks at trying to explain what separates repeat offenders from one-timers.

The last approach (Sections 6 and 7) is to restrict attention to individuals already in the database up to a certain month in order to find out how likely they are to reoffend during the next month (or three months), and if so how often? Section 6 uses regression models to build a linear combination of covariates, and suggests that once the linear combination is defined, the covariate information can be continuously updated, giving a solution to the 'my store, now' problem. Section 7 considers some alternative approaches.

Some concluding remarks are given in Section 8.

3 Networks approach

In order to answer the Auror challenge "Who is most likely to offend in my store now" the most important information in the dataset we can exploit might be 'who' has committed an incident at 'which' store? That corresponds to the so-called incidence/transaction matrix used in the Complex Networks community, and here we use the similarity/proximity approach to tackle this challenge. The rationale of this approach is that, if some factors have influence on the occurrence of the incidents, then (statistically) these factors must be encoded in the incidence matrix. In other words, we use the incidence matrix to capture the effect of various factors on the occurrence of the incidents even though we do not really know what these factors are.

3.1 Incidence matrix and similarity matrix

We consider a toy example. The incidence matrix for the situation

Person	store 1	store 2	store 3	store 4	store 5
A	1	1	0	0	0
B	0	1	1	0	0
C	0	0	0	1	1

(1)

is defined as

$$\mathrm{I} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}. \tag{2}$$

From this incidence matrix, we form a "similarity matrix" for the Persons:

$$\mathsf{S} := \mathrm{I} \times \mathrm{I}^{\mathsf{T}} = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{3}$$

From matrix S, we can say that the similarity between Persons A and B is 0.5, where A is 100% similar to itself, and the similarity between Persons A and C is zero. But what does this "similarity" mean? We interpret it here as they are similar in committing incidents in the same stores.

Similarly, we define the similarity matrix for stores:

$$\tilde{\mathsf{S}} := \mathrm{I}^{\mathsf{T}} \times \mathrm{I} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}. \tag{4}$$

We interpret this "similarity" as saying, for example, that Stores 4 and 5 are quite similar in the way they attract shoplifters, because they have suffered from the same offender.

3.2 Predictions

From the similarity matrix and incidence matrix, we can easily compute the likelihood of future incidents.

Predictions of the likelihood for offenders are calculated as

$$\mathsf{P}_1 := \mathsf{S} \times \mathsf{I} = \begin{bmatrix} 2 & 3 & 1 & 0 & 0 \\ 1 & 3 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \end{bmatrix}. \tag{5}$$

This means for Person A, the top two stores he would like to shoplift at next are Store 2 and Store 1. But for Person B, the top two stores are Store 2 and Store 3.

Similarly, we can compute the likelihood prediction for stores, which is simply the transpose of P_1,

$$\mathsf{P}_2 := \tilde{\mathsf{S}} \times \mathsf{I}^\mathsf{T} = \begin{bmatrix} 2 & 1 & 0 \\ 3 & 3 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 2 \end{bmatrix}. \tag{6}$$

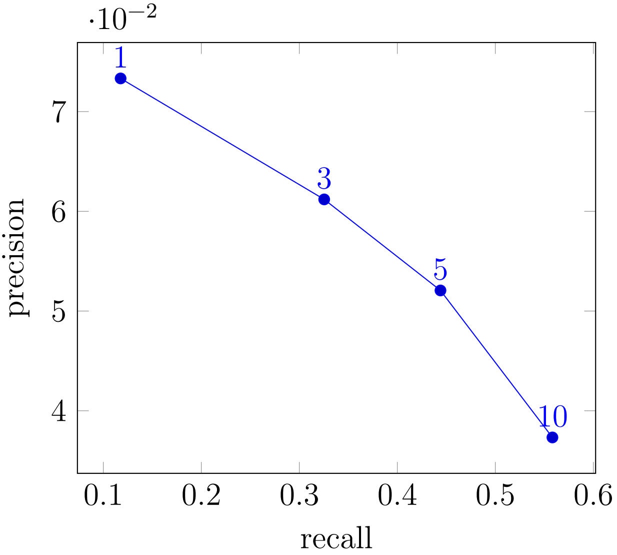
This means, for Store 2, Persons A and B have the same chance of shoplifting in this store. For Store 1, on the other hand, the most likely offender is Person A followed by Person B.

From this toy example, we can see that the offenders who have committed incidents only once do not contribute to the final prediction except for those stores where they have committed incidents. This is because for those people, the similarity matrix has non-zero entries only at diagonal positions.

3.3 Prediction results for the current dataset

Having removed the data for offenders with only one offence from the Auror dataset, this approach was tested using four-fold cross validation [10] for 80%

Figure 1: Precision-recall plot for Predictions for Offenders. The numbers $n = 1, 3, 5, 10$ in the plot mean that the predictions are the n Top Stores predictions.



training data (Prediction for Offenders). The evaluation is summarised in Figure 1. Readers not familiar with the concept of precision/recall are referred to [11]. From this diagram, we can see that, if we predict the Top Store (most likely to be victimised by a particular offender), then more than 10% of our predictions are correct, although only about 8% of the victim stores are included in our predictions. If we make the Top Ten store predictions (that is, we predict the ten stores that are most likely to be the victims of shoplifting by offenders), then only about 3% of our predictions are correct, but more than 55% of the stores in the test set are included in our predictions.

3.4 Summary of the networks approach

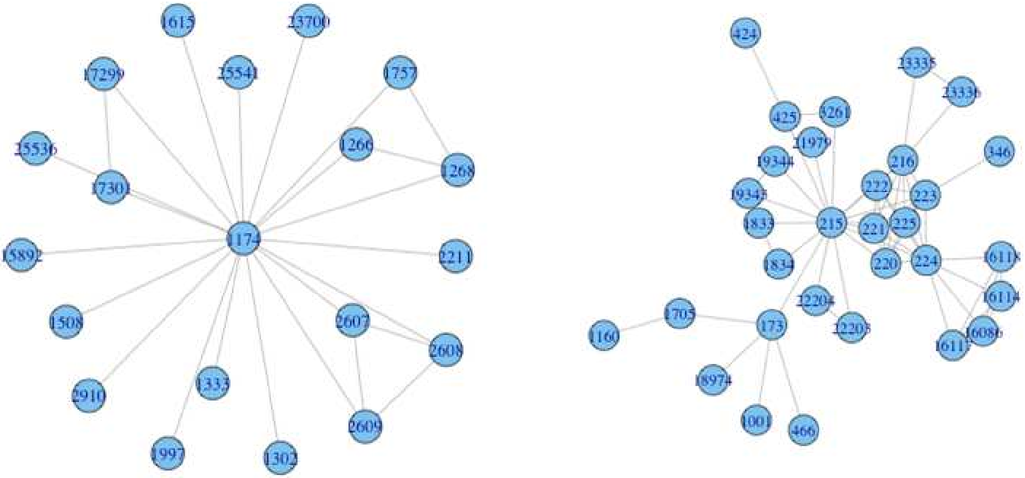
Advantages of this approach:

- The idea is simple. No mathematics other than matrix multiplication is used in the approach. The idea underpinning this approach is the basic idea in so-called personalised recommender systems, although here we are not making recommendations about the offenders.
- We can predict the Top N stores most likely to be hit by each offender, and predict the Top N most likely offenders for each store.
- The predictions are not very satisfactory, but they are reasonably good compared to other methods. In order to have a better prediction, this approach needs more data. The current dataset is only for less than two years and limited stores.

What we could do:

- This is only a preliminary version for the solution to the challenge set by Auror. It is a 'quick and dirty' approach. There are some other things one could try, in future developments, such as different similarity metrics. Here in this short note the similarity we used is the so-called symmetric metric similarity. One could introduce asymmetric metric symmetry with more parameters to tweak.
- The incidence matrix includes 1 for offenders who have committed incidents in some stores, no matter whether that happens once or ten times. We could incorporate the number of incidents into the incidence matrix to quantify much how the offender 'likes' a specific store.
- Currently we cannot predict new offenders. This corresponds to the 'cold start' problem in the Recommender Systems methodology [12]. Future work could examine further techniques used in that paradigm.

Figure 2: Two clusters of accomplices, from network analysis.



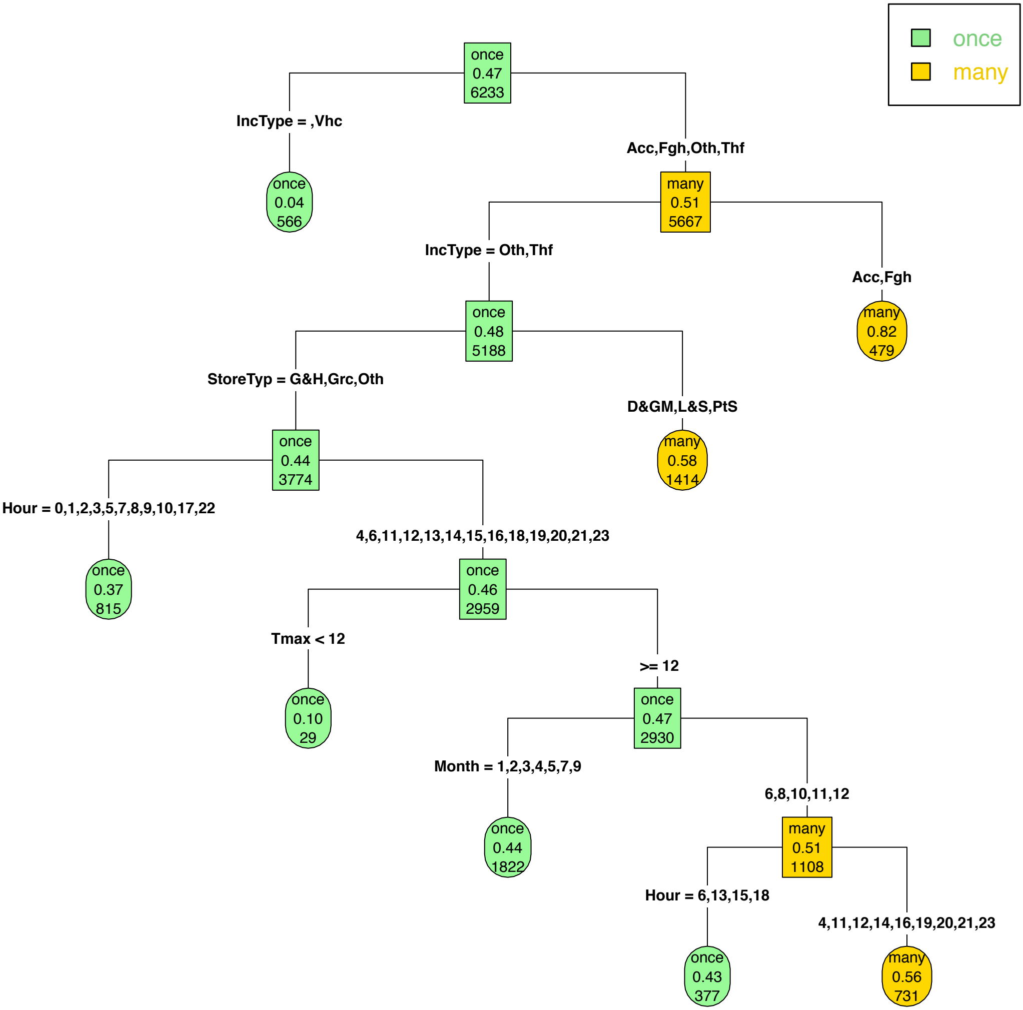
4 Network analysis of accomplices

A question of interest is the extent to which retail theft is a result of organised gang activity. Anecdotally many thefts, for example iphones or manuka honey, are carried out to order. The data had approximately 2700 pairs of individuals who were "connected", meaning they had been in at least one incident together. Social Network Analysis has frequently been used to study organised crime [13], and this methodology was applied to the Auror data, to find individuals who were linked together in groups. Overwhelmingly, these were groups of two or three people, but there were some groups that consisted of more people. Figure 2 shows two clusters (possibly criminal gangs) consisting of 20 and 31 individuals respectively. This brief exploration shows there is clearly scope for mining the relationships between individuals, and exploiting this knowledge for identifying likely offenders.

5 Classification tree for repeat versus one-time offenders

Patterns of behaviour for repeat offenders were investigated using a classification tree model based on the incident when the offender first appeared in the dataset. A binary response variable, "RepeatOffender", was defined with categories "once" for single offenders and "many" to represent individuals subsequently found to be repeat offenders. The data examined consisted of 3314 incidents involving single offenders and 2919 involving repeat offenders. The predictor variables used in the tree model included: day of the week; hour of the day; month of the year; whether the incident occurred during the school term; whether the incident occurred on a week day or at the weekend; whether the incident occurred during normal work hours 8 am–5 pm; the maximum temperature; store type; and incident type. To provide a tree which was interpretable, the number of incident categories was reduced by combining some categories. The classification tree model is shown in Figure 3. In the figure, squares represent nodes (groups) that are later split into subgroups while ellipses represent groups that were not subdivided. The first node indicates that, of 6233 offenders, the majority only appeared once in the dataset: 47% were repeat offenders. The first split separated off 566 individuals who were involved in incidents to do with vehicles: these individuals predominantly appeared only once in the dataset—only 4% were repeat offenders. The remaining individuals had incidents of: forbidden access, such as trespass (Acc); fighting or assault (Fgh); theft (Thf); or other (Oth), for instance a person of interest to the police. The second split carved off the 479 individuals with incidents of access or fighting, of whom 82% were repeat offenders in the database, so that this node predicts individuals are in the "many" category. The remaining individuals (5188, the majority) were then split by store type. Individuals whose incidents involved department/general merchandise stores (D&GM), Lifestyle and sports stores (L&S), and Petrol Stations (PtS) were more likely (58%) to be in the repeat offender category. Individuals whose incidents took place in Garden and Hardware stores (G&H),

Figure 3: Classification tree for repeat offenders.



Grocery stores (Grc) or other stores (Oth) were mostly only once represented in the dataset. The next splits were by hours of the day and temperature (maximum < 12 degrees celsius or $\geqslant 12$ degrees), but all these categories were dominated by single offenders. However for those individuals whose incidents were during the hours 4,6,11–16,18–21,23 (mostly late morning-early evening) *and* when/where daily temperatures exceeded 12 degrees, *and* if their first offence occurred in the months 6,8,10–12 then they were slightly more likely (51%) to be repeat offenders. This is unfortunately rather uninterpretable, especially as month and temperature are likely to be confounded with the growing spread of Auror into new cities, towns and businesses over the course of time. The final split again concerned hour of the day, and there was no obvious pattern.

In summary, then, this method shows that the type of incident and type of store were the most important features in reoffending. The individuals with incidents of forbidden access or fighting, are particularly likely to reoffend. Further data might suggest other specific characteristics associated with repeat offenders, but it will be important to group categories so that they are interpretable.

6 Linear score approach

Conceptually, the task can be thought of as finding a linear combination of individual characteristics, say

$$\eta_{ijt} = \beta_0 + \beta_1 x_{ijt} + \beta_2 y_{ijt} + \beta_3 z_{ijt} + \cdots , \tag{7}$$

for offender i at store j at time t. The covariates $x_{ijt}, y_{ijt}, z_{ijt}, \ldots$ represent characteristics of the individual, store, time, and perhaps interactions of these characteristics, for example that a certain individual i only attempts to shoplift in a particular store j at times t when the shop is relatively busy. We want to be able to predict the η_{ijt}, which we call the score, for each individual. However, we do not have much data on each individual, so to do any prediction we need to borrow information across a large group of individuals.

Suppose the coefficients β_k $(k = 0, 1, 2, \ldots)$ are estimated, somehow. Then, on substituting values for the covariates, one can estimate the η_{ijt}, and then for each store j and time t one can pick the Top Ten individuals i which have the ten largest η_{ijt}'s. So in principle this solves the problem.

The challenge is how to choose the variables to go in the linear combination, and how to estimate the β_k's. Equation (7) is suggestive of a regression approach but to solve it using that method we would need a "response" variable η_{ijt}. It is not obvious how to define this response. For example, we cannot model the probability an incident will be observed, when we only have data on individuals with observed incidents, but no data on the people not observed in incidents.

6.1 Conditioning on individuals already in the database

The approach considered here is to focus exclusively on individuals already in the database up until the end of a particular month, and to use as the response variable whether (or how often) they offend in the next month.

This approach, conditioning only on individuals already in the dataset, means that we cannot model the approximately 90% of events each month which are currently perpetrated by new offenders, that is, ones not already in the Auror database. We still use the new offenders' data, but only for predicting their future incidents, not for predicting their first appearance. Focussing on individuals already in the database means that we are arguing to the brief, which was to find ways to alert security staff to known likely offenders. It is likely that the percentage of new offenders will decrease in size as more and more individuals are entered into the database.

A disadvantage of focussing on individuals already in the database is that this model cannot be used for staff workload planning, such as deciding whether to put more security staff on duty at peak shoplifting times. However, it is possible to explore the database in simpler ways to find such peak incident times for a given type of store and particular suburb. The model used here is solely for ranking known individuals.

Expressing the data this way implies that for each individual in the database we have, as response, a count variable N_i = number of incidents for the ith individual in the next month. Alternatively we could recode the response variable as binary, with response = 1 meaning the individual is recorded in an incident in the next month, and response = 0 meaning no incident for that individual. The present discussion focuses on the count N_i since it does seem informative for ranking purposes to know if the individual has several incidents, rather than just one.

This approach puts the analysis into the purview of a generalised linear model (GLM) [14] but with excess zeros. For each individual included, suppose N_i is an observation from a zero-inflated Poisson (ZIP) distribution or zero-inflated negative binomial (ZINB) distribution. The zero-inflation refers to the fact that the vast majority of individuals in the dataset are *not* observed in an incident during the next month. A logistic regression model is specified for the probability that no incident is observed. A large probability means a large number of excess zeros. Setting aside these excess zeros, the remaining individuals are assumed to have the number of incidents given by a Poisson distribution (if incidents occur randomly, for example opportunistic crime where circumstances make it easy) or a negative binomial distribution if the counts of incidents are overdispersed relative to the Poisson. In the latter case, some individuals may be recidivist offenders which are observed with high frequency, particularly those who steal to order.

Specifically, for those individuals who are not among the excess zeros, we assume the mean of N_i is $\mu_i = e^{\eta_i}$ where

$$\eta_i = \beta_0 + \beta_1 x_i + \beta_2 y_i + \beta_3 z_i + \cdots . \qquad (8)$$

As it stands, the model (8) does not assume time-specific or store-specific information; these are discussed in the next paragraphs. For Poisson data the variance is the same as the mean, whereas for negative binomial data

$$\text{Variance} = \mu_i + \frac{\mu_i^2}{\theta} \tag{9}$$

for some overdispersion parameter θ. If θ is large, then the mean-variance relationship approaches that for a Poisson distribution, and the zero-inflated Poisson model can be used. In both models, not all zeros are excess zeros: some are zeros arising from the Poisson or negative binomial distribution with mean μ.

Looking at data on a monthly basis is convenient for proof of concept. In future it may be feasible to update the model using a shorter time scale, such as data up to the end of the last week, to predict offending the next week. However, *the frequency of model updating does not need to match the frequency of prediction.* The regression model (8) is needed only for estimating the regression coefficients β_k. Once the estimates $\hat{\beta}_k$ are calculated, they can be used to continuously update the linear combination (score)

$$\hat{\beta}_0 + \hat{\beta}_1 x_{ijt} + \hat{\beta}_2 y_{ijt} + \hat{\beta}_3 z_{ijt} + \cdots \tag{10}$$

for each new set of covariates $x_{ijt}, y_{ijt}, z_{ijt}, \ldots$ updated in real time. This is the way that time-specific information will be used. Dummy variables are used to model the different risks for specific stores or locations. The scores are then ranked to find the top ten suspects at any time.

It may not be possible to estimate all the βs in a single equation. For example, suppose parameters β_1 and β_2 are estimated from nationwide data. However, β_3 may only be estimated for a specific subset of the data, such as relating to the effect of distance between a specific location j_0 and the various incidents that an individual was involved in. Despite the fact that the coefficient $\hat{\beta}_3$ is based on this subset of the data, we might nonetheless use it heuristically for any store. That is, define distance to each store j for each individual i, transform this distance into a relevant covariate z_{ijt}, and then just use the term $\hat{\beta}_3 z_{ijt}$ to calculate the score (10) for each (person, store, time) combination. The $\hat{\beta}_3$ we use may not be the best number for a particular store, but it may be an adequate estimate for the purpose of ranking the offenders.

6.2 Illustrative example

In this subsection, rather than just predicting one month's offending, the last six months' offending data were used. That is, all individuals in the database before January 2015 were assessed for reoffending during that January; then all offenders before February 2015 were assessed for reoffending in February; and so on to June 2015. The response data for all six months was then stacked into a single response variable, and the corresponding rows of predictor data stacked into a single set of covariates. This pretends that all responses (rows) are independent, which we know is untrue: the same individual may have up to six responses in the regression, and these responses are correlated. However, this is just for illustration, and hopefully the correlation is not very important in view of the large amounts of data and the other simplifying assumptions that need to be made.

Table 1 summarises incident data for the six months January–June 2015. The first line shows the number of separate (person, incident) combinations for the six months prior to MINZ. If the database contained several lines for the same person and incident (for example, several different items listed as being stolen), then they were counted only once in Table 1. The numbers are gradually rising as the Auror system becomes more widely used, but June's data was for slightly less than a whole month. The bulk (about 90%) of incidents were perpetrated by people not already in the database, so the total number of suspects (people identified from previous months) is rapidly increasing in line 2 of Table 1. Line 3 of the table shows the total number of incidents committed by previously-identified suspects, which is about 10%

Table 1: Numbers of suspects and incidents recorded by Auror in January–June 2015.

	Month	Jan	Feb	Mar	Apr	May	Jun
1	Number of separate person+incidents	449	431	593	694	815	760
2	Cumulative total suspects from database	2026	2427	2780	3280	3869	4574
3	Total count of incidents by suspects	45	51	57	64	75	59
4	Proportion of all incidents by suspects	0.100	0.118	0.096	0.092	0.092	0.078
5	Number of suspects reoffending	25	41	32	44	48	39
6	Proportion of suspects found reoffending	0.012	0.017	0.012	0.013	0.012	0.009

of the incidents each month (line 4). The proportions in line 4 are currently decreasing as the Auror system spreads to new cities and towns, thus finding more first offenders, but this trend should reverse with a longer history of suspects. Line 5 shows the number of suspects detected to be reoffending at all (irrespective of how often). Most reoffenders were only detected once, but a few were detected several times in a month. Only about 1% of suspects were detected as reoffenders in any given month.

Predictors considered for modelling the response were: the offender's gender; the offender's age group; the earliest time of day at which an offence was previously recorded; the proportion of previous incidents that took place during school terms; the number of previous offences by the individual at Grocery stores, Garden centre/hardware stores, Petrol stations, Department stores, Sport/lifestyle stores, or Other stores, respectively; the proportion of previous offences during work hours (8 am–5 pm); the number of previous incidents the individual has been identified in; the locations of these incidents; and the number of days between the individual's latest incident and the fifteenth day of the month being examined. In addition, information was considered about the offender's known accomplices. Accomplices were identified as people with the same incident number. The information included how many accomplices the offender had, how many incidents (in total) accomplices had been involved in, the number of days since the last-recorded incident by an accomplice, and the locations of incidents where the accomplice had been identified.

The small count sizes are a reason for combining the analysis over six months of data. Even so, across the whole database there were only 229 positive responses (at least one reoffence) out of a cumulative total of 18956 rows of data. Intuitively, the number of events seems rather too sparse for predicting many significant variables, but also the cumulative number of 'suspects' is large which would usually mean models with very small p-values. It would be a matter for future research to see how these conflicting intuitions can be resolved. For the moment we simply point out that the p-values produced by this method should be treated with caution until there is rather more data. The results may be informative nonetheless, given that the aim is to rank suspects rather than to estimate probabilities with precision.

If the data were accumulated across thirteen weeks instead of six months, then the number of rows with positive responses (at least one incident in that period) would increase a little, but the cumulative number of zeros (no detected incident in that week) would increase dramatically. Whether this would make much difference to the model is a matter for further research. The question here is whether the $\hat{\beta}_k$'s would change much by estimating based on weeks rather than months. But in either case the linear score (10) can be updated continuously to include new individuals and new events.

The R output in Table 2 uses the zeroinfl package to model the number of incidents per month for each suspect, using the nationwide Auror

Table 2: Modelling nationwide incidents per suspect.

```
zeroinfl(formula = cnumincidentspersuspect ~ propGardnHardware +
propPetrolStation + propDeptStore + propLifeStyleSports +
PropOther + log(SusTimeSince) + log(Susnpriorincidents) |
   log(SusTimeSince) + log(Susnpriorincidents), dist = "negbin")

Count model coefficients (negbin with log link):
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)               1.2711     0.6124   2.076  0.03794 *
propGardnHardware         1.2970     0.3951   3.283  0.00103 **
propPetrolStation        -1.2136     0.4228  -2.871  0.00410 **
propDeptStore             0.1197     0.2315   0.517  0.60502
propLifeStyleSports       0.3933     0.2208   1.781  0.07488 .
PropOther                 1.3885     0.5327   2.606  0.00915 **
log(SusTimeSince)        -0.7010     0.1634  -4.289 1.79e-05 ***
log(Susnpriorincidents)   0.3795     0.1174   3.233  0.00123 **
Log(theta)               -0.2645     0.4467  -0.592  0.55384

Zero-inflation model coefficients (binomial with logit link):
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)               0.8929     0.6828   1.308   0.1909
log(SusTimeSince)         0.4595     0.1792   2.564   0.0103 *
log(Susnpriorincidents)  -1.8428     0.2765  -6.666 2.63e-11 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Theta = 0.7676
Log-likelihood: -1199 on 12 Df
```

data. The probability of excess zeros (no incidents) is estimated to increase with log(SusTimeSince) (the logarithm of the days between the suspect's last incident and the fifteenth day of the month), and to decrease with log(Susnpriorincidents) (the logarithm of the total number of incidents for that suspect prior to that month). The number of prior incidents is the more important effect, judging by the p-value. On the other hand if at least one incident occurs, then the number of incidents is likely to be greater if the latest incident was recent (negative coefficient for log(SusTimeSince)) and if the offender had been involved in many prior incidents. The type of store for previous incidents was also informative. Grocery stores are taken as the 'baseline' category. Suspects with a high proportion of their previous offences at a garden centre or hardware shop, or other shop such as jewellers, were significantly more likely to have repeat incidents (higher mean number of incidents this month) compared to those whose offences were solely at grocery stores. Suspects who offended at a petrol station were less likely to have extra incidents compared to those who solely offended at grocery stores. There was no significant difference in the mean number of incidents for those who offended at department stores or lifestyle/sporting goods stores compared to those offending at grocery stores. The θ parameter is small, indicating considerable overdispersion.

No demographic variables (such as gender or age) or time of day variables were found to be significantly related to the mean number of offences. The data used in Table 2 included offences from several different cities/towns of New Zealand. Neither the city, nor season, nor daily temperature, were taken into account in Table 2 since the effect of these would be confounded with the limited coverage of the Auror database up to the time of MINZ.

6.3 Incorporating information about accomplices

The role of accomplices was considered in Table 3 (R output). It was more significant, in terms of zero-inflation, to use the minimum time between the fifteenth day of the month and *either* the suspect's own latest incident or that of one of their known accomplices (variable log(SusorAccTimeSince)). The longer this time was, the less likely an incident will be observed in the month. On the other hand, the variable Suspropnbyself refers to the proportion of times the suspect was alone at the time of the incident, that is, not accompanied. The negative coefficient means that if a suspect usually worked *alone* then it was estimated as more likely for an incident to occur; however, this effect did not carry over to greater numbers of incidents per month. The AIC (Akaike information criterion [15]) shows that the model in Table 3 fits considerably better than that in Table 2. The number of prior incidents by the offender remained significant.

It can be computationally intensive to include information about accomplices in a predictive model. The present analysis used an extra loop through the data for each suspect, to look for accomplices, which is very inefficient, and a working program for a large database of suspects would need a much more efficient data structure. However, this analysis does provide proof of concept that it is possible to include such information, and that it is worthwhile taking the trouble to do so.

6.4 Analysis incorporating distance

An obvious factor relating to crime is the distance between a store and the location(s) where a suspect has been previously detected. At the moment we cannot calculate distance relationships for each individual store: the number of incidents at any store is too few to give enough cases for generating a regression model that one could have any confidence in. Progress is possible nonetheless. The key idea is that we just need some way of estimating an approximate term, say

$$\beta_d f(\text{distance}_{ij}), \tag{11}$$

to be included in the linear score (10). To compute such an approximation one must decide: firstly, how to measure distance between individual i and store j;

Table 3: Modelling nationwide incidents per suspect, including information on accomplices.

```
zeroinfl(formula = cnumincidentspersuspect ~ propGardnHardware +
      propPetrolStation + propDeptStore + propLifeStyleSports +
      PropOther + log(SusorAccTimeSince) + log(Susnpriorincidents)
      | Suspropnbyself + log(SusorAccTimeSince)
      + log(Susnpriorincidents), dist = "negbin")

Count model coefficients (negbin with log link):
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -2.00701    0.31755  -6.320 2.61e-10 ***
propGardnHardware        1.03391    0.38212   2.706  0.00682 **
propPetrolStation       -1.17774    0.41657  -2.827  0.00470 **
propDeptStore            0.23970    0.22799   1.051  0.29309
propLifeStyleSports      0.53057    0.21263   2.495  0.01259 *
PropOther                0.75210    0.49630   1.515  0.12967
log(SusorAccTimeSince)   0.11701    0.06705   1.745  0.08097 .
log(Susnpriorincidents)  0.81460    0.13052   6.241 4.34e-10 ***
Log(theta)              -0.07972    0.38758  -0.206  0.83704

Zero-inflation model coefficients (binomial with logit link):
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.1090     0.6867  -1.615  0.10633
Suspropnbyself           -0.5760     0.1856  -3.103  0.00192 **
log(SusorAccTimeSince)    1.0770     0.1360   7.917 2.43e-15 ***
log(Susnpriorincidents)  -1.0566     0.1561  -6.769 1.30e-11 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Theta = 0.9234
Log-likelihood: -1180 on 13 Df
> AIC( zinbmodel1, zinbmodel2)
           df      AIC
zinbmodel1 12 2421.011
zinbmodel2 13 2386.185
```

secondly, how to functionally relate distance to the increased expected number of incidents at a given location; and thirdly, what number to use for the multiplier β_d. Once these three facts are established, at least approximately, then the term (11) may be included in the linear score even for stores that have not yet had any incidents.

The distance_{ij} (in km) between store j and the location of a prior incident for individual i can be estimated using geographical coordinates

$$\text{distance}_{ij} = 106.5 \times \sqrt{(\text{lat}_i - \text{lat}_j)^2 + (\text{long}_i - \text{long}_j)^2} \qquad (12)$$

where 'lat' and 'long' are the latitude and longitude of the locations (in degrees). It would make more sense to use road distance or walking distance, but latitude and longitude were the measures available. It is debateable whether to use the minimum distance between the store and all prior incidents, or the mean or median distance to prior incidents, or the distance to the most recent incident. If a repeat offender usually resides in one city, but happens to commit an offence when visiting a different city close to store j, then it is not clear which measure would be the most appropriate: this is something for future investigation.

For the function $f()$ relating distance to risk, it would seem sensible to use some curve that exaggerates small distances but treats all large distances as being pretty much the same. Two possibilities were investigated:

$$f_1(\text{distance}) = \log(1 + \text{distance}),$$

and

$$f_2(\text{distance}) = D_0/(D_0 + \text{distance}),$$

for some constant D_0. Here $f_1(0) = 0$, and a negative coefficient β_d ensures the risk decreases with distance. For $f_2()$, if $D_0 = 5$, say, and $\beta_d > 0$, then this implies that when the nearest incident was 5 km away the risk is half that of when the nearest incident is 0 km away (and the risk is one quarter when the nearest incident is 15 km away, etc.) A suitable choice of D_0 could be found by trial and error.

6.5 Example: Using Pukekohe data to explore the effect of distance

Rather than modelling distance to a particular store, this example considers distance to the centroid of incidents in Pukekohe, a town in south Auckland with a relatively large number of incidents. Pukekohe had 439 data lines in the six months from January to June 2015. Some of these data lines refer to different items taken during the same incident, so eliminating these supplementary lines reduces the data to 308 separate (person, incident) combinations. For these, there were 264 separate individuals, of which 242 individuals (92%) had never appeared before in the database. The number of separate incidents was 258, in which the individuals were working alone 215 times (83%), with one accomplice 37 times (14%), with two accomplices 5 times, and with three accomplices once. Altogether there were only 33 (suspect, incident) combinations which involved people from earlier in the database (20 individuals in all, perpetrating 7.6% of the incidents). This small number of cases means there is limited scope for finding significant predictors, despite there being thousands of potential suspects in the database.

In Table 4, distance is measured from the centroid of all incidents in the database that were observed in Pukekohe. The variable Susminlog1DistSuburb refers to the minimum log(1+distance) from a suspect's locations of offending to this Pukekohe centroid. The response variable is solely the number of incidents in Pukekohe. The output shows that the probability of zero incidents increases with the distance from suspect's previous offending, increases with the time since either the suspect or an accomplice last had an incident, and decreases when the suspect has more prior offences. Conversely, if it is not an excess zero, then the mean number of offences this month decreases with distance and increases with the number of prior incidents. This agrees with intuition. On the other hand, the mean increases if there has been a big time gap since the offender's (or accomplice's) last incident. Perhaps this indicates suspects keeping a low profile in Pukekohe for a while, then hitting the shops when they think they have been forgotten: but there may be other

Table 4: Modelling incidents per suspect in Pukekohe.

```
zeroinfl(formula = cnumincidentspersuspect
~ log(SusorAccTimeSince) + log(Susnpriorincidents)
+ Susminlog1Suburb   )

Count model coefficients (poisson with log link):
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              -2.0248     1.0027  -2.019  0.04346 *
log(SusorAccTimeSince)    0.6454     0.2343   2.755  0.00587 **
log(Susnpriorincidents)   0.5239     0.3163   1.656  0.09764 .
Susminlog1DistSuburb     -0.7141     0.3545  -2.014  0.04396 *

Zero-inflation model coefficients (binomial with logit link):
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.8723     1.5080  -0.578  0.56297
log(SusorAccTimeSince)    1.0889     0.3921   2.777  0.00548 **
log(Susnpriorincidents)  -0.8533     0.3386  -2.521  0.01172 *
Susminlog1DistSuburb      0.7557     0.3763   2.008  0.04465 *
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Log-likelihood: -118.9 on 8 Df
```

explanations. Table 4 shows a ZIP analysis rather than a ZINB one. A ZINB model was fitted but it had a large $\theta = 3.27$, meaning the overdispersion in the mean-variance relationship (9) is not great. The AIC for the ZIP was 253.74 on 8 df, which was better than for the ZINB with the same predictor variables (AIC = 255.27 on 9 df). Although the coefficient for log(Susnpriorincidents) is not significant in the ZIP model for the counts, it is retained both because it makes sense and because the AIC with this variable included is slightly lower than the AIC with this variable dropped.

6.6 Identifying top suspects

For illustration, we consider using the models from Tables 3 and 4 to identify the Top Twenty likely offenders for the last six months of data, both nationwide and using Pukekohe-only data. The predicted values (Scores, η_{ij}) for each model were saved, and the twenty individuals with the highest scores are shown in Tables 5 and 6 respectively.

Table 5 shows the individuals with the Top Twenty scores nationwide. Individual 1174 was prolific, with 22 recorded incidents before 1 January 2015, and a further 22 within the next six months. This was a male who frequently attempted to shoplift meat and other goods, all over suburban Auckland. However, this individual never visited Pukekohe, never coming closer than Papakura, the southernmost piece of continuous suburbia in Auckland city. The seventh and eighth ranked individuals had no recorded incidents prior to 1 January 2015, but eleven in the following six months, so their more recent events cause them to be ranked higher than, say, the 14th individual, who had fifteen incidents before 1 January 2015 but none more recently.

Figure 4 shows the number of incidents in the six months, for the top-scoring 100 individuals, plotted against their score according to the model in Table 3. The dashed line separates off the ten top-ranked offenders (right) from the others. The graph shows that the score does not predict all the high-frequency offenders, but on the other hand most of the Top Ten *are* repeat offenders, which suggests the model is indeed beginning to do what it is supposed to do: highlight persons to watch out for.

Table 6, using the ZIP model of Table 4 based on Pukekohe-only incidents, ranks the individual 22098 highest, due to his high frequency of incidents in Pukekohe during the six months. The prolific offender 1174 is ranked eleventh, and is the only high-ranking individual to never visit Pukekohe. Figure 5 shows the number of incidents within Pukekohe, versus the score based on Table 4 for the 100 highest-scoring individuals. The predicted Top Ten individuals are to the right of the dashed line. Again, the score seems to

Table 5: Individuals with the Top Twenty Scores nationwide, calculated without taking distance or suburb into account (model in Table 3); observed number of events in database; observed number during last six months nationwide and in Pukekohe, and minimum distance from suspects' incidents to central Pukekohe.

Rank	Suspect ID number	Score	Total incidents before six months	Incidents in six months nationwide	Incidents in six months in Pukekohe	Minimum distance to central Pukekohe
1	1174	4.13	22	22	0	15.4
2	197	1.80	13	9	0	31.3
3	14518	1.42	19	0	0	342.5
4	202	1.41	9	9	1	1.0
5	215	1.05	9	3	0	15.7
6	12978	1.00	4	5	0	418.5
7	22098	0.96	0	11	11	0.6
8	23540	0.95	0	11	0	16.5
9	2389	0.79	3	4	0	717.5
10	1135	0.78	7	7	1	0.6
11	391	0.75	1	9	6	0.6
12	1746	0.75	5	0	0	15.7
13	21843	0.66	2	4	0	32.8
14	2996	0.65	15	0	0	48.8
15	2793	0.64	5	2	0	715.1
16	1136	0.63	6	1	0	0.6
17	3172	0.63	5	4	0	700.8
18	173	0.63	6	1	0	29.0
19	16047	0.61	1	4	0	324.8
20	14548	0.57	2	0	0	322.7

Figure 4: Number of incidents detected nationwide, in last six months, versus prediction score without taking distance into account: top-scoring 100 individuals.

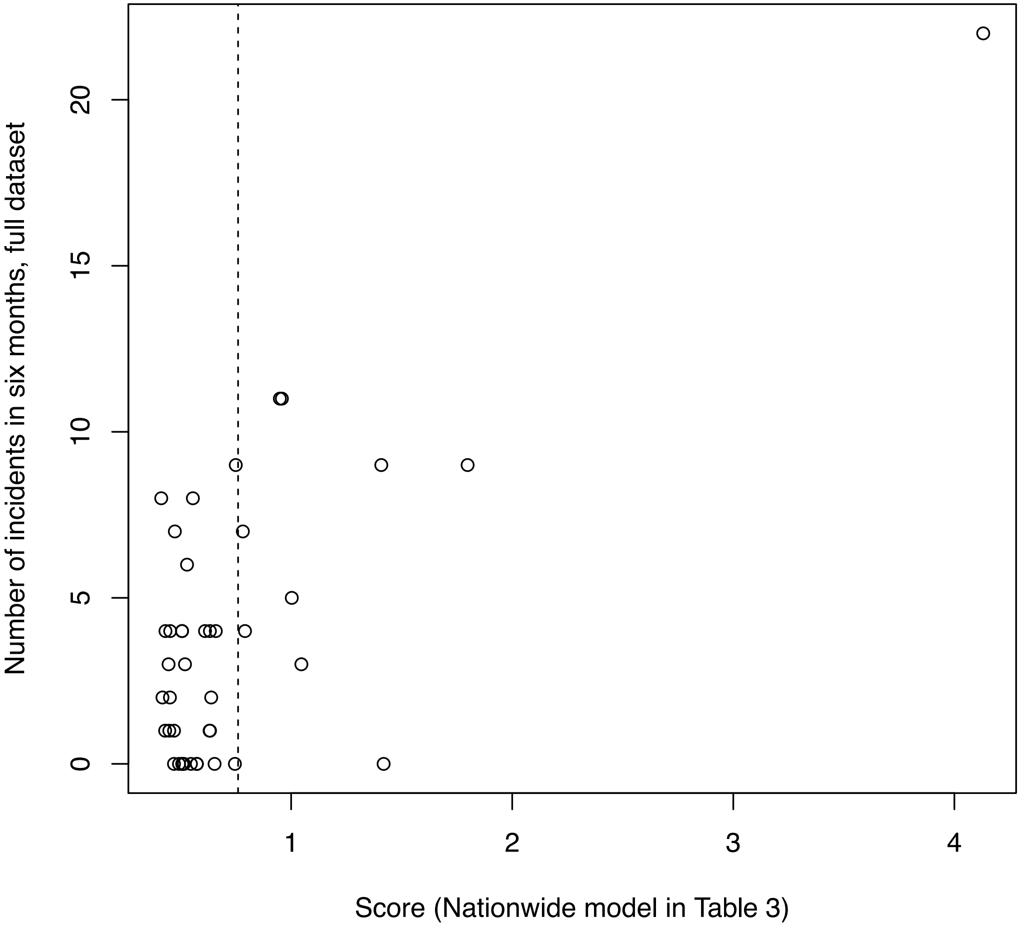


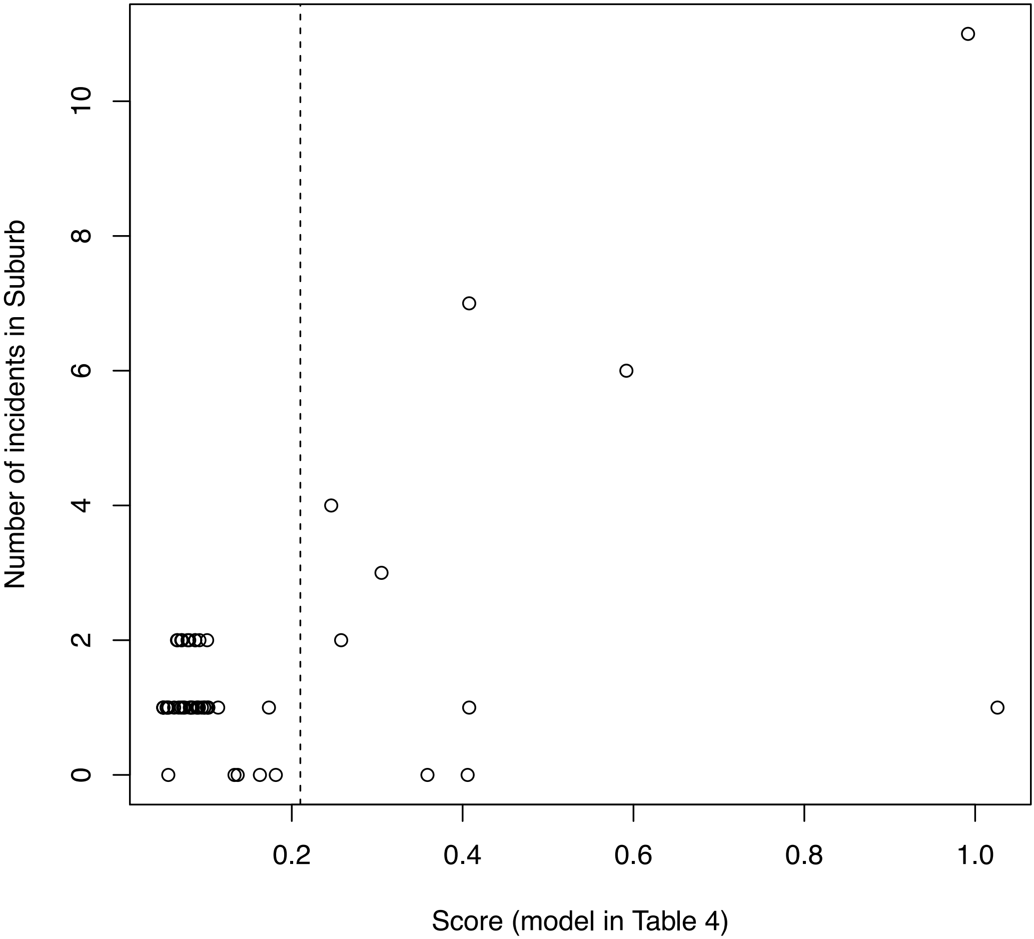
Table 6: Individuals with the Top Twenty Scores calculated using the model in Table 4; observed number of events prior to six months; observed number during last six months nationwide and in Pukekohe, and minimum distance from suspects' incidents to centroid of Pukekohe incidents.

Rank	Suspect ID number	Score	Total incidents before six months	Incidents in six months nationwide	Incidents in six months in Pukekohe	Minimum distance to central Pukekohe
1	22098	0.77	0	11	11	0.6
2	1135	0.72	7	7	1	0.6
3	391	0.60	1	9	6	0.6
4	23305	0.42	0	8	1	0.6
5	27032	0.42	0	7	7	0.6
6	1136	0.38	6	1	0	0.6
7	22047	0.30	1	4	3	0.6
8	20783	0.26	1	3	2	0.6
9	22111	0.22	0	4	4	0.6
10	223	0.19	2	0	0	0.5
11	1174	0.18	22	22	0	15.4
12	25351	0.16	0	3	1	0.4
13	19037	0.10	0	2	2	0.4
14	21918	0.10	0	2	2	0.6
15	23595	0.09	0	2	2	0.6
16	24816	0.09	0	2	2	0.6
17	23027	0.09	0	1	1	0.6
18	25493	0.09	0	2	2	0.6
19	23379	0.08	0	2	2	0.6
20	26119	0.08	0	2	2	0.6

Table 7: Individuals with Top Twenty scores based on Table 3 but adjusted for distance from Pukekohe; total number of incidents by suspect in database; number of incidents by suspect in Pukekohe during last six months; and minimum distance of suspects' incidents from the centre of Pukekohe.

Rank	Suspect ID number	Score	Total incidents before six months	Incidents in six months nationwide	Incidents in six months in Pukekohe	Minimum distance to central Pukekohe
1	1174	0.18	22	22	0	15.4
2	22098	0.77	0	11	11	0.6
3	1136	0.38	6	1	0	0.6
4	391	0.60	1	9	6	0.6
5	1135	0.72	7	7	1	0.6
6	22047	0.30	1	4	3	0.6
7	23305	0.42	0	8	1	0.6
8	27032	0.42	0	7	7	0.6
9	223	0.19	2	0	0	0.5
10	22111	0.22	0	4	4	0.6
11	346	-0.06	1	0	0	0.5
12	17228	-0.06	1	0	0	0.4
13	15975	-0.07	1	0	0	0.4
14	23027	-0.07	0	1	1	0.6
15	15894	-0.07	1	0	0	0.4
16	23030	-0.09	0	1	1	0.6
17	20783	-0.09	1	3	2	0.6
18	22225	-0.10	0	1	1	0.6
19	22167	-0.10	0	1	1	0.6
20	21918	-0.11	0	2	2	0.6

Figure 5: Number of incidents detected in Suburb (Pukekohe), in the last six months, versus prediction score taking distance, time since last incident, and number of incidents up to preceding month, into account: top-scoring 100 individuals.



be doing a reasonable job of highlighting individuals who have a high number of incidents.

6.7 Adjusting the nationwide score for distance

Now suppose we want to use a nationwide score, such as from Table 3, but adjust it to incorporate information about the distance from an individual's location to a shop or suburb's location. This simulates the situation where we do not have enough specific information about offences in a particular shop or suburb to build up a specific database of local offenders, but instead need to use the nationwide database. Here we use the nationwide database to predict likely offenders for Pukekohe. We compare the ranked results to the results based solely on offenders in Pukekohe (Table 4).

Since the coefficients for log(Susminlog1DistSuburb) in Table 4 have modulus 0.71–0.75 a simple approach is to calculate a modified score:

$$\text{Modified Score} = \text{Score(Table 4)} - 0.75 \times \text{Minlog1DistSuburb}\,. \tag{13}$$

Table 7 shows the Top Twenty results for this modified score. This Top Twenty includes thirteen of the Top Twenty suspects from the Pukekohe-specific model in Table 4, and 41 incidents within Pukekohe. This compares to 51 incidents among the Top Twenty in Table 6, using Pukekohe information only. Although this is only one example, it is pleasing that the modified score has netted 80% as many incidents as were netted by a model tailored to the local data alone.

There is some doubt whether the multiplier 0.75 in equation (13) is the best one to use. Table 8 explores the effect of varying the multiplier for Minlog1DistSuburb in equation (13) from 0.75 to some other number. It shows, for instance, for a multiplier of 0.25, there were 27 incidents in Pukekohe involving Top Ten suspects according to equation (13), and 39 incidents among the Top Twenty suspects. This compares to 35 (respectively 51) incidents among the ten (respectively twenty) highest-scoring offenders in Table 6, where the regression model was based exclusively on Pukekohe data. So 0.75 gave the best results here, but Table 8 shows similar numbers of incidents being detected for multipliers between 0.5 and 1.0. Future research could perhaps use trial and error to choose a satisfactory multiplier for log distance to a particular store or suburb. The trial and error could proceed by replicating Table 8 to find the multiplier which gives the maximum number of incidents at that location. This strategy does not even require fitting a model based solely on data from that location.

In conclusion, this example suggests that it is a practical strategy to use a regression model based on nationwide data (as in Table 3) to identify variables predictive of high levels of repeat offending. The regression coefficients of this model are extracted, and applied to the latest data from each individual to obtain that individual's score (ignoring distance). This score is then modified by, say, $-0.75 \times \log(1 + \text{distance})$, where distance is measured between any given store and the individual suspect. Distance for the individual could be based the last known sighting of that person or their accomplice or their vehicle (if recent, say in the last eight hours) or otherwise based on the minimum distance the individual has been from the store. In this way recent information on distance can be continuously incorporated into the model to find the top likely offenders at a particular store.

7 Other suggestions

7.1 Time to event method

The ZIP and ZINB regression models in Section 6 looked at how often an individual reoffended during the specified month(s). Another suggestion made for the Auror analysis was to treat the problem as time-to-event data, with the "event" being a detected offence. The objective would be to model the relative hazard of reoffending, with time being measured between offences or

Table 8: Number of Pukekohe incidents among those scored in the Top Ten and Top Twenty, using score for Table 3 adjusted for distance, by multiplier used in adjustment.

Multiplier	Number of Incidents in Pukekohe (Top Ten)	Number of Incidents in Pukekohe (Top twenty)
0	13	19
0.1	19	30
0.25	27	39
0.50	29	39
0.75	33	41
1.00	33	40
1.25	26	40
1.50	22	40

until the observation is censored at the time of computation.

This approach would fit the regression into the context of survival analysis [16], such as is used for modelling time until mechanical failure, and volcanic eruption. It might be fairly complex modelling, but may be worth considering as an option for these data.

7.2 Baseline comparison

In order to know whether a new method is doing better than an old method, one needs to have some idea how the old method performs. The method currently used by Auror was not disclosed, but as a baseline we considered creating our own list of Top Ten offenders (for most offences) for each suburb, based on all but the last three months of data. On average, 3.2% of offences in a suburb during these last three months were perpetrated by a Top Ten offender, as identified from previous data. The suburb with the highest proportion of offences carried out by a previous Top Ten offender was Pakuranga, at 66%, but this was based on only four out of a total of six offences in the three months. Many suburbs had 0% accuracy, due to no offences having been perpetrated by previous Top Ten offenders. Large suburbs (with over 100 incidents) generally had low accuracy (0–5%) but Pukekohe with 100 incidents had 19% of offences carried out by previous Top Ten offenders. Again, this may be because it is a separate town some 20 km distant from the nearest other suburbs of Auckland.

This baseline analysis may be a useful way of comparing the predictive ability of other proposed methods.

7.3 Naïve Bayes

Similar to the baseline comparison, a Naïve Bayes analysis (sometimes called "Idiot Bayes" [17]) was attempted, to try to predict reoffending on a suburb by suburb basis, during the last three months of data. Naïve Bayes looks at incidents (prior to those three months) which were/were not reoffences, and the rates at which various background factors or covariates occur. For example, what are: the rates of male/female for reoffences/new individuals; the rates of workhour versus non-workhour incidents for reoffences/new individuals; or the rates of offence at different types of store. The method assumes the factors are conditionally independent of each other, given the status of being from reoffences or new individuals. Then Bayes' theorem is used to estimate the probability that an incident with certain background factors is a reoffence. These probabilities are ranked, to find the incidents most likely to be reoffences, and the individuals involved in these incidents are identified. However, just because an incident has a high chance of being a reoffence does not necessarily mean that the perpetrator is a top reoffender, so there is a mismatch here between the goal and the method. Nevertheless, it is a way of picking out ten individuals for each suburb, by selecting the ten individuals with the highest ranked probabilities. The efficacy of the Naïve Bayes approach is measured by looking at the incidents in the next three months in the suburb, and finding the proportion which were carried out by the selected ten individuals. Disappointingly, the Naïve Bayes results had less than 1% accuracy on average, less than the baseline approach. This may be because of the mismatch between goal and method. The result suggests the Naïve Bayes approach is not worth pursuing.

8 Concluding remarks

The methods described in this paper are a work in progress. Since the different approaches are all picking up on different features of the data, it is likely that the best predictions will come from combining results across different approaches. In the end, it may be less important to have a precise formula for ranking, than to be able to update results in a timely manner. The linear score approach of Section 6 may be the best way forward in this regard, as once coefficients are determined, the linear score can be continuously recomputed whenever new data is entered, and the Top Ten list updated for any given store. The difficulty will be to choose appropriate covariates and estimate the coefficients. But as illustrated above, the terms used in the linear score do not all need to come from the same model, or even the same type of analysis.

Work needs to be done on how to validate the different methods, and if necessary combine and tweak them, by seeing how well they predict ‘Who is most likely to offend in my store now’. The baseline analysis (Subsection 7.2) and the cross-validation used in Subsection 3.3 suggest looking at the incidents that occur in a window of time, and finding the proportion of these incidents that involve a previously selected Top Ten individual. However, whether this proportion is to be reported on a store-by-store basis, or suburb-by-suburb basis, or nationwide, is something still to be considered. To report on a store-by-store or suburb-by-suburb basis will require much more data.

In conclusion the study suggests that there are several approaches that have good potential for alerting stores, or police investigators, to likely offenders, and thus fulfilling Auror's goal.

Acknowledgements Thank you to Auror for presenting this problem, and particularly Auror staff J'aime Laurenson and James Choi for their tireless assistance. Thank you to Penny Bilton for Section 5 and for the suggestion in Subsection 7.1, to Golbon Zakari for the results in Section 4, to Zoë Williams for Subsection 7.2, and to Mat Pawley and Zoë for the results in Subsection 7.3, and thank you to others who participated.

References

[1] http://www.auror.co/we-have-exciting-news/ (Accessed 9 Apr 2017). M291

[2] http://www.stuff.co.nz/national/crime/9806920/High-tech-blitz-on-2m-a-day-shoplifters (Accessed 9 Apr 2017). M291

[3] http://theregister.co.nz/news/2015/06/retail-crimewatch (Accessed 9 Apr 2017). M291, M292

[4] http://www.saynotoshoplifting.org/the-issue-why-care (Accessed 9 Apr 2017). M291

[5] http://www.auror.co/the-global-picture-of-shoplifting/ (Accessed 9 Apr 2017). M291

[6] http://www.scoop.co.nz/stories/BU1410/S00104/connecting-the-dots-on-retail-theft.htm (Accessed 9 Apr 2017). M293

[7] http://www.auror.co/two-at-a-time-is-great-prevention/ (Accessed 9 Apr 2017). M293

[8] http://www.auror.co/collaboration_success/ (Accessed 9 Apr 2017). M293

[9] http://www.innovators.org.nz/winners-a-finalists/winners-2015 (Accessed 9 Apr 2017) M293

[10] Seni, G. & Elder J. F. (2010) *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions* Morgan & Claypool, California. doi:10.2200/S00240ED1V01Y200912DMK002 M298

[11] Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies* **2 (1)** 37–63. M299

[12] Ricci F., Rokach, L. & Shapira, B. (2015) *Recommender Systems Handbook*. Springer, New York. ISBN 978-1-4899-7637-6 M300

[13] McGloin J. M. & Kirk D. S. (2010) An Overview of Social Network Analysis *Journal of Criminal Justice Education* **21 (2)** 169–181 doi:10.1080/10511251003693694 M301

[14] McCullagh, P. & Nelder, J. A. (1989) *Generalized Linear Models (second edition)* Chapman and Hall/CRC London. ISBN 9780412317606 M306

[15] Akaike, H. (1974), A new look at the statistical model identification *IEEE Transactions on Automatic Control* **19 (6)** 716–723. doi:10.1109/TAC.1974.1100705 M313

[16] Moore, D. F. (2016) *Applied Survival Analysis Using R* Springer International, Switzerland. eBook ISBN 9783319312453 M326

[17] Hand, D. J. & Yu K. (2001) Idiot's Bayes: Not So Stupid after All? *International Statistical Review* **69** 385–398 M327

Author addresses

1. **B. W. McDonald**, Institute of Natural and Mathematical Sciences, Massey University, Auckland 0632, New Zealand.
   orcid:0000-0002-8954-3313

2. **L. M. Hall**, School of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch 8041 New Zealand.
   mailto:lmh125@uclive.ac.nz

3. **X. P. Zhang**, Computational and Data Sciences, Callaghan Innovation, Lower Hutt 5040, New Zealand.
   mailto:philip.zhang@callaghaninnovation.govt.nz