

Faculty of Actuarial Science and Statistics

A Systematic Approach for the Analysis
of Health and Social risk at
Neighbourhood Level

Dr Les Mayhew

Actuarial Research Paper No. 144

December 2002

ISBN 1 901615 66 9

Cass Business School

106 Bunhill Row

London EC1Y 8TZ

T + 44 (0)20 7040 8470

www.cass.city.ac.uk

Lesmayhew@blueyonder.co.uk

A systematic approach for the analysis of health and social risk at neighbourhood level

Abstract

Poor health and social malaise often go together and are more likely to occur in deprived rather than rich areas. Official data sources make it difficult to analyze health inequalities at an appropriate geographical scale or to link data relating crime, health and education in a meaningful way due to the inflexibility of official statistical spatial units and of the potential for ecological fallacy. This paper explores the use of administrative data sets and data matching techniques to profile households and neighbourhoods to any size or shape. A systematic framework is developed for evaluating risk based on probability theory, employing GIS techniques to produce customized neighbourhood units using newly available local property registers. Based on the framework, a case study of police-reported domestic violence is presented in which data are combined from the police, education, health and local authority and systematically evaluated using new presentational methods and logistic regression. The results show a risk gradient based on risk factors applied at a household level and clear spatial patterns of risk for a deprived area of London. The results can be used to direct local services and government funded programmes to areas of greatest need.

Key words: Risk framework; household level data; logistic regression, GIS applications; domestic violence

1. Background and introduction

Whilst the need for better health care services is undeniable, much need is determined through the interaction of the social structure and the health economy. Evidence for this is apparent in differences in life expectancy between economically deprived and more affluent areas, between the usage of services and wider social phenomena such as crime, domestic violence, drug abuse, and unemployment. The Acheson report on inequalities in health (Acheson, 1998), for example, brings together recent thinking on the causes of inequalities and provides examples of how health and social phenomena can interact.

A key problem is how to use available information on inequalities to target government or local interventions, because typically available data cannot show how localized these interactions are or indeed to what extent different risk factors combine within individual households or neighbourhoods. Potential bias known as the 'ecological fallacy' may be introduced, and refers to the danger of drawing inferences about individuals from aggregate data (Robinson, 1950; Greenland and Robins, 1994). Practical problems arise due to unstable or conflicting administrative boundaries and the use of non-contemporaneous data (Openshaw and Taylor, 1981). Surveys, a potential alternative

source of local data, suffer from the inherent weaknesses that they are expensive to undertake and are generally based on small samples.

For years statistical data have been a by-product of administrative processes and are used in the production of key statistical series on trade, health care, unemployment etc. (Brackstone, 1987). They include registers that keep track of events or inventories such as vital events or patient lists based on statutory or other standard definitions. They are used to support services and record events such as crime, calls to emergency services, social security payments and so on. The minimum information contained in such databases includes basic demographic details like gender, date of birth and possibly ethnicity, and a record of an event such as a hospital visit. Most databases also contain a street address that can be matched with other database addresses. In this paper we use addresses to link data at the level of the household rather than to pre-defined areas, such as wards or post-codes.

The decision to use household addresses for linking purposes requires further clarification. A household may be defined as a group of individuals occupying a house, apartment, group of rooms or a single room that is considered a housing unit. The Office for National Statistics defines a household as people ‘living at the same address with common housekeeping – that is sharing either a living room or at least one meal a day’ (ONS, 2003). Although an address does not necessarily correspond to a household, it is a reasonable approximation for most cases, except where there is more than one household per address. By using the geo-coded property addresses contained in the locally available property gazetteers, it is possible to analyze household types and map neighbourhoods of any shape or size. The fact that administrative data sources are kept up to date provides another advantage over traditional data sources for many purposes.

For some applications, it may be more appropriate to link events and contextual information to individuals. An example of this would be clinical studies in which information about the individual is more relevant than, say, information about the household. A theoretical preference for households rather than individuals exists where events occurring *within* households to one individual may affect events happening to the

same or other individuals in the same household. Examples of this could include studies where there may be interactions at a household level in terms of crime, poverty or educational attainment. Later we shall discuss some of the practical issues and limitations that arise in linking data at a household level. For a discussion of issues arising in the use of individual level data see Coombes and Raybould (1997).

Because administrative data sets are generally very large (typically there are around 100,000 residential addresses in a London borough), systematic methods are needed to analyze patterns in the data. In the first part of this paper, we develop an approach in which risk is contingent on other events or circumstances such as crime, educational attainment, or living in a certain category of housing. If risk can be systematically evaluated, its practical value is greatly enhanced in terms of providing evidence that can be used for taking appropriate preventive or remedial action at the local level. Service providers, for example, would be able to quantify how frequently different risk factors occur together in the same areas and the extent to which joint working with other agencies would be necessary or desirable.

A research question might be whether some households are more 'at risk' of crime or adverse health events if certain risk factors are present and to what extent they affect social and physical well being. Using the framework, it is possible to calculate how many households fall into each risk category and also their spatial concentration. In each case, the choice of risk factors would be partly determined by practical considerations including what data are available and accessible, partly on guidance from professionals working in the area, and finally on research contained in the literature on deprivation and health inequalities.

In applying the approach, it is important to stress that the association of one risk factor with a particular event does not necessarily imply a causal link, although inevitably some will draw that conclusion. Linking events in time to individuals rather than to households is one way to approach this question, although even where this can be done questions of interpretation will always remain. Furthermore, a causal factor could be overlooked altogether if the incidence of that factor is such that it occurs in every household. As Rose

(1985) points out: "...the more widespread a cause the less it explains the *distribution* of cases". Thus, suppose we knew that, for cultural reasons, all households in an area were predisposed to domestic violence, we would not be able to isolate that particular variable unless we were somehow able to compare incidence across *different* cultures. This is simply to say that caution is always needed in interpreting the results.

The process of matching household addresses requires the use of person identifiable data from different agencies, which may be reluctant, or legally constrained, in sharing such information with each other or with third parties. For these reasons the advice of the Data Protection Agency (DPA) was sought. Their advice was that personal data may be processed under section 33 of the 1998 Data Protection Act for research purposes, providing personal data are not disclosed or the results are not used to support decisions about individuals. Further, since the intended use of the data would be ultimately to improve local services, the DPA also advised that this would be within the expectation of individuals and so their permission to use their personal details would not be necessary. This guidance was followed throughout.

This research arose from an initiative financed by the Brent and Harrow Health Authority under the aegis of the Brent Health Action Zone (HAZ). HAZs can be regarded as the Labour Government's response to improving the way local services are delivered and in addressing health and social problems. Brent, a northwest London borough, is often portrayed as an outer suburb but with inner city problems, and so was ideal as a testing ground for the techniques described. The paper begins with a description of the risk framework adopted, which for the most part is based on probability theory. The main new elements concern the way that information is handled, tabulated and analyzed rather than the specific techniques employed.

This section is followed by an illustrative case study on the subject of domestic violence (abbreviated to DV), which is currently of great interest in the health and social policy field. This case study provides a fair test of the concepts and practical aspects of DV since it brings together diverse data from four agencies; the local authority housing department, police, mental health trust, and education services. The results show a clear

risk gradient as well as neighbourhood concentrations of DV, depending on the absence or presence of certain factors. A concluding discussion then reviews other applications of the techniques that are currently under development and some longer-term data issues raised by the study.

2. Methods of measuring and conceptualizing risk at the household level

We begin with an outline of the statistical methods used to estimate household risk. We use 'risk' here to define the chance of an adverse or disadvantageous event happening, for example, a bereavement or unemployment. We define a risk factor as a situation or an event that could increase or be associated with the probability of occurrence of an adverse event. Note that a risk factor does not therefore have to be the cause of an event to be associated with it. We use the term probability in a more categorical way, for example the probability of living in social housing. These distinctions, though, are not always hard and fast. For a general descriptive introduction to risk and probability see any standard textbook such as Chou (1972), Armitage and Berry (1987), or Altman (1999).

The methodology operates in stages: firstly relevant data from different agencies are matched at address level to the local authority property gazetteer using a purpose-designed address-matching algorithm. Each household is assigned a geographic coordinate and data are then anonymized to suppress any identification of particular households or persons. Individual factor combinations are tabulated and exhaustively enumerated. The risk of a given event occurring with every factor combination is calculated and confidence intervals are then derived based on the observed ratios of events per household combination (Barnett, 2002). A normally distributed sample is assumed for these purposes, but exact methods are needed where household factor combination sizes fall below acceptable levels (see Annex 1, and for example Freund 1973).

One way to define accuracy is by the probability of the observed value of risk being within +/- y% of the true value of risk with a specified level of confidence. It may not be possible to ensure every factor combination meets the desired criterion, as many combinations will contain only a small number of households (see below). A typical

decision rule would be to express the width of confidence band at the 95% level of confidence as a ratio of the observed risk value. Where the ratio is less than some multiple of the observed risk an indication to this effect may be included in tabular outputs. For example, if the width is less than twice the observed risk it will be significantly different from zero at the 95% level of confidence.

Other benchmarks such as the average risk for the whole population may be used as a comparator instead of zero, as well as using different decision rules depending on preference. Similar considerations apply to the determination of relative risk, which is defined as the ratio of two observed risks. In an M -factor model there will be 2^{2M} such measures of relative risk including the same factor combination with itself, in which case the relative risk will be one. The method for calculating confidence intervals for relative risk may be found in Altman (1999).

The above procedure relates to the measurement of risk. It does not say whether a factor is itself a statistically significant explanatory variable of a given event, only that its presence can be measured. Logistic regression techniques are used to evaluate the contribution of each risk factor to an event, for which there are several possible specifications and estimation techniques. Which method is adopted is less important than the need to specify a suitable model with an appropriate number of factors based on a sound hypothesis. Greater transparency is achieved if the model coefficients have the same polarity. For example, in some applications it is found that living in social housing added to the risk of an event occurring, while in others it reduced it. In one model therefore the variable for this risk factor would be defined as 'living in social housing' and in the other as 'not living in social housing'.

The choice of factors normally depends on the prior hypothesis, on previous research and on the availability of suitable data. There are limitations to the number of factors that can be incorporated in a single model because the number factor combinations grow rapidly at a rate of 2^M . So, for example, with five factors, factor combinations rise to 32, while for 9 factors they increase to 512. As M increases the number of events associated with a

given risk are spread over a rapidly increasing number of household factor combinations, so that many factor combinations may have no entries or events. Consequently, confidence intervals around individual risk estimates diverge and in some cases they cannot be calculated.

Several considerations are involved in obtaining a suitable model. These include the statistical properties of individual regression coefficients and possible interaction effects between individual factors, and so an iterative approach is often appropriate. There are a number of possible strategies for developing decision rules for the inclusion or exclusion of factors at the design stage and in subsequent iterations. A reasonable aim is to ensure that the number of factor combinations with no ‘events’ is kept low by eliminating systematically the number of included factors based on their coefficients and standard errors.

The customary way of checking if the number of households in a given factor combination is unusual is to compare the expected number that would have arisen by chance with the number that actually occur and then use the chi-squared (χ^2) to test for independence. Each case is different but where risk factors are statistically independent of one another, the *expected* risk of an event will be identical for every factor combination. Suppose there are M factors in a population of households, then it is easily seen that the risk of x is a constant:

$$r_x = \frac{P_1 P_2 \cdots P_m \cdots P_M P_x}{P_1 P_2 \cdots P_m \cdots P_{M-1} P_x + P_1 P_2 \cdots P_m \cdots P_{M-1} (1 - p_x)} = p_x$$

Where p_m is the probability of occurrence of factor m .

Although independence is not a likely outcome in actual applications, the result is useful for obtaining rough prior estimates of how many households one would *expect* to find with 0, 1 or more than 1 risk factor present. Consider a neighbourhood in which M factors are selected for analysis, each with an equal chance p of occurrence. The proportion of

households with one or more factors present increases with the prevalence of each factor, whereas the proportion with no factors declines.

It may be shown that the expected proportion of households with k factors is given by $P(k) = C_M^k p^k (1-p)^{M-k}$. This has a maximum value of $P(k) = (1 - \frac{k}{M})^{M-k}$, when $p = 1/M$. For example, if $M = 5$ the factor probability is 0.2, and the maximum percentage of households with one factor is shown to be 41.1%.

The relationship between factor occurrence and the expected distribution among households is given in Figure 1. It shows that the proportion of households with no factors declines as the proportion with one factor increases, until the latter reaches a maximum before declining. A third curve shows how the probability of finding households with more than one factor increases reaching a maximum when $p=1$, which denotes the special case when all factors are present in all households. Further analysis shows that as M increases indefinitely, the proportion of households with one factor converges to e^{-1} or 36.8% and the proportion of households with no factors converges to the same. Finally, analysis can also show that the proportion with 2 or more factors converges to $1 - \frac{2}{e}$, which equates to 26.4%.

Constant factor prevalence would normally be considered unusual. Actual factor prevalence is easily obtained from the base data in which case simulation using the above approach may be used to obtain prior estimates of the frequency occurrence of each factor combination. Hence, this can give an indication of the likely confidence intervals for any given event and of the concentration of risk factors. The results obtained, however, are still only a guide. The actual frequency distribution among factor combinations can be quite different, although sufficiently close for this technique to add value and to lead to improvements at the design stage of any project.

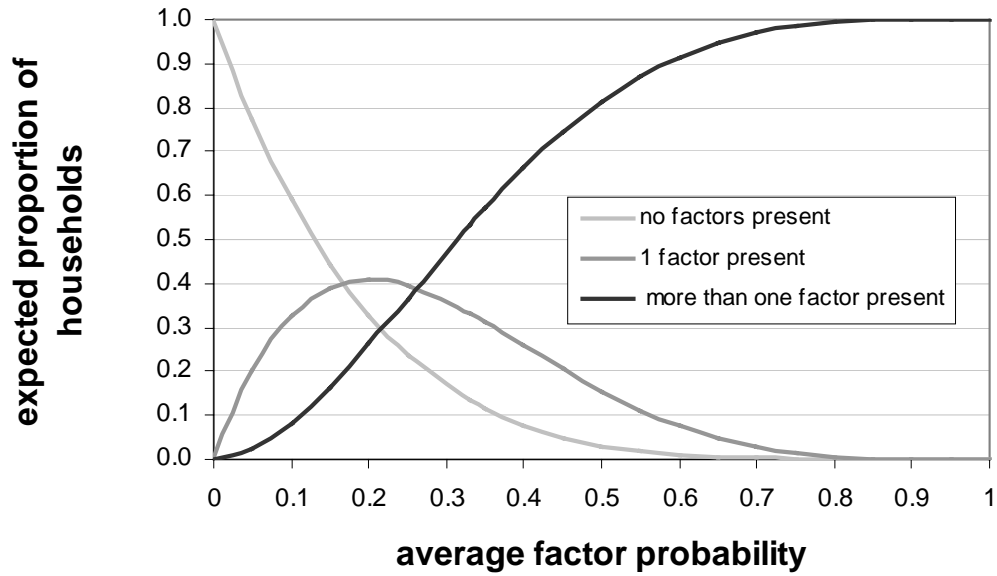


Figure 1: The expected proportion of households with zero, one or between two and five factors present.

Tabulating risk

Because of the potentially large number of factor combinations, systematic methods are needed for organizing large quantities of data and for presenting results. In this section we consider how risk combinations are defined, enumerated, uniquely identified and categorized. Consider initially a simple risk analysis based on 3 factors, A, B and C as shown in the Venn diagram in Figure 2, where n refers to the number of households in each factor combination. Consider the risk of A occurring with or without the presence of B and refer to Table 1. For example, $A \cap B$ means the set of households in which factor A and factor B are both present whilst $A \cup B$ means 'either or'. The symbol Ω is defined as the universal set of all households. Column three in Table 1 shows how they are defined and enumerated in terms of n .

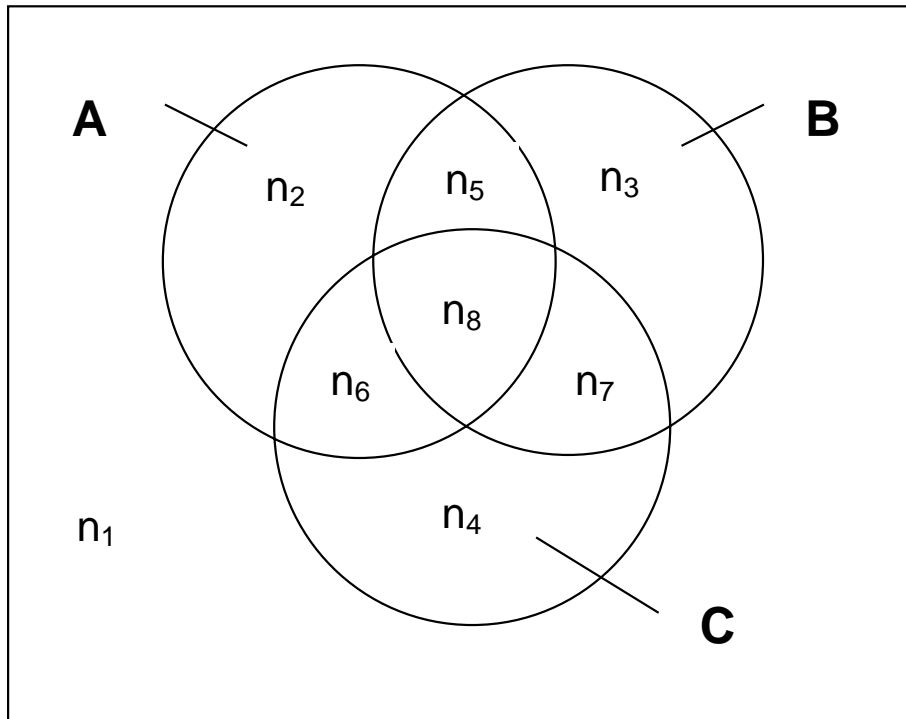


Figure 2: Counting basis for evaluating risk in the three-factor case.

Table 1 is split into two parts. The upper part, consisting of 4 rows, is termed the ‘mutually exclusive set’, because the number of households in each combination is mutually exclusive and adds up to the total number of households in the set $(\sum_{i=1}^{i=8} n_i)$.

There are 3 factors, 8 possible combinations and four possible measures of risk. Each is designated a ‘level’ (see column 1): no risk factor is designated level zero, one risk factor level 1, for which there are two cases, and two risk factors level 2. By contrast, the lower part of Table 1 is termed the ‘overlapping’ set and also has three levels and four measures of risk. The difference is that in the overlapping case, the risk of A occurring with B is defined *regardless* of the presence of C, whereas in the mutually exclusive case the presence of C is explicitly ruled out.

Each level and risk factor combination represents different subsets or aggregations of risk, which expand in a binomial progression. Figure 3 shows an example of an

expansion based on four risk factors using a 0-1 notation, in which zero indicates the absence and 1 indicates the presence of a given factor. Note that a bar over the overlapping sets is used to differentiate between each type of set. This notation may be thought of as analogous to a ‘bar-code’ for identifying household types. It is useful as a tool for sorting and filtering large data sets and generating tabular outputs. In this four-factor example, there are 16 mutually exclusive and 16 overlapping risk factor combinations, giving a total of 32 or 2^M where M is the number of risk factors in the set. In the case study that follows we develop these concepts further and show how results may be analysed and presented.

Data considerations

Thus far only the types of data that may be used have been indicated, but no detail has been given of the practical issues arising from their use. For the most part the administrative data sets used are locally produced and maintained and are held by the local authority, police, or health services. A detailed analysis of the strengths and weaknesses of the data is beyond the scope of this paper. However, in general we found the quality to be good, with the key problem being the non-standard recording of addresses rather than the completeness or accuracy of the data. In this section we are primarily concerned with prior selection considerations, and how to conceptualize and use the data provided by these sources.

It is practical to distinguish three types of variable, which may loosely be described as ‘categorical’, ‘event’, or ‘flow’. Categorical data are data that do not change very often or at all within a time window. Examples include household tenancy or the ethnicity of a subject. Event data are data of the one-off variety like a birth, bereavement or crime, whereas flow data refer to activities with a measurable duration. This could include spells on means tested benefits or entitlement to free school meals (typically abbreviated by FSM), for which the household base changes quite frequently as people move in and out of work. Some variables, such as age, are continuous and may need to be discretized (e.g. whether a person is over or under 18 years). Inevitably this can result in some loss of information.

The three types of data are fundamentally different in the sense that the stock of social housing at a point in time does not alter over the short term. The stock of households receiving FSM, by contrast, is related to the flow (incidence) and the duration of entitlement (in the steady state, stock = flow \times average duration). The simplest way of associating data is to pick the most recent time period, say two years and to match the events occurring to the 'stocks' during that time frame; the hypothesis being that factors are 'associated' if they occur together in the same household. Whilst this appears straightforward in principle, there are a number of important issues to consider, especially with 'flow' and 'event' data types.

Administrative data, for example, tend to reflect the current caseload rather than lapsed cases. Access to lapsed cases that terminate within the chosen time frame may, at the least, be extremely difficult or impossible. This inevitably introduces an element of imprecision into calculations involving these types of data and also a source of bias if lapsed cases are not typical. Furthermore, some subjects change address during the interval and so working with 'households' rather than named individuals can be another source of inaccuracy. This drawback is potentially surmountable if names, as well as addresses, are used for cross checking purposes.

The main issue with event data concerns multiple occurrences of events at the same address. A typical example of this would be multiple visits from a health worker or several reported crimes. Since the measure of risk is defined on the scale from zero to one, multiple events at a single address could violate this condition. One option is to count multiple events as single occurrences or alternatively each event could be counted individually and treated as a separate risk occurrence, for example two *different* types of crime.

<i>Level</i>	<i>Risk set</i>	<i>Evaluation</i>	<i>Risk or probability</i>
<u>Mutually exclusive set</u>			
0	$\frac{A \cap \bar{B} \cap \bar{C}}{\bar{B} \cap \bar{C}}$	$\frac{n_2}{n_2 + n_1}$	Risk of A occurring with B and C absent
1	$\frac{A \cap B \cap \bar{C}}{B \cap \bar{C}}$	$\frac{n_5}{n_5 + n_3}$	Risk of A occurring with B with C absent
	$\frac{A \cap \bar{B} \cap C}{\bar{B} \cap C}$	$\frac{n_6}{n_6 + n_4}$	Risk of A occurring with C with B absent
2	$\frac{A \cap B \cap C}{B \cap C}$	$\frac{n_8}{n_8 + n_7}$	Risk of A occurring with B and C
<u>Overlapping set</u>			
0	$\frac{A}{\Omega}$	$\frac{n_2 + n_5 + n_6 + n_8}{n_2 + n_5 + n_6 + n_8 + n_1 + n_3 + n_4 + n_7}$	Risk of A occurring at all
1	$\frac{A \cap B}{B}$	$\frac{n_5 + n_8}{n_5 + n_8 + n_3 + n_7}$	Risk of A occurring with B regardless of whether C present
	$\frac{A \cap C}{C}$	$\frac{n_6 + n_8}{n_6 + n_8 + n_4 + n_7}$	Risk of A occurring with C regardless of whether B present
2	$\frac{A \cap (B \cup C)}{(B \cup C)}$	$\frac{n_5 + n_6 + n_8}{n_5 + n_6 + n_8 + n_3 + n_4 + n_7}$	Risk of A occurring with B or C

Table 1: Levels of risk in three-factor example based on set theory linked to the notation in Figure 2.

Mutually exclusive set

			1100		
			1010		
		1000	1001	1110	
		0100	0110	1101	
		0010	0101	1011	
0000	0001	0011	0111	1111	
<i>level 0</i>	<i>level 1</i>	<i>level 2</i>	<i>level 3</i>	<i>level 4</i>	
$\overline{0000}$	$\overline{0001}$	$\overline{0011}$	$\overline{0111}$	$\overline{1111}$	
	$\overline{0010}$	$\overline{0101}$	$\overline{1011}$		
	$\overline{0100}$	$\overline{0110}$	$\overline{1101}$		
	$\overline{1000}$	$\overline{1001}$	$\overline{1110}$		
		$\overline{1010}$			
		$\overline{1100}$			

Overlapping set

Figure 3: Risk hierarchy for four factors for the mutually exclusive and overlapping risk sets. A bar is used over the overlapping set to distinguish between the two types of case.

For certain phenomena only proxy variables may be available. For example, a proxy for heavy drinking may be a previous police caution or conviction that is alcohol-related, but clearly its use would not capture heavy drinkers who have never been cautioned or convicted. This is an example of a more general problem where it is not mandatory to report something to an authority. The fact that not all crimes or relevant incidents may be reported can introduce bias or under estimation of the risk or risks involved. A health service example arises where for reasons of resource constraints the methodology may not be directly translatable into need. It should be noted, however, that to some extent these problems are inevitable with any type of data set.

An important deficiency in many administrative systems occurs with respect to ethnicity. Where it is reported in more than one administrative data set, the definitions often differ (e.g. crime data and school pupil registers). Where data are incomplete, it may only be feasible to evaluate risk in certain factor combinations and at certain levels in a risk hierarchy. However, such analyses may be misleading. Take for example the case of

burglary. Crime data record the ethnic group of victims, the accused and their addresses, but suppose there is no information on the ethnicity of households *not* victimized or accused of burglary. In such situations it would be possible to analyze the risk of being victimized or accused by ethnic group, while controlling for other factors (for example, household tenancy) to see which are more susceptible.

However, such an analysis could not provide an overall risk assessment of the likelihood of a particular ethnic group being involved in that crime unless their total numbers can be enumerated through other data. We may illustrate this by returning to the three-factor example in Figure 2 and the hierarchy in Table 1. Suppose factor A represented households with suspected burglars, factor B households belonging to a particular ethnic group, and factor C social housing. However, estimates would not normally be available for n_3 , n_7 , n_4 , or n_1 . These represent:

- n_3 the number of households in the chosen ethnic group *not* involved in burglary and *not* living in social housing,
- n_7 the number in the chosen ethnic group that *are* living in social housing but *not* involved in burglary,
- n_4 the number *not* in the ethnic group that *are* living in social housing but *not* involved in burglary and
- n_1 the number *not* in the ethnic group *not* living in social housing and *not* involved in burglary.

Note, however, we will generally have or can deduce information on n_4+n_7 or n_3+n_1 . More to the point, if we had information on just one of the missing unknowns, all the others could be calculated. In practical terms this means *it may not be possible to calculate risk in certain levels or combinations*. In this particular example only two can be imputed from the list in Table 1 with the assumed information available. These are:

- Level 0, overlapping set, the risk of A occurring at all (i.e. the overall risk of burglary),
- Level 1, overlapping set, the risk of A occurring with C regardless of whether B is present (i.e. burglary with social housing).

Neither provides a risk assessment based on ethnicity. If such data are missing some of the possible analyses that can be carried out will be misleading. For example, if we consider all burglaries only (circle A in Figure 2) and examine the ethnic composition of all the accused, we might find one group over represented, but this may not be a surprising result if that group represent a majority of all households. Thus, it is preferable to have information about all households and not just those accused or suspected of burglary. Note that the issue of ethnicity is also part of a more general problem regarding its definition, whether it is recorded at all in administrative databases; although given consistent definitions it can be dealt with by the framework.

Not all administrative systems provide data at address level to enable address matching to the property database. Hospital activity data, for example, only record post-code of residence. In these cases an intermediate data set containing addresses may be appropriate, in which names and dates of birth are used as primary link variables to populate the original data-set with the missing addresses. The ambulance service, on the other hand, provides only a location reference relating to the place on an incident, which could refer to the patient's home or elsewhere. The geo-referencing system used is not as precise as the local property gazetteer and does not for example distinguish between flats in high-rise buildings. Where addresses are non-standard or incomplete, manual data matching rather than an address-matching algorithm may be needed to complete the matching process.

Finally, a general issue arises with service-based data sources, such as hospital activity, rather than residential-based data sources like council tax. Post-codes can be used to screen households into the selected geographical frame of reference relatively simply, but it is obviously more difficult to include residents accessing services in neighbouring areas

or authorities. This may necessitate data pooling arrangements with neighbouring service providers or local authorities (e.g. in the case of educational data), although the extent of the problem will depend on the nature of the service and its geographical range.

3. Illustrative case study

Attention now turns to a case study on domestic violence (abbreviated as DV) in which most of the issues identified in the preceding sections are highlighted. Other case studies could have been selected but DV turns out to be a classic example of why data matching techniques may be warranted and standard data sources are likely to be deficient. DV is increasingly recognized as a health as well as a social problem, whose current high profile can be gauged by the launch of a London-wide domestic violence strategy (Greater London Authority, 2001).

The literature on DV is extensive but it differs in three key respects from the analysis presented here. Firstly, this research adopts a broader definition of DV than is usual in published research, which usually deals with specific sub-types like violence against women or between partners (e.g. see Jewkes, 2002). Here we use the police practice of classifying any incident as ‘domestic’ if it involves any member or member of a household including children or there is some sort of close relationship, including former partners.

Secondly, the research is dependent on an incident being reported to the police in the first place; however, it can safely be assumed that the underlying level of DV (reported and unreported) will be higher, although how it varies in this more general case is beyond the scope of this paper (for wider discussion and evidence see, Mirlees-Black, 1998; and Felson et al, 2002). Thirdly, risk is based on routinely held administrative data rather than data obtained in surveys where there is more control over variable definition, and so caution is necessary to select the right types of proxies.

There is ample agreement in the DV literature as to which risk factors appear to be important both within the UK as well as overseas, although attempts at combining them in a comprehensive framework have only been partial. The factors include drug and

alcohol abuse, mental health problems, poverty, and young children in the household (Brown, 2002; Coker et al, 2002; Cunradi et al, 2002; Moreno et al, 2002; Tolman and Raphael, 2000).

For this case study, we used five administrative sources from five different departments or agencies to try to capture these dimensions. After examining and testing various possible factors, the model was reduced to five that appeared to capture the main dimensions required. These were: 1) households in receipt of free school meals from the education department (abbreviated to FSM); 2) households with registered mental health patients from the local mental health trust; 3) noise complaints to local authority from the housing department; 4) drug offending from the police and 5) social housing from the local authority (council tenancy or housing association).

Two of these, noise complaints and drug offending, may be considered as ‘event variables’, FSM and mental health patients as ‘flow-variables’ and social housing as a ‘categorical variable’. Note that FSM seeks to capture poverty *and* children since eligibility depends on the receipt of Income Support¹ and a child being of school age (5-16 years), whilst social housing is included to capture material circumstances and possible contextual risks.

Based on these five factors, there are 64 possible risk factor combinations giving rise to 32 each in the mutually exclusive and overlapping risk sets. The following analysis is split into two parts: first the risk of DV occurring at all and secondly the risk of DV being repeated more than once in a household. The reason for the difference was to see whether the influence of the risk factors in each case is consistent, and if there was a ‘hardened case’ effect. The procedure used was the same in both cases. First, the number of households in each risk factor combination was tabulated, and then the risk of DV was calculated by observing the proportion of reported incidents for each one. Confidence intervals were also calculated at this stage both for each risk observation and for relative risk, using the techniques described earlier.

¹ A financial benefit paid to qualifying households on low income

Several specifications of the logistic model and means of estimating the parameters were tested in the course of this research. These included ordinary least squares (OLS), grouped or ungrouped, and maximum likelihood. In the grouped approach each risk category represents one observation regardless of the frequency of households. Consequently OLS treats each group as equal in weight and there are only as many observations as there are risk categories (5 factors equates to a maximum of 32 observations). In the ungrouped approach there are as many observations as there are households; over 100,000 observations in this case study.

A consequence of these differences is that regression diagnostics for ungrouped data show very high values for R-squared and low standard errors for the coefficients. In the grouped model, R-squared is still high but standard errors are higher, although still significant. The principal difference between the grouped and ungrouped approach is that the grouped model will predict the observed risk in factor combinations with small numbers of households more accurately than the ungrouped approach. However, although variance is minimized, the coefficients are biased. A more detailed comparison of each method is beyond the scope of this paper. In what follows we restrict our attention to the results obtained using grouped and ungrouped OLS, since the coefficients obtained using maximum likelihood model are very similar to those using ungrouped OLS.

The basic OLS model is specified as follows.

$$\hat{L}_i = \ln \left[\frac{\hat{r}_i}{1 - \hat{r}_i} \right] = \beta_0 + \sum_{m=1}^{m=M} \beta_m x_{mi} + u_i$$

where i is the i^{th} factor combination, $i=1,32$ and M is the number factors in the model, $M=5$, and \hat{L}_i is the sample log-odds in category i of a reported DV occurrence.

For each factor combination:

$\hat{r}_i = \frac{a_i}{n_i}$ sample risk of DV occurring in a household

$a_i =$ the number of occurrences of DV in factor combination i

$n_i =$ the total number of households in factor combination i with reported DV

$x_{mi} =$ the presence or absence of factor m in factor combination i

$\beta_m =$ model coefficient for the m^{th} risk factor

$\beta_0 =$ model constant

$u_i =$ error term

Since by definition

$$risk = \frac{odds}{1 + odds}$$

The predicted risk of occurrence of an event in factor combination i may be estimated from:

$$\bar{r}_i = \frac{\exp(\alpha + \sum_{m=1}^{m=M} \beta_m x_{mi})}{1 + \exp(\alpha + \sum_{m=1}^{m=M} \beta_m x_{mi})}$$

The model was fitted separately for the mutually exclusive set (case a) and the overlapping set (case b) for any reported incident. This was then repeated in the case of multiple reported incidents of DV, giving eight sets of regression results altogether. Interaction effects were also investigated to examine if the presence of one factor with another could cause second-order effects by either raising or lowering risk. However, when this possibility was tested for all possible interaction pairs, only weak effects were observed and so the detail is omitted from the results.

Measuring the risk of any DV incident

Results for comparing the risk of *any* reported DV incident are shown in Table 2 in a format known as a ‘risk ladder’. There were 2,582 cases of DV reported altogether in the given time window, distributed among 102,427 households. The totals at the foot of each risk factor in Table 2 show the number of households with each factor present. Factor combinations showing the number of households within each combination are given in descending order of risk for the mutually exclusive set and overlapping set (shown in italics). Observed risk ranges from zero (no factors) up to five, although in practice the observed maximum was only 3. Four categories shown in the table had no occurrences of DV but contained some households, while a further twelve higher-level combinations (three or more factors) had no households or incidents of DV and were therefore omitted.

In general, results for case (a) indicate that the more factors there are, the higher the observed risk of DV. Note, however, that the number of households in high-risk categories is markedly smaller, which is to be expected given the previous discussion on the likelihood of finding factors together. Figure 4 shows risk plotted on the vertical axis and each factor combination on the horizontal axis using the results in Table 2, case (a). Vertical lines display the 95 % confidence intervals, which apply in each case. The results for case (b) generally follow a similar pattern to case (a). Note, however, that the household column in case (a) sums to the households in the study area, which does not apply to case (b).

Consider case (a) in Table 2 in more detail. The households at highest risk are those in receipt of FSM and where there have been noise complaints. The risk turns out to be

100%, although this factor combination is mainly of interest because noise complaints are relatively rare. The second risk category, FSM and drug offending, has a risk of 33% but here also there are only three observations in case (a). Moving down the table, interesting cases include social housing and drug offending (risk 11.8%), social housing and FSM (9.6%), FSM by itself (6.7%) and social housing by itself (3.8%). The largest category in terms of households (74,356) is that with no factors present (1.8%). This result may be compared with the average risk of DV of 2.5%, taking all 102,427 households together.

There are some consistencies but also some inconsistencies between the risk categories in the sense that more factors should generally equate to greater risk if the model is well-specified. For example, the risk of noise, social housing and FSM is greater than the risk of noise and social housing, and social housing and FSM, and noise and FSM. However, the evidence for there being a difference is inconclusive, as there are so few cases in either category. Conversely, if we take the combination of drug offences and FSM, the risk is nearly 6% higher than the risk category of social housing, drug offending and FSM. Such anomalies are to be expected since no model is perfect, as long as there is the clear suggestion of a risk gradient and a pattern among the factors. Note that hypothesis of DV being independent of household category would be rejected by a chi-square test, as can be easily demonstrated.

An important issue is by how much one factor combination is at greater risk than another, since this may be influential in targeting resources. Relative risk is obtained by dividing one observed risk by another for every factor combination. So for example the observed risk of DV in the presence of drug offending and social housing compared with the risk for FSM and social housing would be $11.8/9.6$. This risk is 1.22 or 22% higher, with associated 90% confidence intervals of 0.92 to 1.63. Table 3 tabulates relative risk for each factor combination, with confidence intervals omitted for brevity (see Altman, 1999 pp266-268 for method of calculation). The binary 'bar code' system, as previously described, is employed to signify the presence or absence of a factor in a given combination.

Risk Level	Case a Households	Case b Households	Noise Complaint	Mental Health	Social Housing	Drug Offence	Free school Meals	% risk of DV a	%risk of DV b
3	2	2	Y		Y		Y	100.0	100.0
2	3	13		Y		Y		33.3	15.4
2	19	53				Y	Y	26.3	22.6
3	34	34			Y	Y	Y	20.6	20.6
3	15	15		Y	Y		Y	13.3	13.3
2	323	367			Y	Y		11.8	12.5
3	10	10		Y	Y	Y		10.0	10.0
2	1265	1316			Y		Y	9.6	10.1
2	11	26		Y			Y	9.1	11.5
1	405	794				Y		6.9	10.1
2	366	393		Y	Y			6.8	7.1
1	1209	2558					Y	6.7	8.6
2	35	39	Y		Y			5.7	10.3
1	374	783		Y				5.3	6.4
1	23944	25996			Y			3.8	4.2
0	74356	102427						1.8	2.5
1	49	93	Y					0	4.3
2	2	4	Y	Y				0	0.0
2	3	5	Y				Y	0	40.0
3	2	2	Y	Y	Y			0	0.0
total	102427	-	93	783	25996	794	2558	-	-

Table 2: Risk of DV incident being reported to the police according to different factor combinations based on 102,427 households. Case (a) the mutually exclusive set; case (b) the overlapping set.

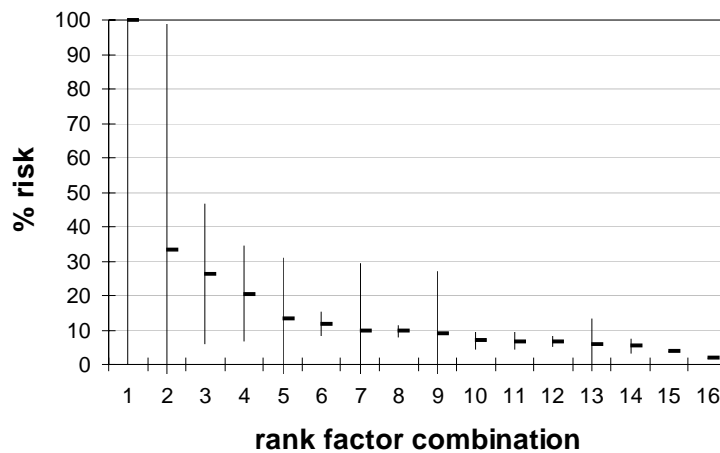


Figure 4: Ordered factor combination risks based on case (a) with 95% confidence intervals as indicate by vertical bars.

Regression results

Table 4 shows the results of the regression analysis for cases (a) and (b) in Table 2, using grouped and ungrouped OLS. A plot of observed and predicted risk values in Figure 5 based on case (a) for the grouped and ungrouped cases shows that most observations fall on a roughly 45-degree straight line through the origin. The main outliers tend to lie among the higher risk categories, for which the frequencies are much lower. The results for case (a), grouped OLS, suggest that drug offending is the highest risk factor, increasing the odds of DV 3.9 times ($e^{1.35}$), followed by FSM (2.6), noise (2.0), mental health (2.0), and social housing (1.1). For the ungrouped model, case (a), the odds increase by 3.7, 3.2, 1.6, 2.0 and 2.1 respectively. The general trend is that in the ungrouped model, odds increase among factors with a higher prevalence and decrease among those with a lower prevalence. The results for case (b), grouped and ungrouped, are generally consistent with case (a), and similar arguments apply.

A key difference between the grouped and ungrouped model is the effect of social housing. In the ungrouped model it increases the odds of DV by almost two-fold. Further analysis shows that social housing is the most important discriminator based on cases reported between finding and not finding DV in this population. Figure 6 is a 'spider' diagram showing the average occurrence of each risk factor, including DV, according to housing type. As is seen the overall level of occurrence of DV, FSM, mental health problems and noise is higher in social housing, but the relative occurrence of each factor is approximately the same. The main difference is that in social housing the chances of finding the other factors increases by two to three-fold.

ABCDE	10101	01010	00011	00111	01101	00110	01110	00101	01001	00010	01100	00001	10100	01000	00100	00000
10101	1.0	3.0	3.8	4.9	7.5	8.5	10.0	10.4	11.0	14.5	14.6	14.9	17.5	18.7	26.5	55.3
01010	0.3	1.0	1.3	1.6	2.5	2.8	3.3	3.5	3.7	4.8	4.9	5.0	5.8	6.2	8.8	18.4
00011	0.3	0.8	1.0	1.3	2.0	2.2	2.6	2.7	2.9	3.8	3.9	3.9	4.6	4.9	7.0	14.6
00111	0.2	0.6	0.8	1.0	1.5	1.8	2.1	2.1	2.3	3.0	3.0	3.1	3.6	3.9	5.5	11.4
01101	0.1	0.4	0.5	0.6	1.0	1.1	1.3	1.4	1.5	1.9	2.0	2.0	2.3	2.5	3.5	7.4
00110	0.1	0.4	0.4	0.6	0.9	1.0	1.2	1.2	1.3	1.7	1.7	1.8	2.1	2.2	3.1	6.5
01110	0.1	0.3	0.4	0.5	0.8	0.9	1.0	1.0	1.1	1.4	1.5	1.5	1.8	1.9	2.7	5.5
00101	0.1	0.3	0.4	0.5	0.7	0.8	1.0	1.0	1.1	1.4	1.4	1.4	1.7	1.8	2.6	5.3
01001	0.1	0.3	0.3	0.4	0.7	0.8	0.9	0.9	1.0	1.3	1.3	1.4	1.6	1.7	2.4	5.0
00010	0.1	0.2	0.3	0.3	0.5	0.6	0.7	0.7	0.8	1.0	1.0	1.0	1.2	1.3	1.8	3.8
01100	0.1	0.2	0.3	0.3	0.5	0.6	0.7	0.7	0.8	1.0	1.0	1.0	1.2	1.3	1.8	3.8
00001	0.1	0.2	0.3	0.3	0.5	0.6	0.7	0.7	0.7	1.0	1.0	1.0	1.2	1.3	1.8	3.7
10100	0.1	0.2	0.2	0.3	0.4	0.5	0.6	0.6	0.6	0.8	0.8	0.9	1.0	1.1	1.5	3.2
01000	0.1	0.2	0.2	0.3	0.4	0.5	0.5	0.6	0.6	0.8	0.8	0.8	0.9	1.0	1.4	3.0
00100	0.0	0.1	0.1	0.2	0.3	0.3	0.4	0.4	0.4	0.5	0.6	0.6	0.7	0.7	1.0	2.1
00000	0.0	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.5	1.0

Table 3: Table of relative risk based on case study (a) observed risk values, first 16 rows of Table 2².

² The emboldened table entry shows that household in social housing with a drug offender and school aged child receiving free school meals is 5.5 times more at risk of being reported for domestic violence than one in social housing alone.

	<i>Model</i>	<i>Constant term</i>	<i>Noise</i>	<i>Mental health</i>	<i>Social housing</i>	<i>Drug offending</i>	<i>Free school meals</i>	R^2
a) Mutually exclusive case								
- grouped OLS	$\hat{\beta}$	-3.59	0.7	0.68	0.09	1.35	0.96	0.80
	<i>s.e.</i>	0.29	0.56	0.26	0.25	0.26	0.26	
- ungrouped OLS	$\hat{\beta}$	-3.99	0.44	0.85	0.74	1.3	1.17	0.99
	<i>s.e.</i>	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	
b) Overlapping case								
- grouped OLS	$\hat{\beta}$	-3.40	1.10	0.44	0.10	1.06	1.14	0.80
	<i>s.e.</i>	0.25	0.31	0.23	0.21	0.23	0.21	
- ungrouped OLS	$\hat{\beta}$	-3.64	0.72	0.8	0.51	1.34	1.61	0.99
	<i>s.e.</i>	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	

Table 4: Regression results for any DV incident – mutually exclusive and overlapping cases, where *s.e.* is the standard error of the coefficient $\hat{\beta}$.

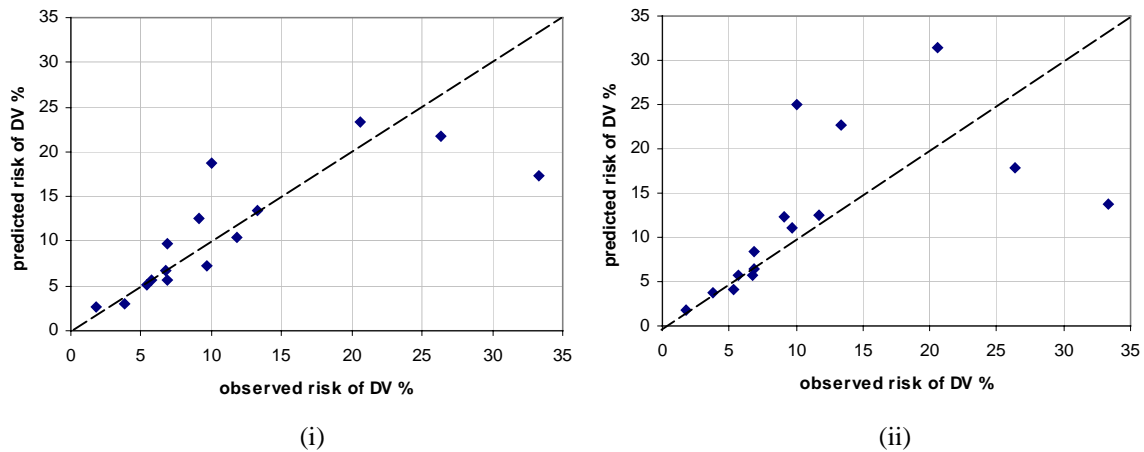


Figure 5: Predicted risk versus observed risk, case (a), Table 2: (i) grouped OLS and (ii) ungrouped OLS

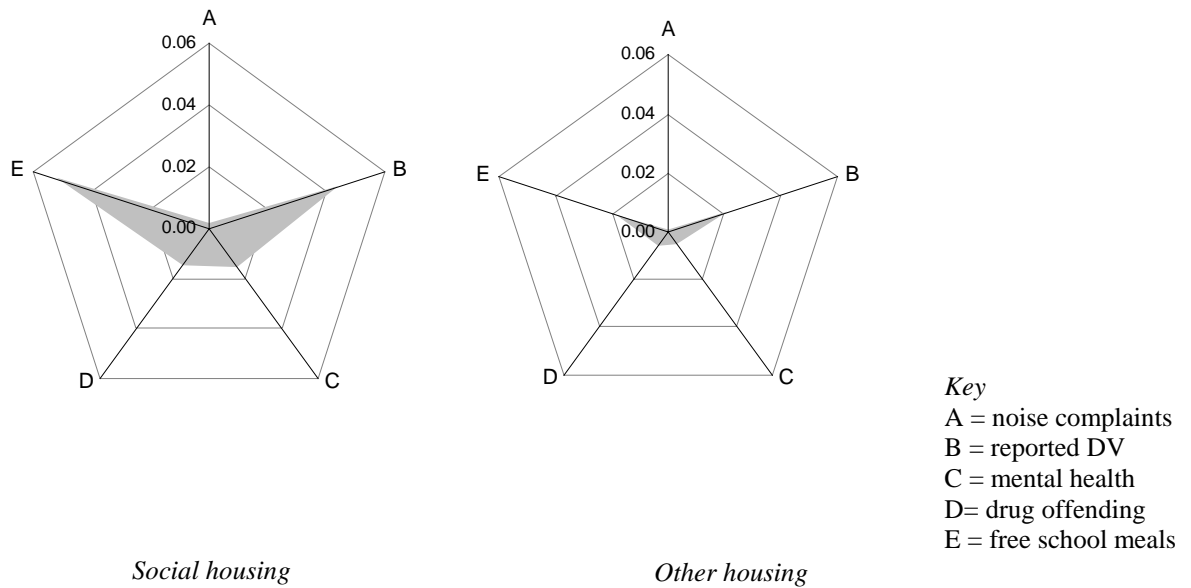


Figure 6: Spider diagrams showing the average occurrence of each factor split by 'social housing' and 'other' housing.

Multiple incidents of domestic violence at the same address

This section considers the case where there is more than one reported DV incident in a household. For brevity, the analysis of relative risk is omitted from the description, as well as other detail. The data show that multiple occurrences of DV at the same household are not uncommon and that there are sufficient observations to test this possibility. The results for cases (a) and (b) are reported in Table 5 and follow the same format as before. They show that the risk of reported multiple incidents falls on average by nearly two-thirds as compared with previously reported results, suggesting that police reporting may have led to a reduction in incidents. However, they also show that rank order is more or less preserved compared to before, with the exception of one or two risk categories involving small numbers of households.

Table 6 shows the comparable regression model results for this case. The main differences as compared with previous results are that noise, drug offending and mental health now present a somewhat higher risk. Based on case (a) the mutually exclusive set, grouped OLS, noise complaints increase the odds of DV 6.4 times compared with a

multiplier of 2 previously, drug offending 5.1 (3.9), mental health 3.7 (2.0) and FSM 2.9 (2.6) and social housing 1.0 (1.1). For the ungrouped model case (a), the same comparison yields 4.3 (1.6), 3.9 (3.7), 3.2 (2.3), 3.7 (3.2), and 2.1 (2.1). The results seem intuitively acceptable, since it would seem to confirm that multiple reported cases are in some sense more hardened and difficult to stop than isolated incidents.

<i>Level</i>	<i>No. Of Households</i> <i>a</i>	<i>No. Of Households</i> <i>b</i>	<i>Noise Complaint</i>	<i>Mental Health</i>	<i>Social Housing</i>	<i>Drug Offence</i>	<i>Free School Meals</i>	<i>% Risk Of Multiple DV a</i>	<i>%Risk Of Multiple DV b</i>
3	2	2	Y		Y		Y	100.0	100.0
2	3	13		Y		Y		33.3	15.4
3	34	34			Y	Y	Y	11.8	11.8
2	19	53				Y	Y	10.5	11.3
3	10	10		Y	Y	Y		10.0	10.0
3	15	15		Y	Y		Y	6.7	6.7
2	35	39	Y		Y			5.7	10.3
2	323	367			Y	Y		5.3	6.0
2	1265	1316			Y		Y	4.1	4.5
2	366	393		Y	Y			3.0	3.3
1	374	783		Y				2.9	3.2
1	1209	2558					Y	2.8	3.7
1	405	794				Y		2.5	4.4
1	23944	25996			Y			1.4	1.6
0	74356	102427						0.6	0.9
1	49	93	Y					0	4.3
2	2	4	Y	Y				0	0.0
2	3	5	Y				Y	0	40.0
2	11	26		Y			Y	0	3.8
3	2	2	Y	Y	Y			0	0.0
<i>total</i>	102427	-	93	783	25996	794	2558	-	-

Table 5: The risk of a DV incident being reported more than once to the police from the same household. Case (a) the mutually exclusive set; case (b) the overlapping set. Risk sets with no households are not shown.

	<i>Model</i>	<i>Constant Term</i>	<i>Noise</i>	<i>Mental Health</i>	<i>Social Housing</i>	<i>Drug Offending</i>	<i>Free school meals</i>	<i>R²</i>
a) Mutually exclusive case								
- grouped OLS	$\hat{\beta}$	-4.71	1.86	1.32	0.04	1.62	1.07	0.84
	<i>s.e.</i>	0.34	0.65	0.33	0.31	0.31	0.33	
- ungrouped OLS	$\hat{\beta}$	-5.04	1.47	1.16	0.76	1.36	1.30	0.99
	<i>s.e.</i>	<10 ⁻³	<10 ⁻³	<10 ⁻³	<10 ⁻³	<10 ⁻³	<10 ⁻³	
b) Overlapping case								
- grouped OLS	$\hat{\beta}$	-4.48	2.18	0.79	0.17	1.48	1.07	0.85
	<i>s.e.</i>	0.28	0.34	0.26	0.23	0.26	0.23	
- ungrouped OLS	$\hat{\beta}$	-4.66	1.72	1.05	0.54	1.49	1.27	0.99
	<i>s.e.</i>	<10 ⁻³	<10 ⁻³	<10 ⁻³	<10 ⁻³	<10 ⁻³	<10 ⁻³	

Table 6: Regression results for the multiple incident case.

Figure 7 is a contour map of DV based on the number of incidents per hectare. It clearly shows that there are marked concentrations in certain areas which are partly related to population density and partly to the concentration of risk factors. Figure 8, by contrast, shows the concentration of risk factors by neighbourhood (defined as 600 m x 600m cells). Each cell is shaded according to whether it falls into the top decile of occurrence on each of four factors – FSM, DV, drug offending and mental health (noise complaints have been omitted since there are too few to create a meaningful map). The darker the cell the more factors apply.

The results are remarkable since they show a clear concentration in the south of the borough that could not have been anticipated on the basis of the previous contour map. However, they also show some unexpected pockets of apparent problem neighbourhoods elsewhere, which would not have been identifiable using units based on ward or postcodes. Interpretation is important however, since the analysis has shown DV can be prevalent where no risk factors are prominent.

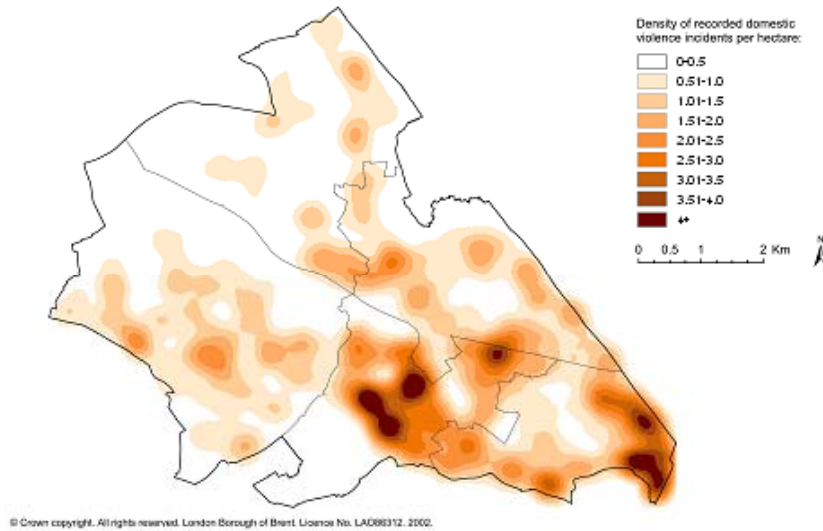


Figure 7: Map showing the density of reported DV incidents per hectare in relation to borough and locality boundaries.

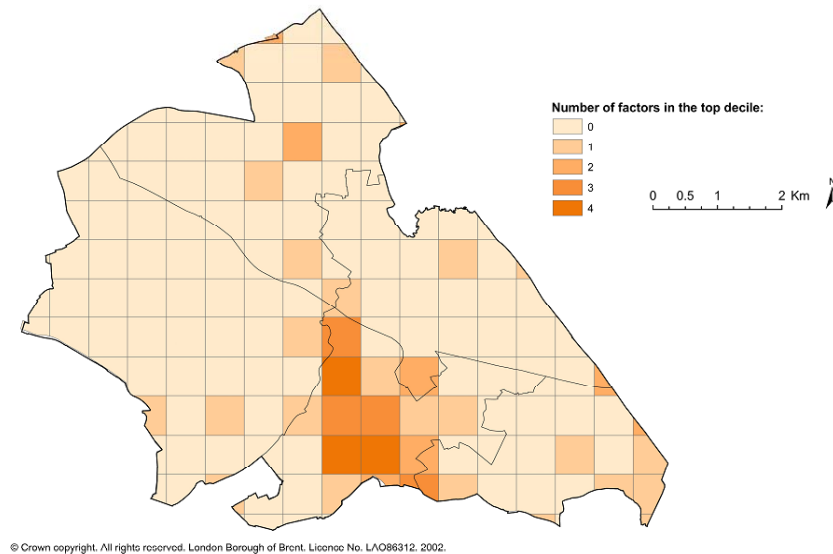


Figure 8: Neighbourhoods(600m x 600m cells) in the top decile, measured in terms of occurrences per household of DV, drug offences, FSM and mental health problems.

4. Discussion and conclusions

It has been demonstrated that the approach used enables a more detailed consideration of the exposure of different household types to various health and social risks, and the areas in which they tend to concentrate, than would have been possible using conventional data sources. At the very minimum, it has also shown that there is much information ‘locked up’ in administrative and publicly available data sets that could be exploited more than it currently is. The methodology itself is capable of development and modelling work is continuing using a range of statistical techniques. More generally the methodology has also been shown to be useful for dealing with long-standing research problems, such as the ecological fallacy and the modifiable areal unit problem (Openshaw and Taylor, 1981).

A risk factor does not have to be the ‘cause’ of an adverse event to be associated with it. However, the literature on causation is complex and is relevant to this paper only to the extent that it reminds us to be cautious in terms of any conclusions we draw (Parascandola and Weed, 2001). A good example of the pitfalls that can occur is the use of ‘counterfactual hypotheses’ in which the probability of event A in the presence of factor B is compared with the probability of event A in the absence of B. If A only occurs with B present it may be reasonable to assume some causal association; however, this is based on the assumption that other possible factors remain constant, which may not be true. For example, we may assert that children are expelled from school because they live in low-income households, but this would be to ignore simultaneous variations in other risk factors, such as school-aged crime.

Addresses have been used as proxies for households, and so this overlooks the fact that some addresses may contain more than one household. Further, since households are composed of individuals, one cannot entirely eliminate ‘ecological fallacy’ although arguably a household is a more appropriate unit of analysis than the individual in many circumstances. To strengthen any of these findings it would also be necessary to incorporate other variables. For example, the raw data indicate that certain age ranges are more at risk than others. To extend the number of factors, either the ‘time window’

covered by the analysis or an extension of the study area to adjoining administrative authorities would need to be considered.

In further work a time dimension has been added to the framework, for example to determine whether it is more common for one event to precede another or to follow it. An example might be whether school exclusion tends to lead to crime or vice versa. This has opened up a new area of investigation based on an analysis of time-sequenced factor combinations of events. However, due to the nature and stage of the development of administrative systems, it is generally not possible to go back very far in time and there are also practical constraints in terms, for example, of sample size. However, initial results are encouraging, but it will be a while before administrative systems could compete, say, with established longitudinal surveys for this kind of analysis. However, the potential certainly exists.

It would be premature to say definitively whether the approach described in this paper will reveal a greater understanding of interactions between the health and social economy and so lead to better predictive models, although the results from this case study seem promising. They show that such phenomena are not distributed randomly but that there are systematic associations between the risk factors considered. Whilst this is hardly a new finding, the fact the methodology has enabled detailed quantification of different risk combinations, at the local as well as large scale, must be considered an improvement over comparable methods. At the same time it has opened the way to a more systematic evaluation of risk patterns and potential regularities or associations.

Whilst there is obvious scientific logic in taking forward such a research agenda, it is important not to disregard the immediate potential applications for which the methodology can add value. For example, local area analysis is in increasing demand for making cases to central Government or the European Commission for funding regeneration and other initiatives. As such initiatives become more targeted and focused so the demand for detailed information and analysis increases. The methodology has

already been used in one such successful exercise and proved useful in changing perceptions about what the health needs were in a particular locality.

In terms of policy analysis, governments are increasingly demanding an evidence base for policy interventions, but as far as local initiatives are concerned this has proved problematic for a range of reasons, most often data inadequacies. This is precisely where the methodology offers a way forward. For example, imagine a scenario where systematic analyses take place at different points in time and across different neighbourhoods. The kind of approach described could help to ascertain whether a policy intervention has been effective or not.

The final class of applications concerns issues around resource allocation. Currently this can be very 'hit and miss' at local geographical scales. Health workers do not always have a clear picture of local needs and so determining training needs or prioritizing services is an inexact science. Similarly with so many agencies operating in an area it becomes virtually impossible to know what the services are costing overall, or the extent of any duplication of effort, particularly in gathering and processing information that would allow greater co-ordination of resources.

The techniques have potential applications that go beyond the scope of this paper. For example, if all data were assembled in this way, dependence on centrally produced ward-based statistics might, in theory, be significantly reduced. Local authorities could construct their own data sets from their administrative data into whatever geographical units were deemed appropriate, if necessary under central guidance. Statistical information produced this way could be more readily updated than equivalent statistics currently disseminated by central government, which are often years out of date. The question is to determine what statistics would be produced more quickly and reliably by this method, what new statistics could be published that are not currently available, and what the conventions would be about making them generally available for research purposes.

References

Acheson, D. (1998) *Independent Inquiry into Inequalities in Health* (The Stationery Office, UK).

Armitage, P. and Berry, G. (1987) *Statistical Methods in Medical Research* (Blackwell, Oxford).

Altman, D. G. (1999) *Practical Statistics for Medical Research*. Chapman & Hall/CRC, London.

Barnett, V. (2002) *Sample Survey: Theory and Methods* (3rd edition, Arnold, UK).

Brackstone, G. (1987) "Issues in the use of Administrative Records for Statistical Purposes" *Survey Methodology*, **13(1)** 29-43.

Brown, R. M. (2002). "The development of family violence as a field of study and contributors to family and community violence among low-income fathers", *Aggression and Violent Behavior* **7(5)** 499-511.

Child, J. S. (2000) "Mapping child maltreatment: looking at neighborhoods in a suburban county", *Child welfare*, **79(5)** 555-572.

Chou, Y. (1972) *Probability and Statistics for Decision Making* (Holt, Reinhart and Winston, New York).

Coker, A.L., Davis, K.E., Arias, I., Desai, S., Sanderson, M., Brandt, H.M. and Smith, P.H. (2002) "Physical and mental health effects of intimate partner violence for men and women", *American Journal of Preventive Medicine* **23(4)** 260-268.

Coombes, M.G. and Raybould, S. (1997) “Modelling the influence of individual and spatial factors underlying variations in the levels of secondary school examination results”, *Environment & Planning A* **29** 641-658

Cunradi, C.B., Caetano, R. and Schafer, J. (2002) “Alcohol-related problems, drug use, and male intimate partner violence severity among US couples”, *Alcoholism: Clinical and Experimental Research* **26(4)** 493-500.

DiBartolo, L. (2001) “The geography of reported domestic violence in Brisbane: a social justice perspective”, *Australian Geographer* **32(3)** 321-342.

Felson, R.B., Messner, S.F., Hoskin, A.W. and Deane, G. (2002) “Reasons for reporting and not reporting domestic violence to the police”, *Criminology* **40(3)** 617-647.

Freund, J.E. (1973) *Modern Elementary Statistics* (Prentice-Hall International, London).

Greater London Authority (2001) *The London Domestic Violence Strategy* (The Greater London Authority, London, UK).

Greenland, S. and Robins, J. (1994) “Invited commentary: ecological studies – biases, misconceptions, and counter examples”, *American Journal of Epidemiology*, **139** 742-60.

Jewkes, R. (2002) “Intimate partner violence: Causes and prevention”, *Lancet* **359(9315)**: 1423-1429.

Mayhew, L.D. (2002) “The neighbourhood health economy: A systematic approach to the examination of health and social risks at neighbourhood level”, Actuarial Research Paper 144, CASS Business School, City University, London.

Mirlees-Black, C. (1998) *Domestic Violence: Findings from a new British Crime Survey self-completion questionnaire* (Home Office Research Study No 192, London: Home Office).

Moreno, C.L., El-Bassel, N., Gilbert L. and Wada, T. (2002) “Correlates of poverty and partner abuse among women on methadone”, *Violence Against Women* **8(4)** 455-475.

Office for National Statistics (2003) *Census 2001 Key Statistics for local authorities in England and Wales* (The Stationery Office, London).

Openshaw, S. and Taylor, P. J. (1981) “The Modifiable Areal Unit Problem”, in N. Wrigley and R.J. Bennett (editors), *Quantitative Geography: A British View*, Routledge, London.

Parascandola, M. and Weed D. L. (2001) “Causation in Epidemiology”, *Journal of Epidemiology and Community Health* **12** 905-912.

Robinson, W. S. (1950) “Ecological correlations and the behaviour of individuals”, *American Sociological Review* **15** 351-57.

Rose, G. (1985) “Sick individuals and sick populations”, *Journal of Epidemiology* **14** 32-38.

Tolman, R.M. and Raphael, J. (2000) “A review of research on welfare and domestic violence”. *Journal of Social Issues* **56(4)** 655-682.

Annex 1: Note on the standard error of observed risk estimates and constructing confidence intervals

In typical risk analysis using these methods the number of observations range from a few to thousands, whilst the observed risk can range from zero to 100%. In repeated samples of size n with the risk of an occurrence equal to r the expected number of cases at risk will be nr with a standard error of $\sqrt{r(1-r)/n}$ assuming a normal approximation to the binomial distribution. Thus, if the sample size is 40 and the observed risk is 27% then the standard error of the risk estimate is approximately $\pm 7\%$ (i.e. 20% to 34%) of the mean based on the graph. See point P in Figure A.1, which is a plot of sample size up to 100 against risk (%).

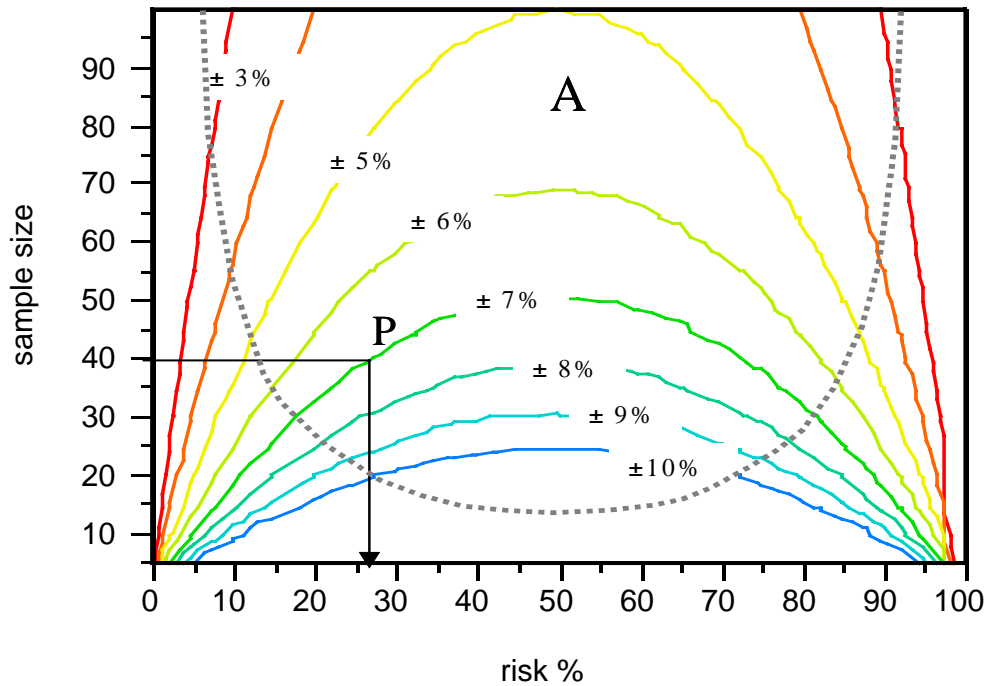


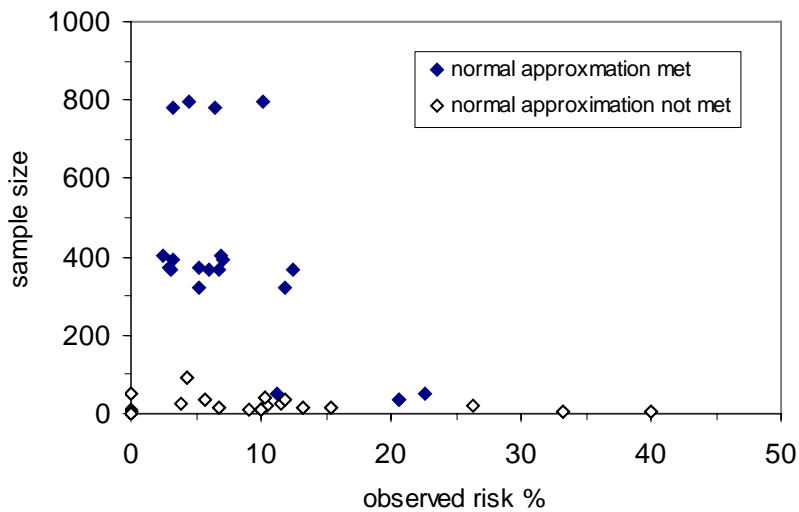
Figure A.1: Graph showing the confidence intervals for different sample sizes and observed levels of risk (measured as a percentage)

It is considered usual practice to use the normal approximation to the binomial distribution only when $nr > 5$ and $n(1-r) > 5$. The boundary condition meeting these criteria is indicated by the dotted line in the domain annotated by A. In this area, confidence intervals based on other specified levels of confidence are generally approximated using:

$$\hat{r} - z_{\alpha/2} \sqrt{\frac{\hat{r}(1-\hat{r})}{n}} < r < \hat{r} + z_{\alpha/2} \sqrt{\frac{\hat{r}(1-\hat{r})}{n}}$$

Where \hat{r} is the observed risk estimate or x/n and $z_{\alpha/2}$ is the number of standard deviations using the standard normal distribution corresponding to a chosen level of confidence $(1-\alpha)100\%$. In the figure above z equals one and the confidence level is 68%, that is the true risk of P in the example above is $27\% \pm 7\%$ with a 68% confidence level. For a 95% level of confidence, substitute 1.96 for z .

Just under half of the 80 risk estimates given in this paper for individual factor combinations generally falls within domain A. Figure A2 is a plot of sample sizes of 1000 or less versus observed risk (0%-50%). The darker points are risk observations that satisfy the normal approximation. Three categories may be recognized: i) large sample sizes of 800 cases or more with low standard errors; ii) sample of between 300 and 400 cases, with slightly higher standard errors; iii) sample sizes of <100 only a minority of which meet the necessary condition. The associated risk in these cases are all >10% with the standard errors of around 7% or higher.



$\hat{r} = 0.25$ with a 95% confidence interval is $0.12 < r < 0.75$, which would be considered quite wide.