Factors That Contribute to the

Transition to MSV Crime by Young People

Data Analytics Lab

January 2020

1 Table of Contents

2 Introduction

This study looks at the driving factors of Most Serious Violent (MSV) crimes in those aged 25 and under, with a specific (and separate) consideration of both knife and firearms crimes. The approach taken here is primarily to discover the main contributory factors that contribute towards a nominal moving from non-MSV crime into MSV crime / use of knives / use of firearms. The analysis is based not on a set of predictive models but rather takes an explanatory approach. It suggests that particular characteristics are associated with a higher chance of being involved in such crimes and that by identifying these, WMP could look to help those at risk of becoming involved. This is potentially advantageous for a number of reasons (if programmes could be successful in steering young people onto a different course):

- The chance of the young person being involved in physical violence would be reduced.
- The probability of officers being injured would reduce if the young person is helped before they display many of the characteristics that increase their probability of becoming involved.
- The costs involved in dealing with the increased harm levels associated with these crimes would be reduced.
- The reduction in the number of people involved in such crimes may be sufficient to reduce the incentives to 'protect' oneself.

There are many apocryphal rationales for knife violence especially:- sentencing, gangs, social media amongst others (The Independent 27 April 2018). The approach used is to consider a number of models, within which one looks for the most important variables and the magnitude of their effects. The focus of study is *youth* crime, i.e. those 25 and under who perpetrate a relevant crime (knife, firearm or MSV in general). We examine the cases up to that point in time. The question is therefore *"what changes occur that a person becomes involved in such crime and can we see a pattern that allows for interventions?"*.

The models will consider a number of the variables at or before the nominal's first occurrence of MSV, including the number of previous offences, by role (victim or suspect/ defendant), the number of nominals associated with that particular crime, whether older or younger, the weapon used as well as family involvement in crime (as a count) with especial consideration of close family.

The potential role of Urban Street Gang membership is considered using intelligence logs to generate metrics for involvement in specific gangs in specific areas. Further consideration of the youngsters' past was also considered using the COMPACT database. Where nominals were known to be missing, information about exposure to CSE, DV, drugs and county lines crimes are also available. The roles of gangs and county lines type movements are captured through the gang involvement and the specific area of that interaction and also a specification of the intelligence logs based on how far they are from the West Midlands. A classification of 'far' is created representing 2 counties and further away to pick up a degree of the cross-county lines movements that is not explicitly considered elsewhere.

The gang influence is based upon a number of mentions in the same log as a known gang name.

There is some contemporaneous information used. This is because certain crime classes are associated with higher levels of violence. Thus if a nominal is involved in say domestic violence then there might be a higher probability of them using serious violence relative to other offence types. It is possible to use this type of information in addition to the past information as the focus is explanation rather than prediction.

The control group was selected from nominals of a similar age with non-MSV type crimes with a similar Cambridge Harm Index value (Sherman, Neyroud, and Neyroud (2016)). The high levels of the CHI for MSV offences meant that the threshold was reduced with sexual violence and other crimes of a similar ilk removed from the control sample, though domestic violence crime flags were still included in the data. This allowed a sample of crimes across the WMP area and younger offenders.

The sample is approximately 80% non-MSV and 20% MSV based, with declining proportions with knife (approximately 3.25%) and firearm crimes (0.62%). The nature of the imbalance in the data is discussed later in the report. However, it is shown in Owen (2007) that only the intercept of the regressions is affected by these imbalances. The underlying regression coefficients are not systematically influenced in any major manner.

One further aspect of this work is to use domains of the Indices of Multiple Depriviation in order to consider the area in which the nominal is living. These give information at the LSOA level of ranks and deciles in the local authorities. These are linked where possible via postcodes to nominals to take into account some level of the broader socio-economic environment that the nominal experiences. This is obviously a rather broad measure and not as granular as one would hope either in space or time. It is however a reasonable proxy for the general state of locations.

The report considers the general background and gives a summary statistical overview of the data, with a focus on particular elements. A brief description of the modelling is given with a more technical appendix included. The results are discussed in light of the overall impact on the offences and nominals involved. There are a number of common elements across each of the explanations of the transition to MSV, knife or firearms crime. These are highlighted as are the commonalities across the specific models. This is due to the potential for greatest and broadest impact of interventions and danger signals from the young people.

In what follows [^1] etc. refer to the notes at the end of this document.

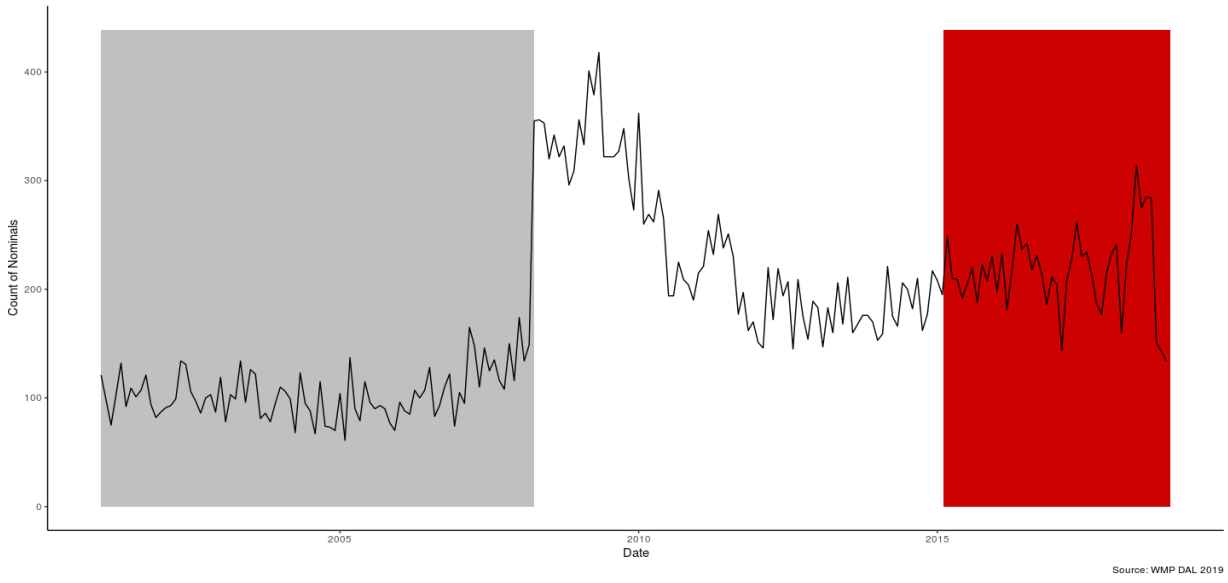
3 Overall MSV Crime Situation

The increase in MSV (and knife) crime amongst the young specifically has been noticeable in the past few years. The number of nominals as defendants, suspects or believed to be responsible for the crime can be seen to have varied considerably since 2001 as presented in the graph below. The data towards the end of the sample is potentially less reliable (increased variance may be introduced due to the various systems taking time to update).

One can see that there have been a number of periods of relatively high growth and there was a substantial leap in April 2008, followed by a decline after this. The average over the period between January 2001 and April 2008 was 104 (100 median) nominals involved in MSV crimes per month and after the break nearly 229 (215 median) nominals (ignoring the break gives an average of 178 & median of 172)[^1]. The greyed out area is carried throughout the initial analysis to ensure that one can see a comparison with the overall picture.

Considering the changes in the MSVs, specifically knife and firearms crimes as well, three periods are specified- early pre-April 2008, late post February 2015 and the middle. We can see that there is an interesting situation where there is a particular increase in MSV and MSV participation in the middle period. In this case the mean number of nominals involved in MSVs per month was 104, 236 and 217 with the medians of 100, 210 and 218 in the early, middle and late periods respectively. The large increase circa 2008 and the difference in the median and means in the middle period suggest that the increase in the mean is due to differences in procedural or reporting processes rather than a structural change in the underlying process.

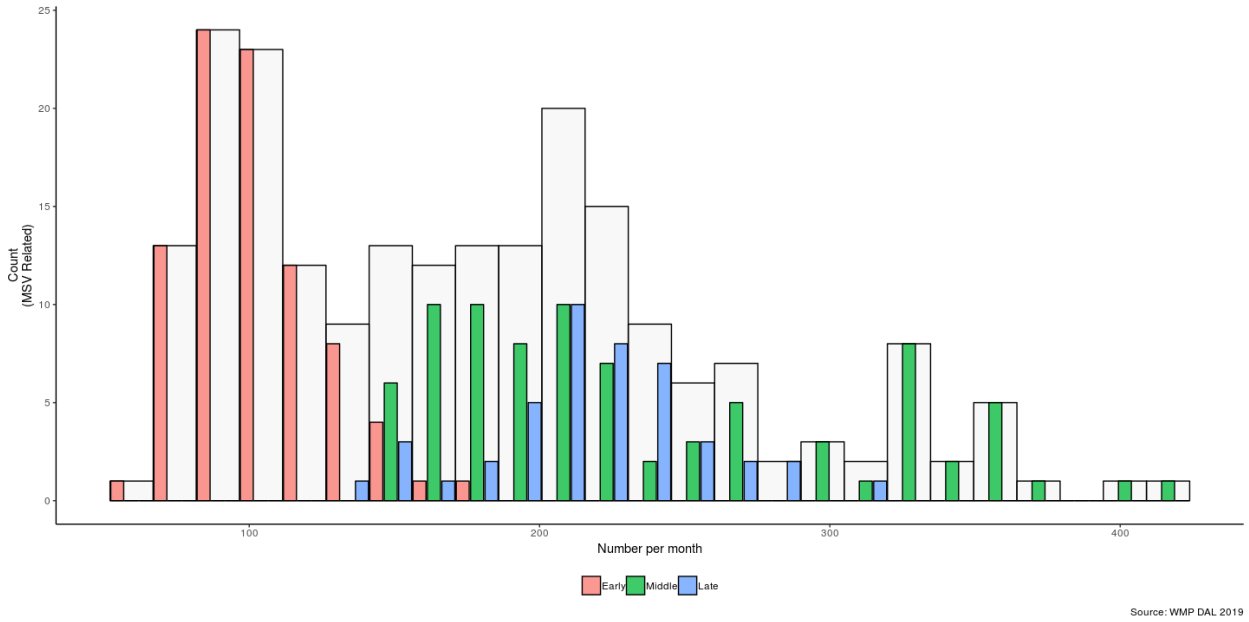
Nominals becoming Involved in MSV Crime



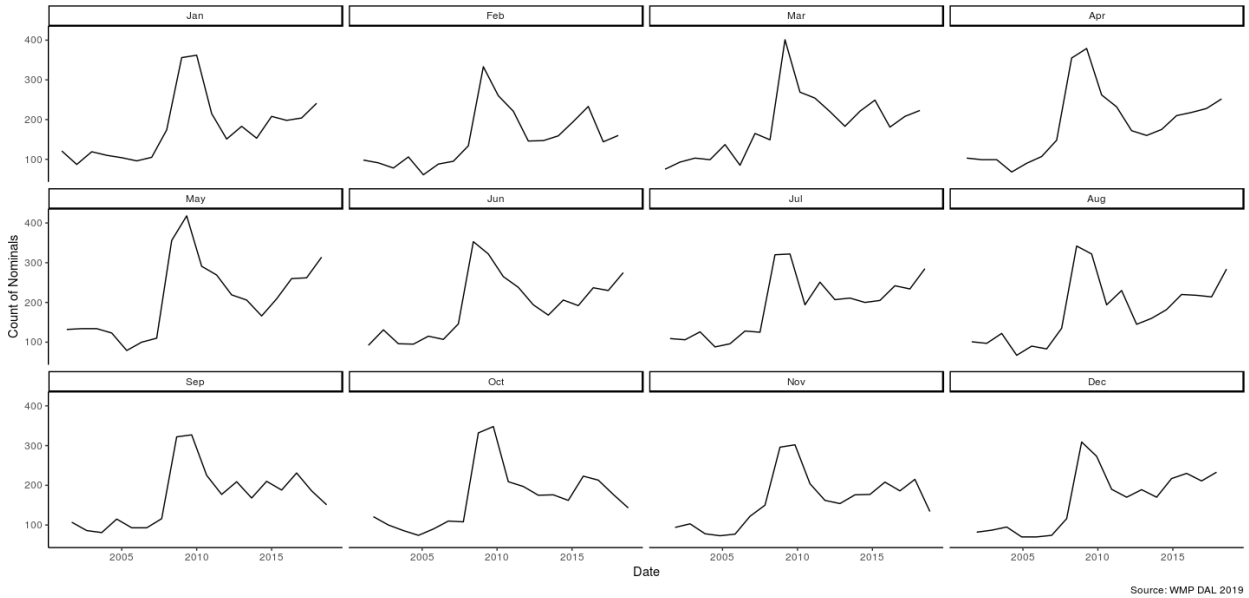
The graph below shows the distribution of the number of nominals involved in MSV crimes. The *grey* area is the total count per month, with the coloured columns representing counts

in each of the various time splits described above. It can be seen that there has been a shift in the number of MSV related crimes throughout the period with the middle regime being more extreme than the other two.

Number of MSV Related First Offences



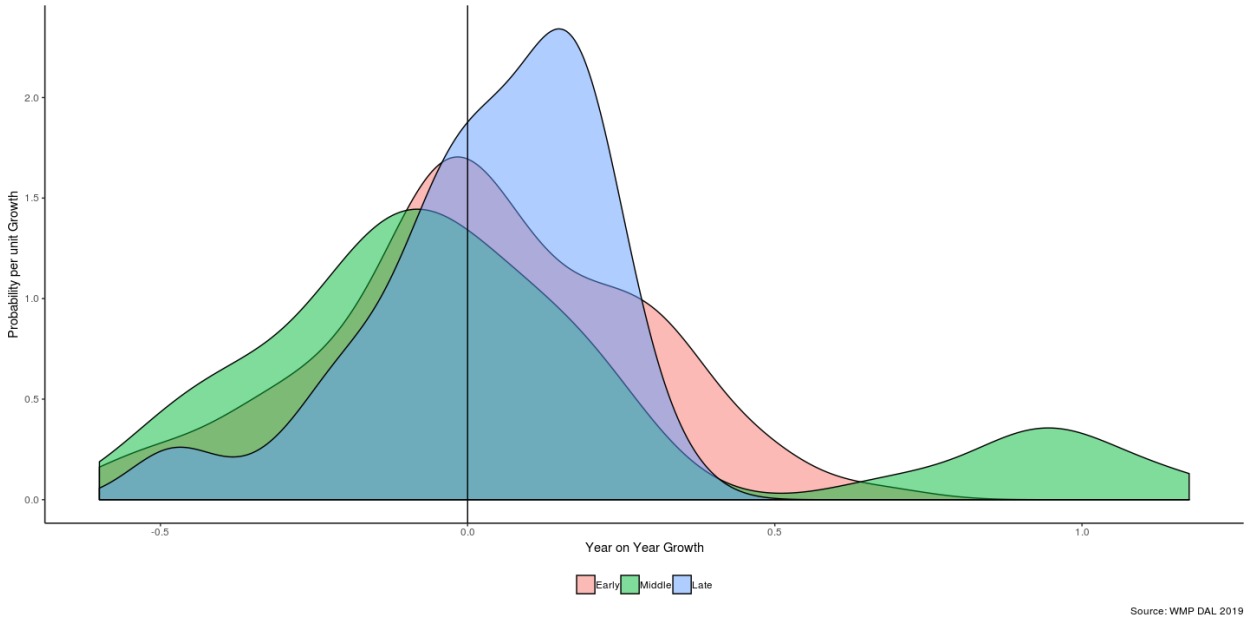
The monthly pattern, plotting counts by month as well as by year, shows some seasonality with March to May being a little higher than the other months. This is presented below.



The structural breaks in April 2008 and February 2015 can be seen using a histogram of the year on year growth rate; the densities presented below normalises this to take into account the fact that there are more observations after April 2008. One can see that there have been more *extreme* growths year on year after the April 2008 split. It is clear that the

growth rate in the later period is positive and the middle period is flat to negative depending on the metric used. The extreme growth periods in the middle section should likely be seen as outliers and their influence discounted to some extent due to the apparent step change (at least in part likely being due to changing recording practices).

Densities of distribution of Growth Rates by Period



3.1 Age of First Offences & Gender

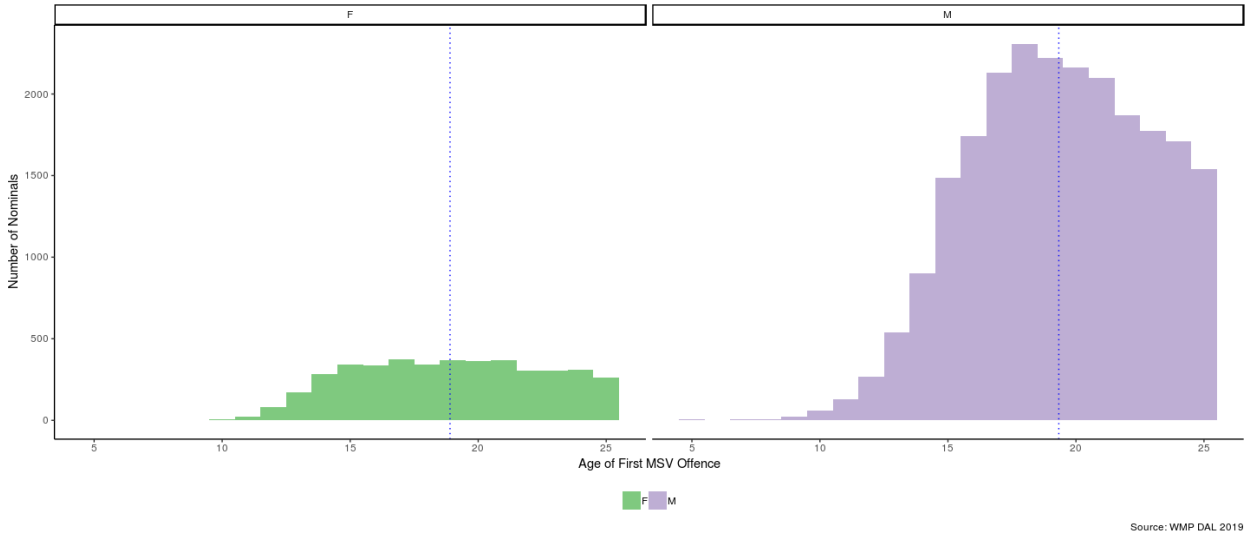
The data considered looks at young offenders. Looking at the post 1999 data, we can consider the ages of those involved in the respective crimes. In aggregate the average age for nominals is around 19 (both mean and median), with gender making little practical difference (males have a mean average 4 months later than females). There are (as might be expected) considerably more male nominals involved in MSVs than females and those who are unknown account for about 0.25% of the total. This initial data considers those involved in the event, thus includes those believed to be responsible as well as suspects and defendants.

Ages of Nominals Involved in Crimes

	SEX	MEAN	MEDIAN	VAR	MIN	COUNT	PROPORTION
MSV	M	19.330	19.000	11.970	5.000	20635	83.840
	F	18.900	19.000	13.310	5.000	3940	16.010
	U	19.260	18.500	12.790	14.000	38	0.150
	TOTAL	**19.260**	**19.000**	**12.210**	**5.000**	**24613**	**100.000**
Non-MSV	M	18.420	18.000	15.360	0.000	71494	75.330
	F	17.840	17.000	15.640	0.000	23218	24.460
	U	18.570	19.000	14.620	10.000	193	0.200
	TOTAL	**18.280**	**18.000**	**15.490**	**0.000**	**94905**	**100.000**

We can see the underlying distributions and the differences in the gender splits most easily from the graph below. The sheer magnitude of difference is clear in the table above. There are a higher proportion of males involved in MSV relative to the non-MSV crime, with over 8 in 10 involved being male as opposed to 3/4 for non-MSV crimes.

Age Distribution of Nominals Involved in MSV Crimes



The age distribution of those involved in their first MSV crime is seen above and is much as one would expect, given the emphasis on the young. The relatively low level of female MSV offences are reflected in the flatter age distribution, which suggests that age is less important for females in general. On the other hand, males have a median age of 19 and a mode slightly below this. This suggests that males' ages have an impact on their likelihood of moving towards MSV. This would also tally with the anecdotal belief that initiation in gangs and the influence of that age group can have an impact primarily on boys in their late teens.

Excluding those not charged, we see the same picture as those involved, with males being a higher proportion of those charged and higher than those of non- MSV crimes.

Ages of Defendants Involved in Crimes

	SEX	MEAN	MEDIAN	VAR	MIN	COUNT	PROPORTION
MSV	M	19.260	19.000	11.850	7.000	14296	84.270
	F	18.620	19.000	12.920	9.000	2638	15.550
	U	19.030	18.500	11.410	14.000	30	0.180
	TOTAL	**19.160**	**19.000**	**12.070**	**7.000**	**16964**	**100.000**
Non-MSV	M	18.360	18.000	15.180	0.000	58249	76.280
	F	17.680	17.000	15.210	6.000	17967	23.530
	U	18.440	19.000	15.180	10.000	147	0.190
	TOTAL	**18.200**	**18.000**	**15.270**	**0.000**	**76363**	**100.000**

3.2 Ethnicity

As part of the exploratory data analysis element of the project we have examined ethnicity irrespective as to its inclusion as a variable in the models.

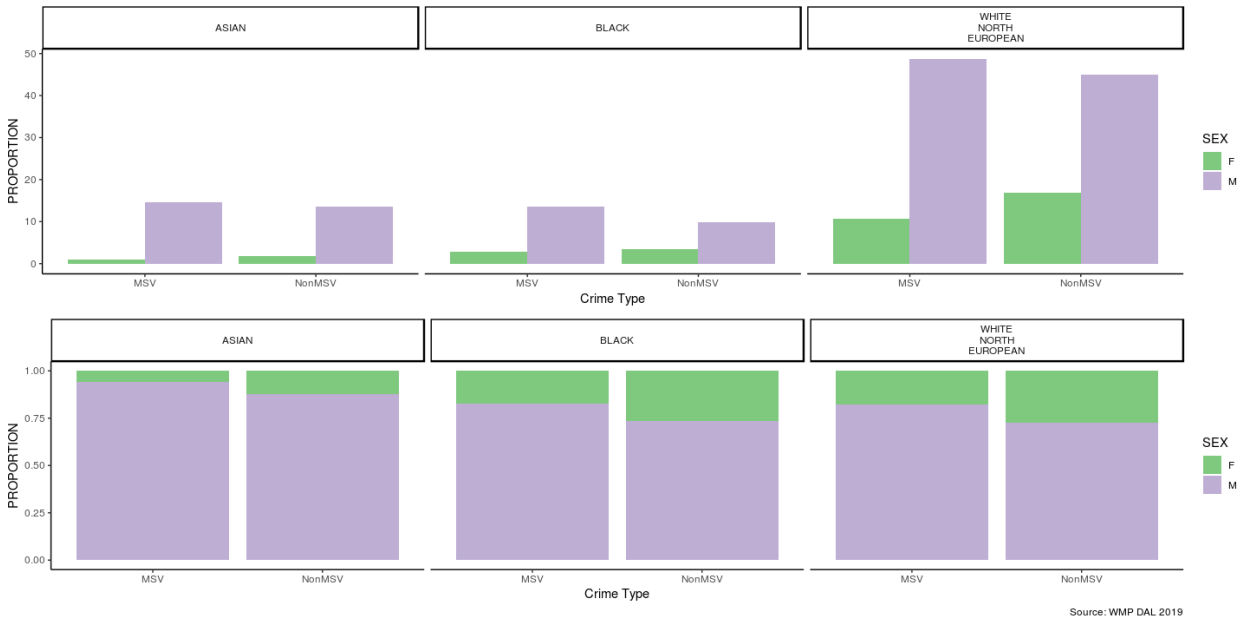
We can see from the table below that the ages are in all but one case the same for the first offence be it MSV or not. We can see that when we consider the ethnic split in terms of gender and sex, Black males constitute 13% of the MSV crimes as opposed to 9% of the non-MSV crimes. White Northern European males have a similar picture 45% against 49%.

Proportions of Nominals Involved in MSV Crimes by Ethnicity and Gender

	Ethnic Background	Proportion Female	Proportion Male	Count
NonMSV	ASIAN	1.880	13.490	14559
	BLACK	3.490	9.780	12563
	CHINESE	0.050	0.160	196
	MIDDLE EASTERN	0.030	0.350	352
	NOT KNOWN	1.510	4.790	5962
	OTHER	0.530	1.260	1694
	WHITE NORTH EUROPEAN	16.920	45.020	58659
	WHITE SOUTH EUROPEAN	0.120	0.650	727
	TOTAL	**24.510**	**75.490**	**94712**
MSV	ASIAN	0.950	14.700	3845
	BLACK	2.890	13.640	4064
	CHINESE	0.010	0.130	36
	MIDDLE EASTERN	0.020	0.400	104
	NOT KNOWN	0.900	3.930	1187
	OTHER	0.520	1.920	600
	WHITE NORTH EUROPEAN	10.650	48.670	14577
	WHITE SOUTH EUROPEAN	0.090	0.570	162
	TOTAL	**16.030**	**83.970**	**24575**

This table shows the split by MSV, ethnicity and gender. The subsequent graph below shows this more clearly. For each of these groups, males are a higher proportion of those involved in MSV crime.

Proportions of Nominals Involved in MSV Crimes by Ethnicity and Gender



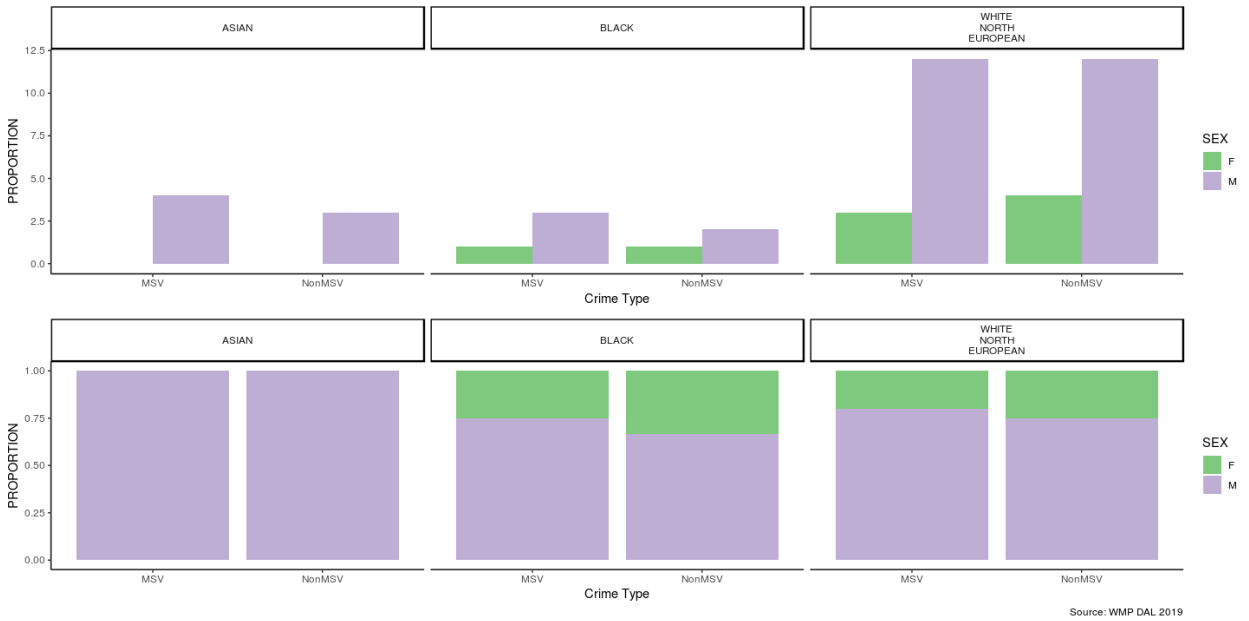
As before, there is a discrepancy in the gender split after charging, though as can be seen in the table the numbers are low. The data below is based on Defendants only. These parallel the data for those involved in the (larger set) of roles in crime noted above .

Table Showing Proportions of Defendants By Ethnicity & Sex

Sex	Ethnic Background	Proportion MSV	Proportion Non-MSV	Count
F	ASIAN	0.870	1.700	1440
	BLACK	2.770	3.500	3140
	CHINESE	0.010	0.050	41
	MIDDLE EASTERN	0.020	0.020	22
	NOT KNOWN	0.900	1.270	1124
	OTHER	0.530	0.450	435
	WHITE NORTH EUROPEAN	10.370	16.460	14298
	WHITE SOUTH EUROPEAN	0.100	0.120	105
	TOTAL	**15.580**	**23.570**	**20605**
M	ASIAN	14.770	13.630	12891
	BLACK	13.480	9.820	9764
	CHINESE	0.170	0.190	170
	MIDDLE EASTERN	0.430	0.360	349
	NOT KNOWN	4.170	4.430	4082
	OTHER	1.740	1.170	1189
	WHITE NORTH EUROPEAN	49.040	46.140	43473
	WHITE SOUTH EUROPEAN	0.640	0.680	627
	TOTAL	**84.420**	**76.430**	**72545**
TOTAL	ASIAN	15.640	15.330	14331
	BLACK	16.250	13.320	12904
	CHINESE	0.180	0.240	211
	MIDDLE EASTERN	0.440	0.390	371
	NOT KNOWN	5.070	5.700	5206
	OTHER	2.270	1.630	1624
	WHITE NORTH EUROPEAN	59.410	62.600	57771
	WHITE SOUTH EUROPEAN	0.740	0.800	732
	TOTAL	**100.000**	**100.000**	**93150**

The proportion of each gender being involved in and being charged for these crimes is similar to those seen in the previous table. This would suggest a relatively flat scaling between these groups with relatively fixed numbers progressing from involvement to becoming defendants.

Split of Crime Types by Gender and Ethnicity for Defendants and Offenders



Considering the independence of these factors is important. As a measure of the relationship between the variables of interest, the Uncertainty Coefficient (aka Theil's U or entropy coefficient (Theil (1971))) tells us, that given one variable (say that the nominal is male) how much of the other variable (e.g. involvement in MSV) can we explain. This is an asymmetric measure (unlike correlations) and so a weighted average is often also reported to give an overall relationship. Looking at the split in the data based upon post 2000 and gender not unknown (i.e. only those reported as male or female) we can consider the information given by various groups of variables. The measure has a range from 0 to 1 (with 1 being perfect information provided). The measure reported as `column` is the uncertainty associated with the column variable, given the row variable.

Table Showing Theil's U Statistic

	Variable	Column Conditional on Row	Combined Measure
Defendants	Sex	0.007	0.006
	Ethnicity	0.001	0.001
	Both	0.012	0.015
Involved	Sex	0.008	0.007
	Ethnicity	0.001	0.001
	Both	0.011	0.013

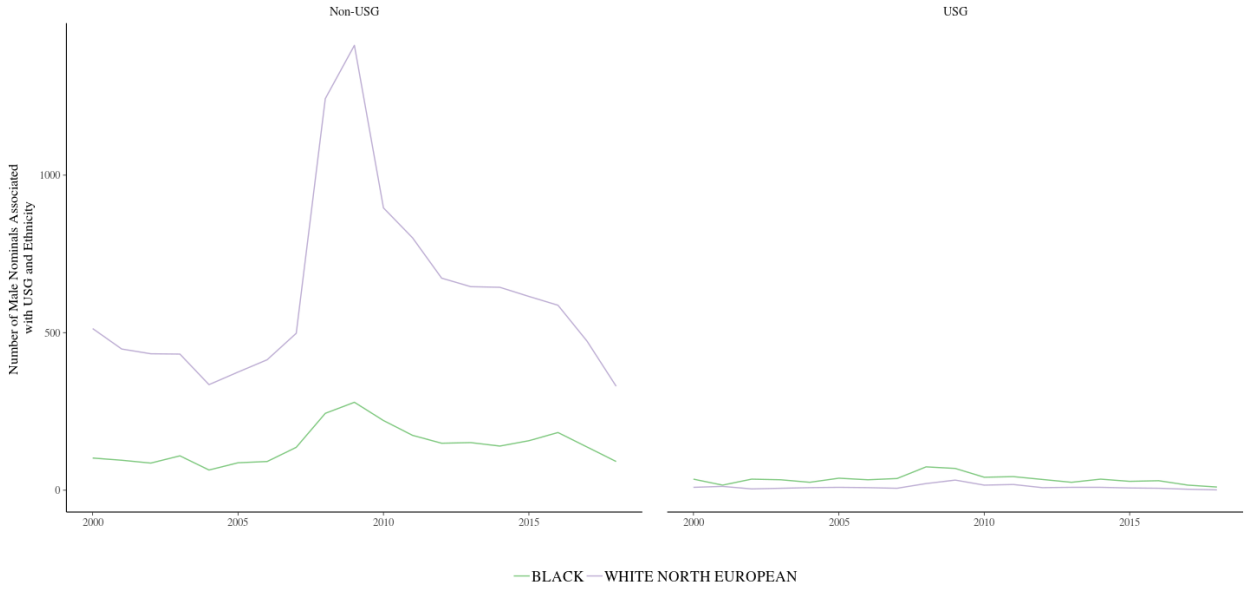
The low level of information held by the individual variables when compared with the combination is useful. They are not the whole story, however they do appear to contribute a minor amount of useful information.

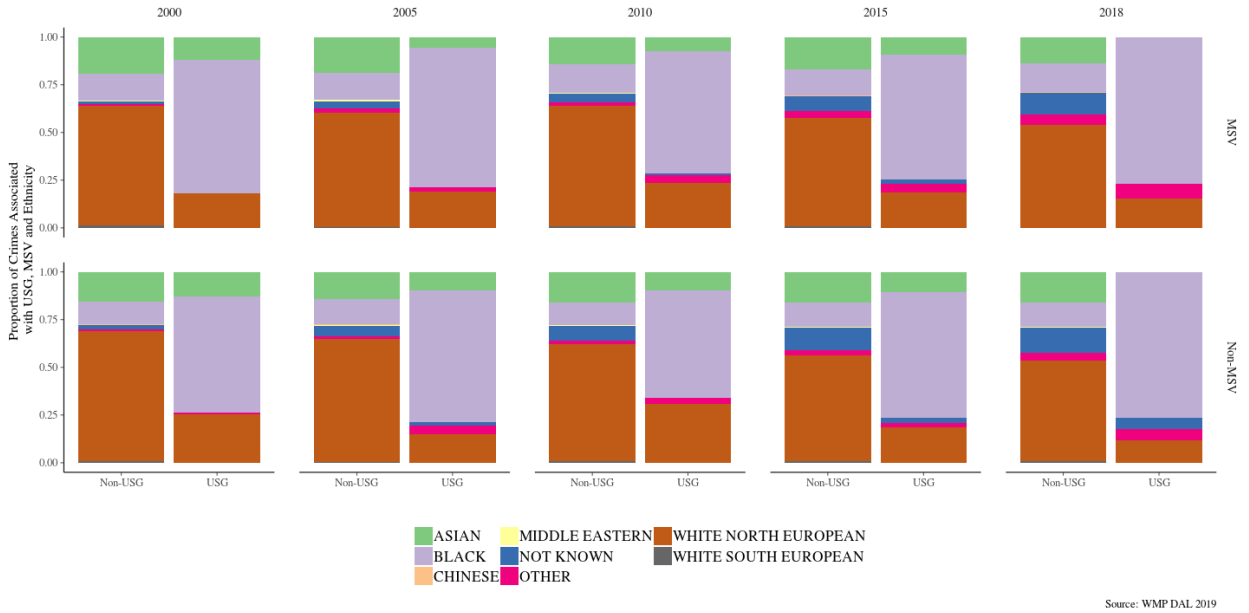
3.3 Gangs

The Urban Street Gang is considered as one of the main drivers of the recent increases in violent crime. Gang names are potentially changing as gangs change and there is no definitive list of these organisations. The IMS logs were used to extract details in which "USG", "Urban Street Gang" or similar was mentioned. These details were searched for also using further key words associated with gangs, such as "member", "solja" or "crew". The relevant details were then n-grammed to a 2-5 word n-gram. This was used so as to avoid ignoring USGs with potentially longer names and to minimise the false positives.

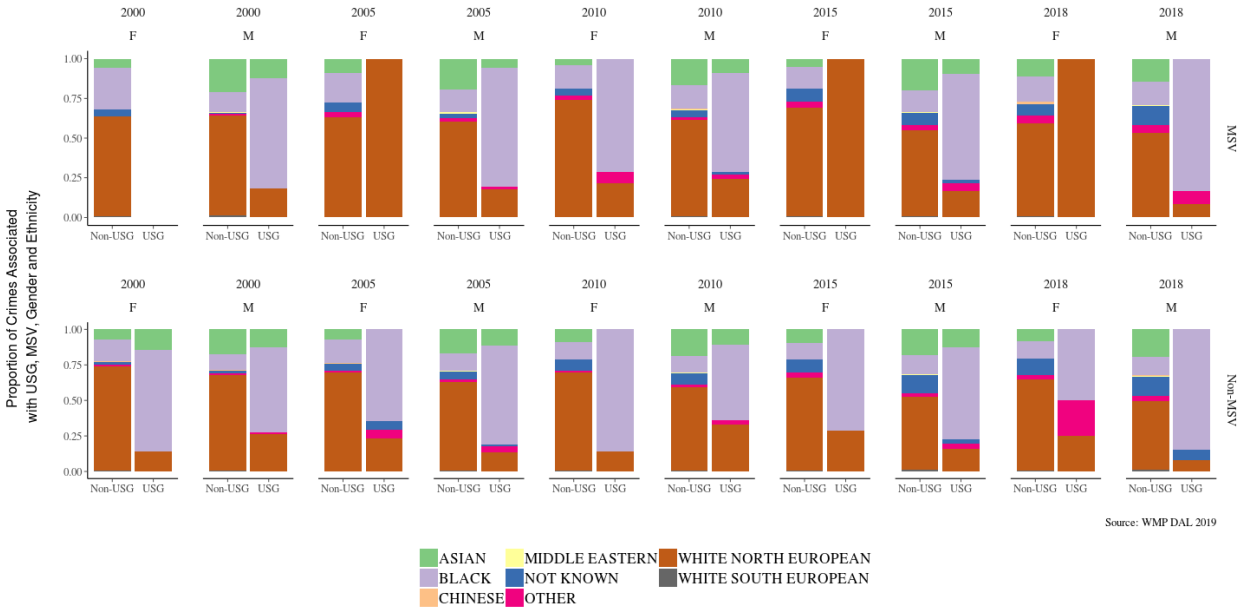
The n-grams were analysed using a collocation algorithm with point wise and log-frequency biased mutual dependency measures calculated. This information allowed the extraction of the main USG names, which have since been sent to a subject matter expert (SME) for consideration. This approach allowed the extraction of the name of the gang if it was multi-part as the n-grams including and up to the name would tend to co-locate. Words after the name tend to be more variable. This was used to link nominals in the IMS logs to entries that mention these gangs. A count was used to determine the amount of association with a particular USG. Forty five gangs were identified though at least 1 was believed to be a replication/ synonym of another name. These were amalgamated. The sparsity of the data relative to the number of nominals meant that USGs were linked to NPUs rather than kept separate. The number of activities in the NPU was used as a metric representing the involvement in gang activity in a particular area.

The story that can be seen here is one of relatively high membership of USGs by black young males who are then undertaking most of the MSVs associated with that demographic. The baseline number of nominals involved is relatively low in comparison to North Europeans, however the USG aspect is relatively high and has been stable over the last two decades. The number of crimes of interest associated with USGs has actually remained relatively constant over the period of interest.





We can see that there is a predominance of black young people in gangs that are involved in both MSV and non-MSV crimes as their first crime. The proportion of these nominals is far greater than in crimes not associated with gangs. There has been a relative reduction in Asian involvement in gangs.

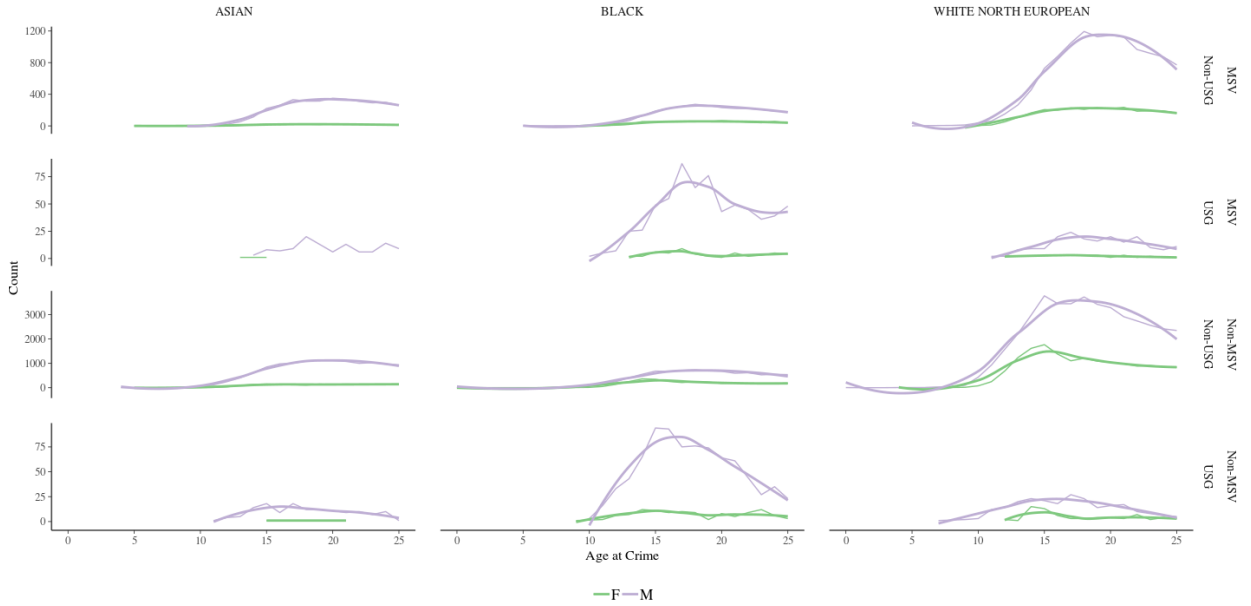


The gender split is also interesting as seen above. Young, black females involved in USGs and MSV are rare, with blips in 2008 and 2010, which also saw higher levels generally amongst black males too. This is the case with non-MSV crime as well, though the peaks are earlier (2002, 2003 and 2005). This pattern is repeated for White North European females less involved in USG crimes. Thus as has been discussed previously, crimes associated with USGs are predominantly committed by males. It should be noted that the numbers of girls associated with USG crimes is generally very low as can be seen from the able below.

Table Showing the Split between The Two Most Involved Ethnicities in MSV (count)

		2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
MSV	BLACK	0	1	1	4	2	0	2	2	9	3	10	1	6	2	4	0	2	1	0
	WHITE NORTH EUROPEAN	0	2	0	2	0	1	3	0	1	6	3	1	1	0	0	1	0	2	1
Non-MSV	BLACK	5	8	10	14	6	11	8	8	6	7	6	7	1	4	1	5	2	1	2
	WHITE NORTH EUROPEAN	1	6	4	8	6	4	5	7	6	4	1	4	1	2	1	2	6	1	1

From the examination of the data, there is a concentration of USG violence amongst black males. The proportion of the crimes associated with MSV and USGs are considerably higher for this group than any other.



The modal ages for the demographic splits are mostly similar with the exception of Asians where the females in particular are involved at a younger age with gangs (though this should not be taken with any great degree of reliability given the *very* low number of observations)

Table Showing The Modal Age of MSV by Gender and Ethnicity Conditioned on Gang Involvement

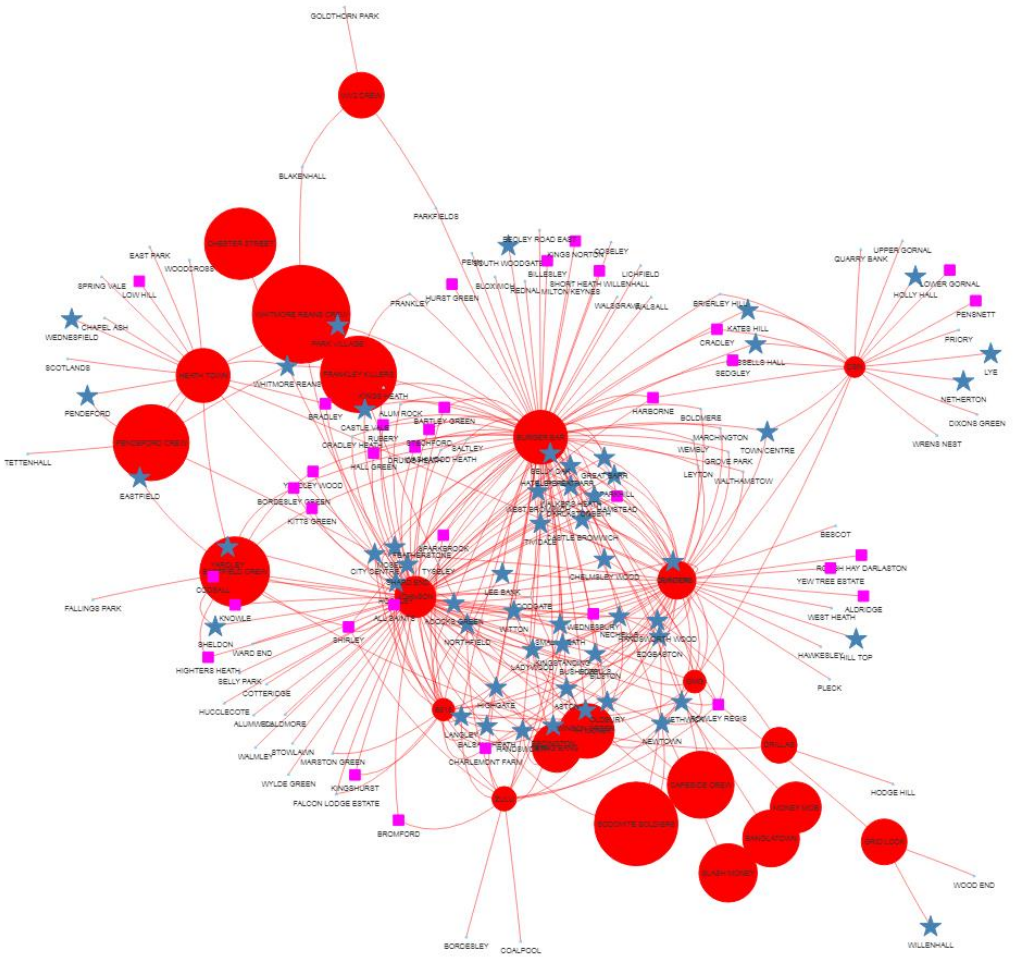
	Ethnicity	Female MSV	Female Non-MSV	Male MSV	Male Non-MSV
Non-USG	ASIAN	20.000	15.000	20.000	19.000
USG		13.000	18.000	18.000	17.000
Non-USG	BLACK	20.000	14.000	18.000	18.000
USG		17.000	14.000	17.000	15.000
Non-USG	WHITE NORTH EUROPEAN	21.000	15.000	18.000	15.000
USG		21.000	14.000	17.000	17.000

Cells in Red are those with fewer than 5 counts
Those in Purple have between 5 and 10

3.3.1 Additional Outputs from Gang Identification

An additional spillover of the work to extract the data for the USGs is that these are now connected with nominals and locations in a manner that gives rise to the creation of a network of nominals and gangs. The current analysis gives links between the gang and the nominal *without* any direction of involvement - currently this approach takes no account of the words around the gang just that it is there; a victim and perpetrator will be considered as both having a link to the gang. This can be mitigated by increasing the number of mentions above a threshold - it is *hopefully* unlikely that someone would be mentioned as a victim more than once. In light of this it is possible to generate a network of nominals involved in the USG scene in the West Midlands. The graphic below shows the main interactions of nominals with the gangs. Where an individual is linked to two gangs in some way this creates an edge between the nodes. The gang nodes are central to the main clusters. One can see that football gangs such as Zulu are separated from the more *traditional* USGs. Similarly, Wolverhampton gangs are limited in their contact with gangs based in Birmingham. The Johnsons and Burger Bar Crews have a lot of interactions/ nominals in common - this would appear to confirm their adversarial nature. The first graph relates nominals to the gangs with 3 or more mentions in the IMS logs. Stars represent more mentioned nominals, and so potentially more *important* people in the gang scene.

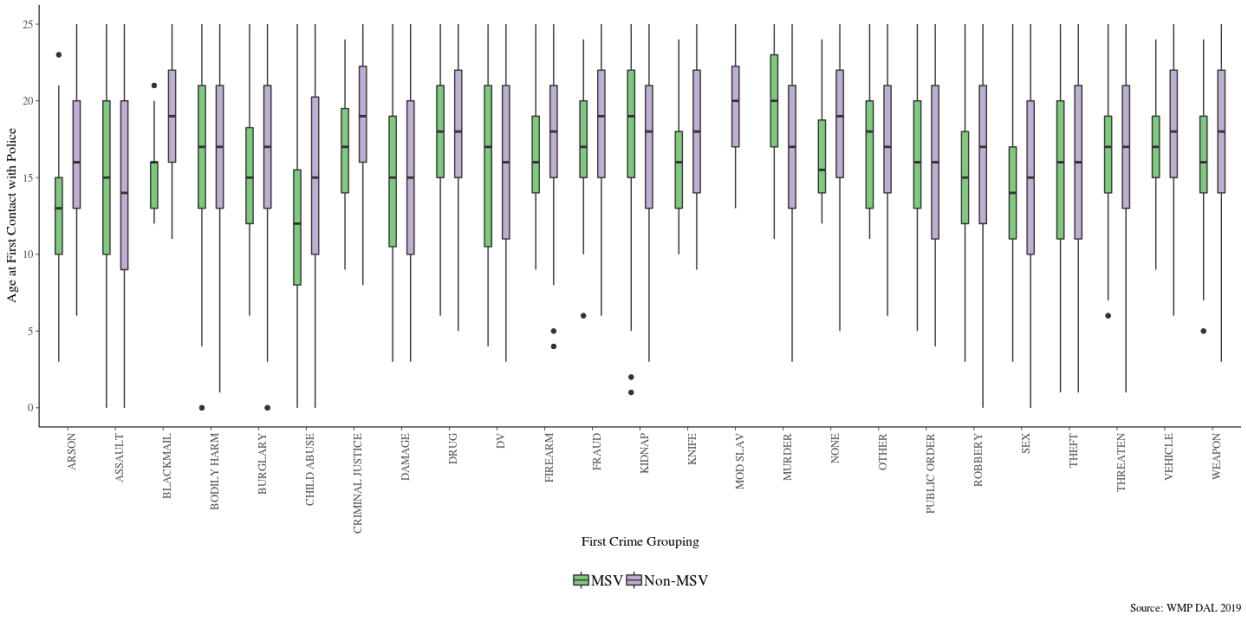
Using a similar technique we can use the locations where the gangs are mentioned as a node. This helps identify the main areas of activity and interaction between the USGs. This is shown in the network below, where the base level is more than 15 mentions in the IMS logs. These identify the main *patches* for the USGs, which though known by SMEs and local members of the force can highlight areas of contention or slightly less ostentatious USG behaviours.



3.4 First Crimes & Age of First Crime

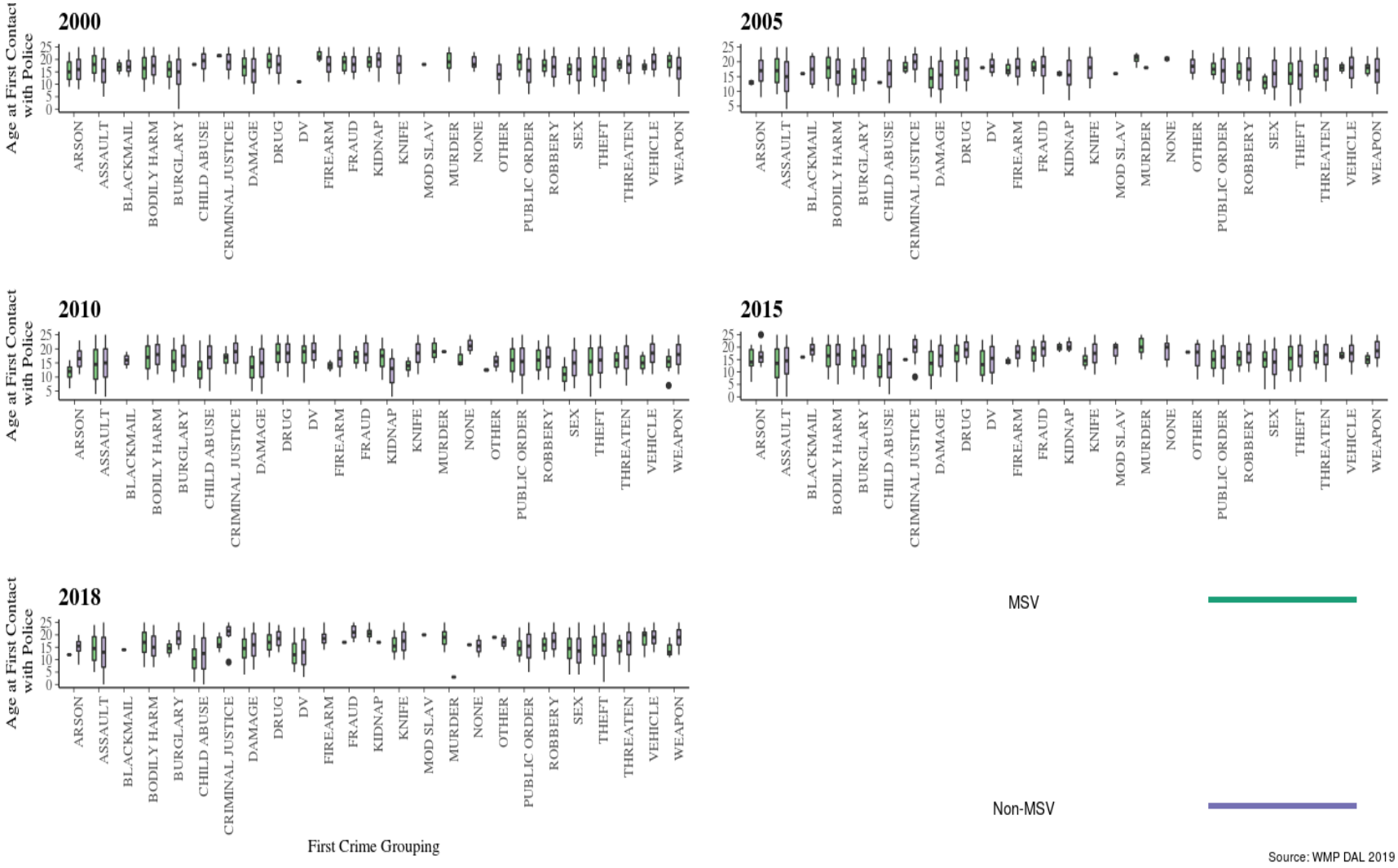
Previous work has shown that in many cases there is a relationship with MSV and an early commencement of interactions with the police. The crimes were mapped into 24 broad categories based upon the crime reported (with an additional *NONE* category). If these factors are found to be important, it gives a natural point of intervention (it should always be borne in mind that if interventions were successful, that over time the nature of the relationship between various factors and MSV would change).

Distribution of Ages of First Contact with the Police by Grouping



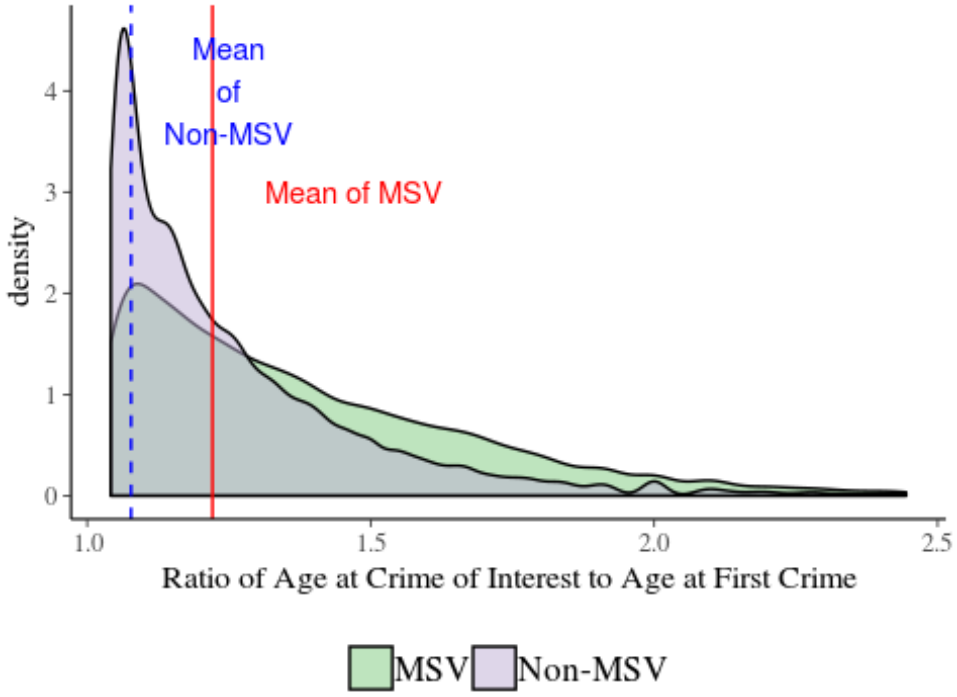
There are some notable points here; that there is little difference in the average ages by MSV except in the case of Arson, Blackmail and the Abuse of children. This might be due to the classification and will be followed up with SMEs. This is the aggregated data. The split by year shows that this is an unstable measure.

Age of First Contact with the Police Conditioned by Major Crime Type



The correlation of the age of the first offence and the age of the MSV is 0.544. There is substantial deviation in the move from the initial crimes to MSV. The ratio of the age at the MSV crime to age at the first crime is informative. The further the ratio

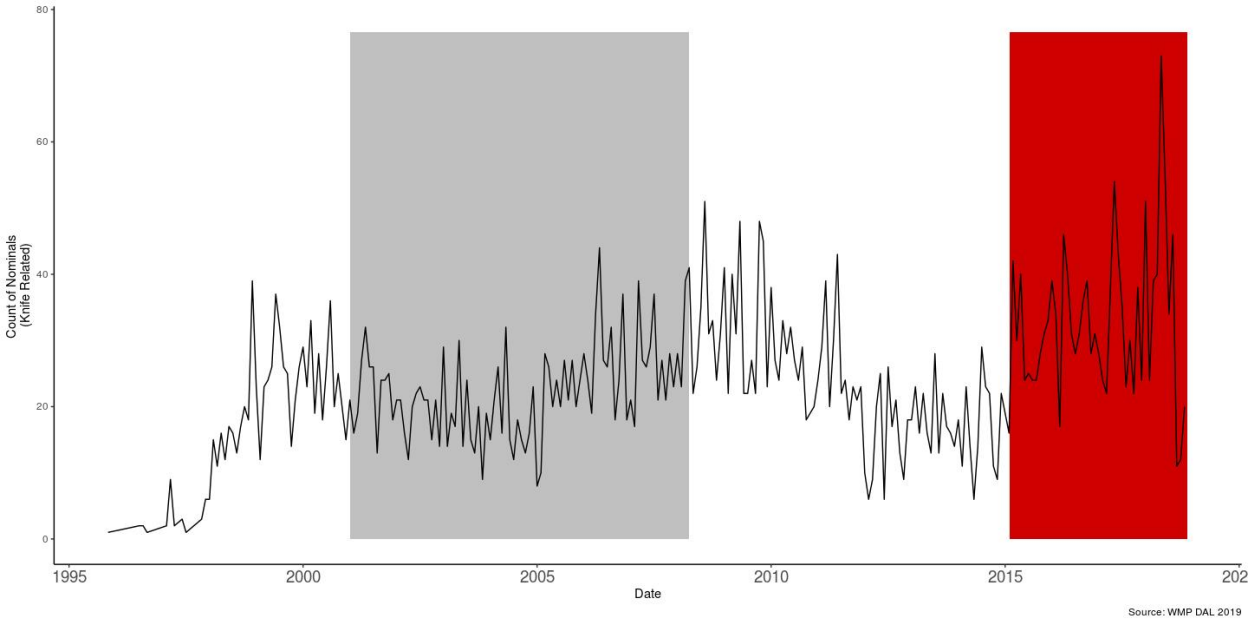
is from 1, the later the MSV occurs relative to the nominal's first crime. The graph presented below shows the differences in the means and the spread in general.

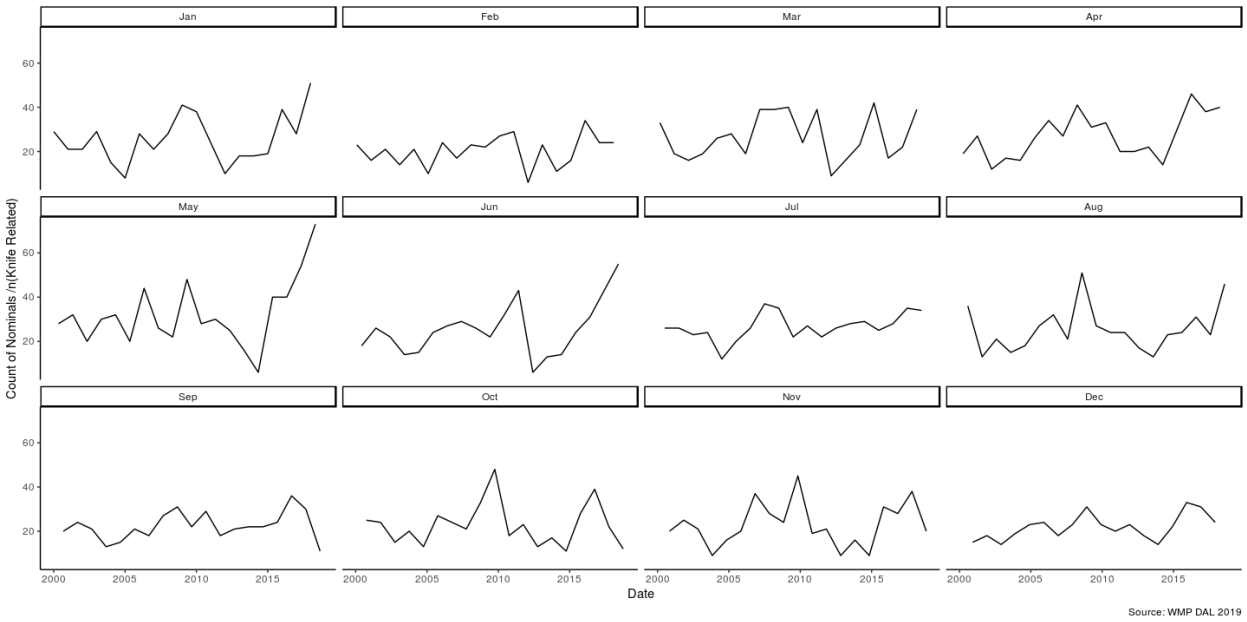


Source: WMP DAL 2019

4 Knives

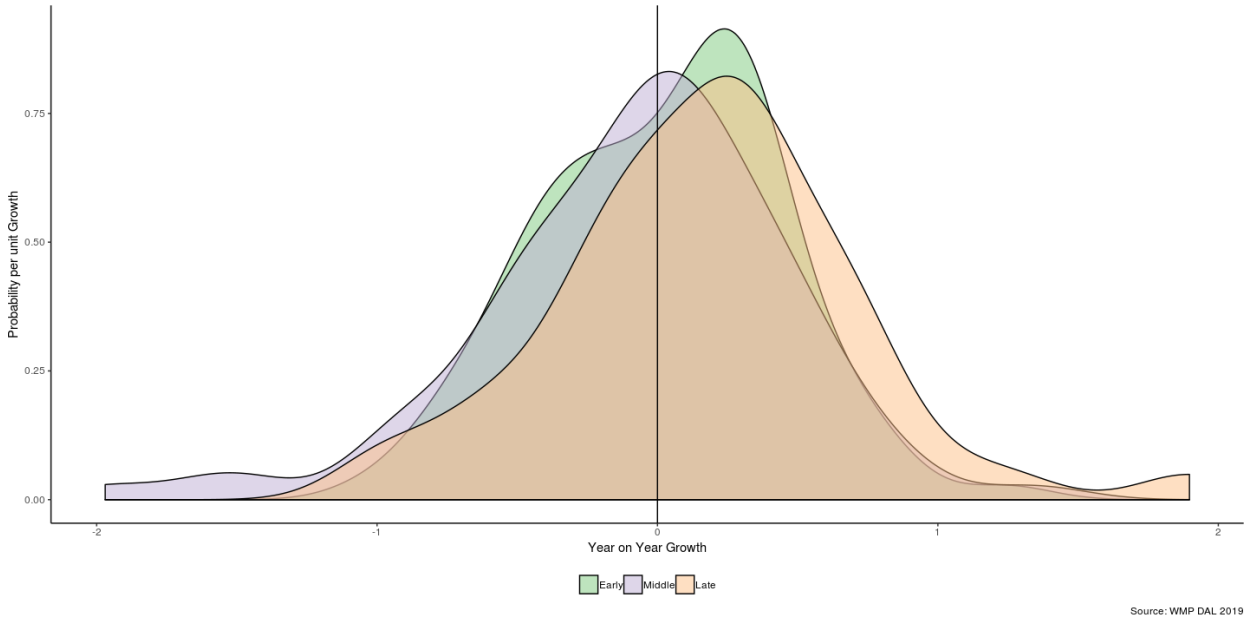
The level of knife crime has seen a slight downward trend since the turn of the century, with the exception of the period after February 2015, when a different step change occurs. The change in the original total / MSV numbers is not important (qv. Footnote 1). Using the three periods, there is an increase in the average from 22 (21) to 24 (22) and finally to 33 (31) in the number of nominals involved in knife crime (medians in parentheses)





As previously, the year on year changes are interesting. We can see that there is an upward, though small shift to the right in the histogram and densities in the Post February 2015

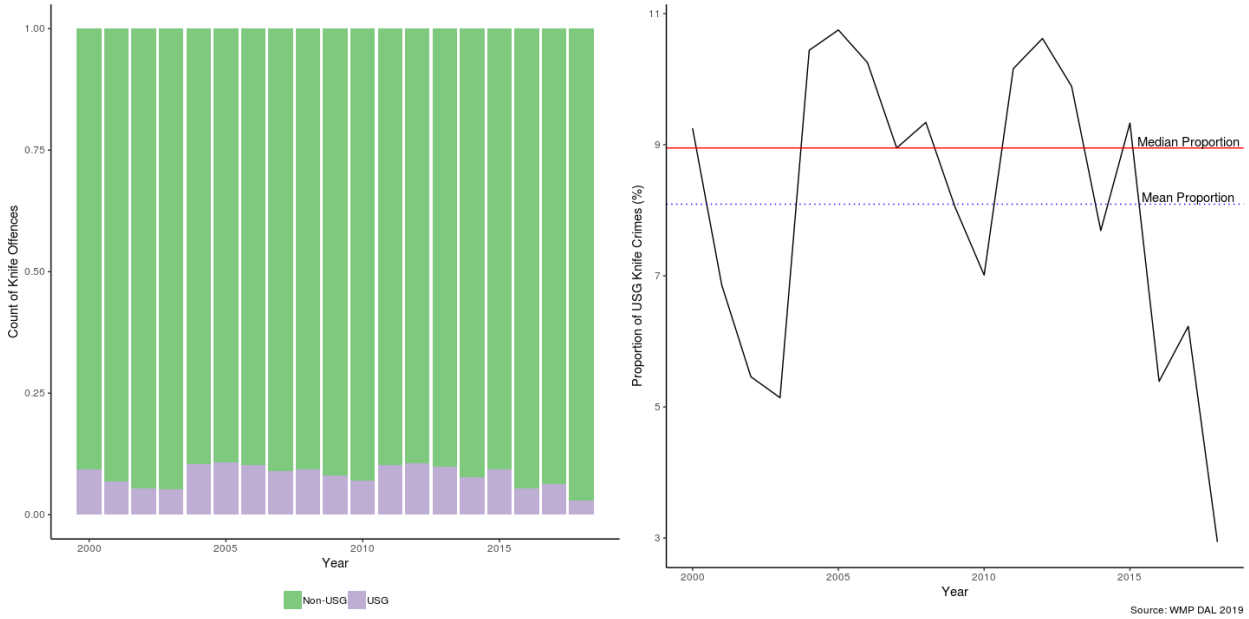
period, though the growth rate year on year looks similar to that of the earlier period. There is a noticeable increase in the growth rate from the middle period until today. The median growth rates for the various periods show the same story more starkly with the year on year growth for the later period being 0.25, for the middle period 0 and for the earliest period 0.07. When using the month on month growth rates (percentage change), which are inevitably more volatile the median growth rates are 0 for both of the earlier periods and 2.53 for the latter. This suggests that there is increasing growth in knife usage amongst these nominals through time and the growth rate itself is increasing.



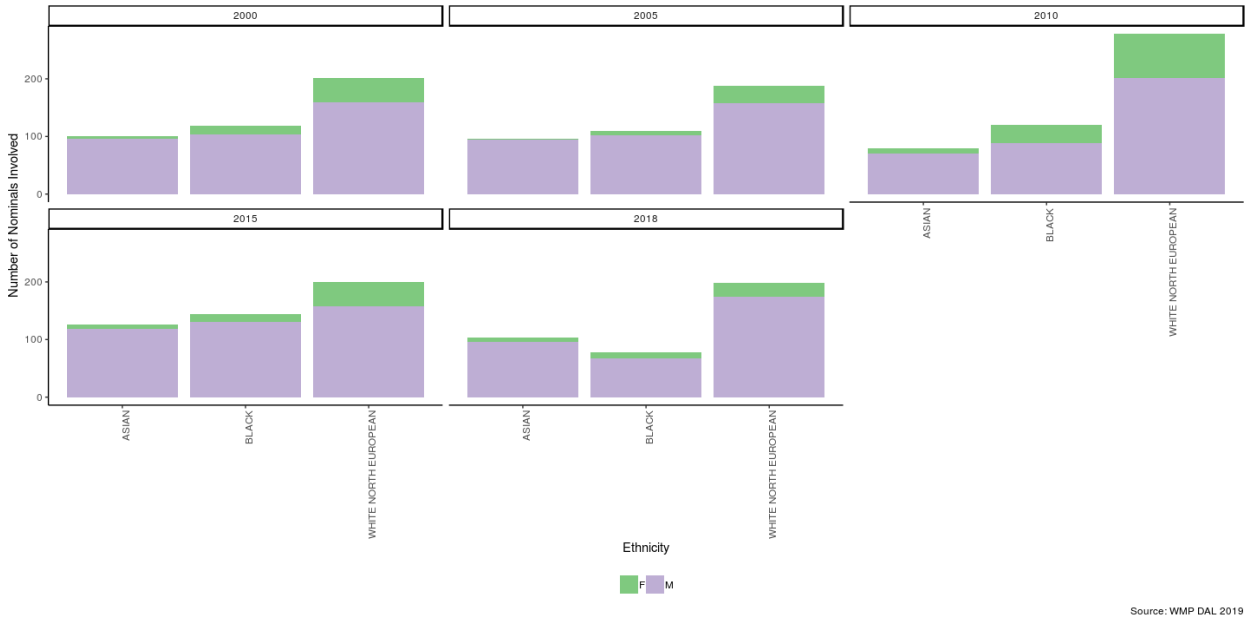
Discussions with SMEs suggest that knife crimes especially are associated with USGs. Looking at all MSVs, the Urban Street Gang influence is 4.47% rather than 8.05% in the case of knife crime (there is a flag relating to USG in the data). There is a higher proportion of USG involvement in knife crime than MSV in general (though it should be noted that there are fewer knife crimes and so the numbers are smaller in absolute terms).

We can see the development of knife based MSVs looking at the following graphs. Note that the 2018 numbers include only up to November and so should be discounted to some extent, not least due to the increases in 2019.

Knife Crime Associated with USGs



As was seen previously, the ethnic split of the USGs is not uniform (There is also an Unknown ethnicity which has not been included in these numbers). Throughout the period considered, there are very few (1 or 2) females involved in USG based knife crimes whereas amongst non-USG there are on average 7 involvements (mean 7.895) for Black females compared with a median of 20 (mean of 21.21) for white north Europeans. Asian females are lower than both of these groups with a median of 3 and mean 2.94 for non-USG knife violence and no USG-connected violence.



The story is different for Males, where USG based knife crime accounts for 25% of the offences.

Table Showing the Male Knife Crime Statistics for Specific Ethnicities

	Ethnicity	Count	Proportion By Ethnicity
Non-USG	ASIAN	1628	95.990
	BLACK	1436	74.170
	WHITE NORTH EUROPEAN	3134	97.150
USG	ASIAN	68	4.010
	BLACK	500	25.830
	WHITE NORTH EUROPEAN	92	2.850

The story is similar for Defendants and Offenders, rather than just those involved. The data barely changes with perhaps 1% more of those as defendants.

Table Showing the Male Knife Crime Statistics for Specific Ethnicities (Defendants / Offenders)

	Ethnicity	Count	Proportion By Ethnicity
Non-USG	ASIAN	506	96.380
	BLACK	477	75.470
	WHITE NORTH EUROPEAN	1020	96.500
USG	ASIAN	19	3.620
	BLACK	155	24.530
	WHITE NORTH EUROPEAN	37	3.500

Only for those Classified as Defendants or Offenders

4.1 Age, Ethnicity & Knives

The age of nominals carrying knives requires consideration - if the nominals start carrying knives at a young age this would be indicative of a different type of intervention than for their older contemporaries. The graphic and tables below show the age distribution, though it is worth remembering that the distribution is truncated at 25 and so the upper measures will be influenced by this. One can see that USG based knife crime is predominantly a male issue. The median & mean ages of those involved with knife crimes is marginally lower when gangs are involved. The first quartile for males involved in gang based crime is lower than non-gang based, suggesting the possibility of an initiation or use of young boys for violence though this cannot be confirmed. In the table below, the groups with means less than the grand mean for each gender or the extreme quantiles are highlighted below in red as these are seen as more at risk. Those coloured blue have the relevant statistic higher than the summary statistic.

Distributions of Knife Crime Ages

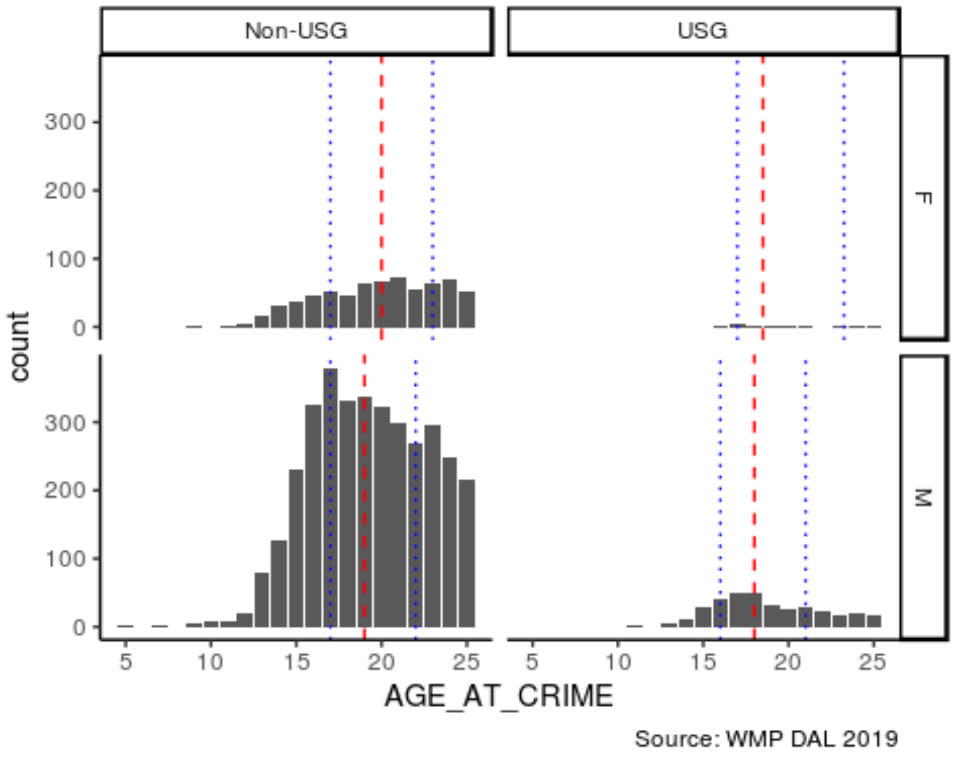


Table Showing the Age Distribution of Involvement in Knife Crime by Ethnicity and Gang Involvement

	Ethnicity	SEX	Mean	Q_05	Q_25	Q_50	Q_75	Q_95
Non-USG	ASIAN	M	19.400	14.000	17.000	19.000	22.000	25.000
		F	19.620	14.000	18.000	20.000	22.000	25.000
	BLACK	M	19.140	14.000	17.000	19.000	22.000	25.000
		F	19.740	14.000	17.000	20.000	22.000	25.000
	OTHER	M	19.000	14.000	16.000	19.000	22.000	24.600
		F	19.640	14.000	17.750	19.500	23.000	25.000
	WHITE NORTH EUROPEAN	F	19.880	14.000	17.000	20.000	23.000	25.000
		M	19.340	14.000	17.000	19.000	22.000	25.000
USG	ASIAN	M	20.320	15.650	18.000	20.500	23.750	25.000
	BLACK	M	18.600	14.000	16.000	18.000	21.000	24.000
		F	20.360	16.500	17.500	19.000	24.000	25.000
	OTHER	M	17.000	14.000	15.000	17.000	18.000	22.000
		F	18.000	18.000	18.000	18.000	18.000	18.000
	WHITE NORTH EUROPEAN	M	19.630	15.250	17.000	19.500	22.000	24.750
		F	19.250	17.000	17.000	18.500	20.750	22.550
Overall		**F**	**19.81**	**14**	**17**	**20**	**23**	**25**
Overall		**M**	**19.25**	**14**	**17**	**19**	**22**	**25**

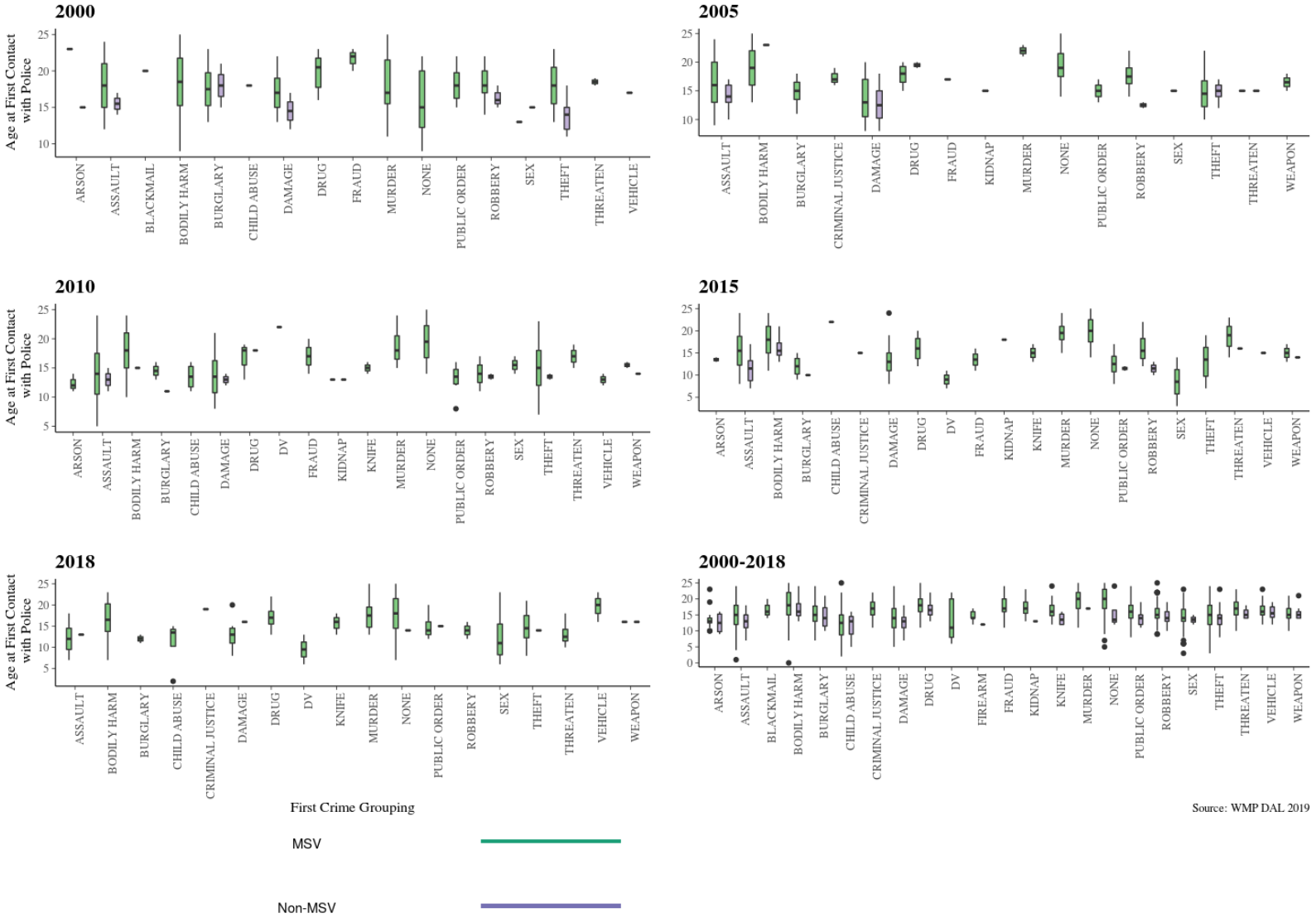
Given the possibility of early recruitment into USGs as a driving impetus of knife violence it is necessary to consider the age of the first crime. Taking the difference between the age at first contact (when the nominals were first involved in a crime) and the age at the first (MSV) knife crime, the mean of this difference will give us the average time between the first and the knife crime, thus giving us an understanding of the history of the nominals.

Table Showing the Mean Time between First Crime and Knife Crime for USGs and Non-USG involved Crimes (Years)

	Sex	Ethnicity	Average Time Betweeen Starting & Knife crime	Average Age of Knife Crime	Average Age of First Contact	Number of Cases
Non-USG	F	ASIAN	1.453	19.623	18.170	53
	M	ASIAN	2.459	19.400	16.984	814
	F	BLACK	2.687	19.740	17.060	150
	M	BLACK	3.344	19.135	15.826	718
	F	OTHER	2.570	19.787	17.254	756
	M	OTHER	3.042	19.245	16.237	3917
	F	WHITE NORTH EUROPEAN	3.329	19.880	16.613	401
	M	WHITE NORTH EUROPEAN	4.264	19.343	15.108	1567
USG	M	ASIAN	4.529	20.324	15.794	34
	F	BLACK	4.364	20.364	16.000	11
	M	BLACK	4.852	18.600	13.796	250
	F	OTHER	3.353	19.824	16.471	17
	M	OTHER	4.858	18.704	13.879	372
	F	WHITE NORTH EUROPEAN	1.250	19.250	18.000	4
	M	WHITE NORTH EUROPEAN	6.087	19.630	13.543	46

This table is interesting for a number of reasons. Firstly the age that nominals are first involved in crime is lower for those involved in USGs. Males on average have about 3.4 years between the first crime and the knife crime considered here, females have an average of less than 2.77. Though Black males have a lower difference in times between first and current crime, they are still involved in their first crimes at the age of 13 which is similar to White North Europeans (the average age for male USG nominals involved in knife crime was 13.96, for non-USG males it is 16). This suggests that those involved in USGs tend to have a longer history of involvement in crime and tend to be somewhat younger than the norm for involvement in knife crime.

Time Development of the Spread of Ages of First Contact by Offence Type

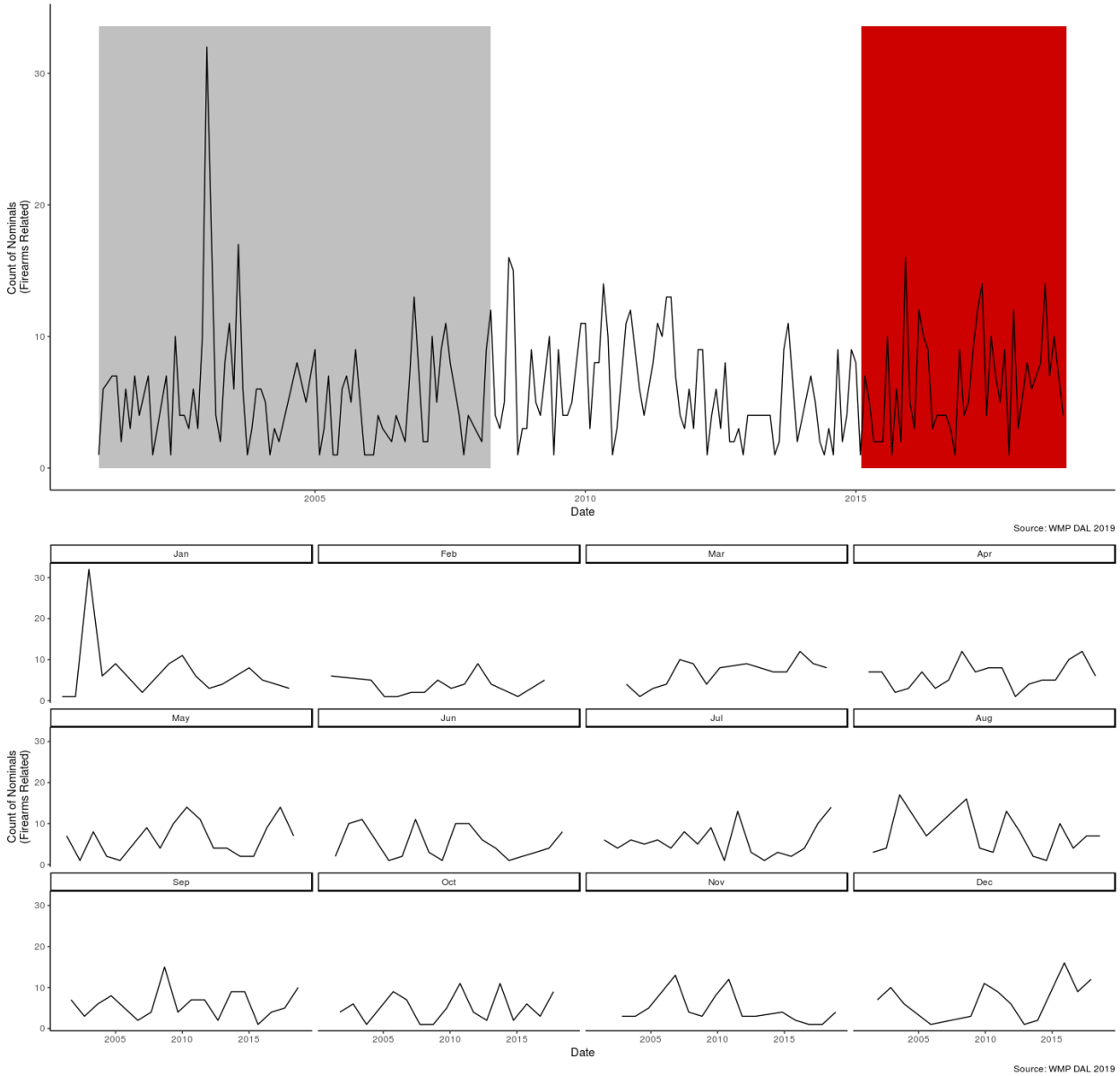


There is a considerable variation in the age of the first offence and the type of offence for those involved in knife crime as is the case for MSVs as a whole. There have been periods where assault, robbery, damage and theft have been associated with younger nominals involved in USGs. This however is not a constant finding. There is evidence presented in the graphs that those associated with USGs mostly have a lower age of first contact and it is more constant, with lower variances than those not involved in the gangs. This is not a surprise given the disparate nature of the data.

5 Firearms

Firearms crimes are far less common than other offences considered here; on average there are only 6 nominals involved in a firearms related crime on average (median 5) per month, compared to 23.58 (average) and 23 (median) in knife crime. The numbers associated with firearms are therefore far smaller and prone to lurches that appear substantial where they are only such as they start from a low base (increasing from 1 to 3 appears as a large growth rate). There are about 1000 observations in this dataset for firearms based crimes.

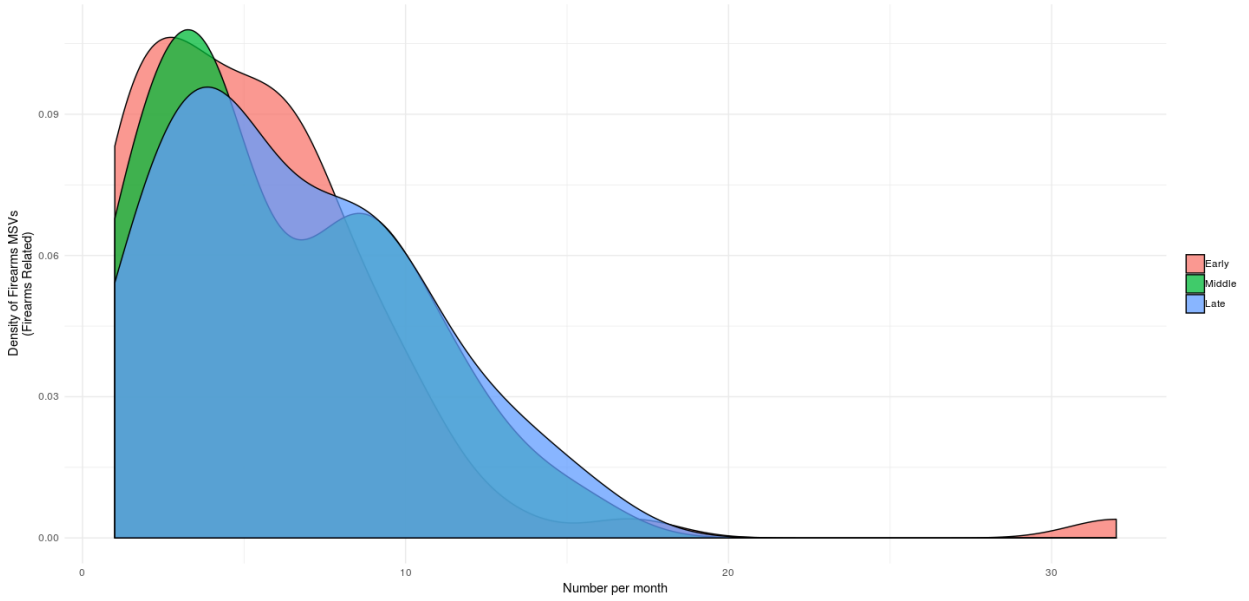
Firearms Crimes over Time



The count of nominals involved in firearms for the first time saw a spike in January 2003.

This is clearly an outlier insofar as there were a number of raids around the country such as those in Finsbury Park and the shootings of Charlene Ellis and Latisha Shakespeare on New Year's Day 2003.

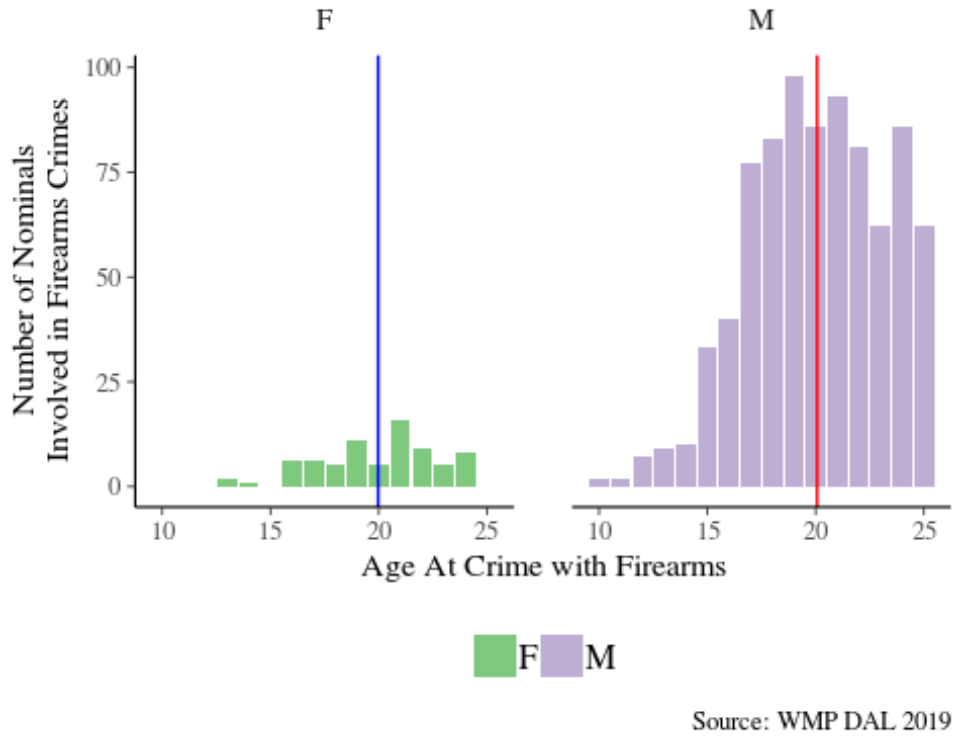
Distribution of Growth Rate in Firearms Crimes per Month



We can see that there is a single outlier in the firearms data, where 30+ nominals were involved. The relatively low numbers of these nominals and offences make the firearms data more traditional (in a statistical sense) - though somewhat more dispersed; the variance is relatively high, but again this is driven by the outliers. The three regimes also appear to tell a story of relatively static firearms use and participation, though there appears to be a slight increase in the growth of nominals participating in firearms crimes (year on year growth mean of 0.06 and median of 0.0 for the whole period and 0.11 (0.05), -0.05 (0.00) and 0.18(0.22)) with a more concentrated increase (i.e. smaller variation in the growth rates) in the later periods, which suggests a more clustered growth in the use of firearms.

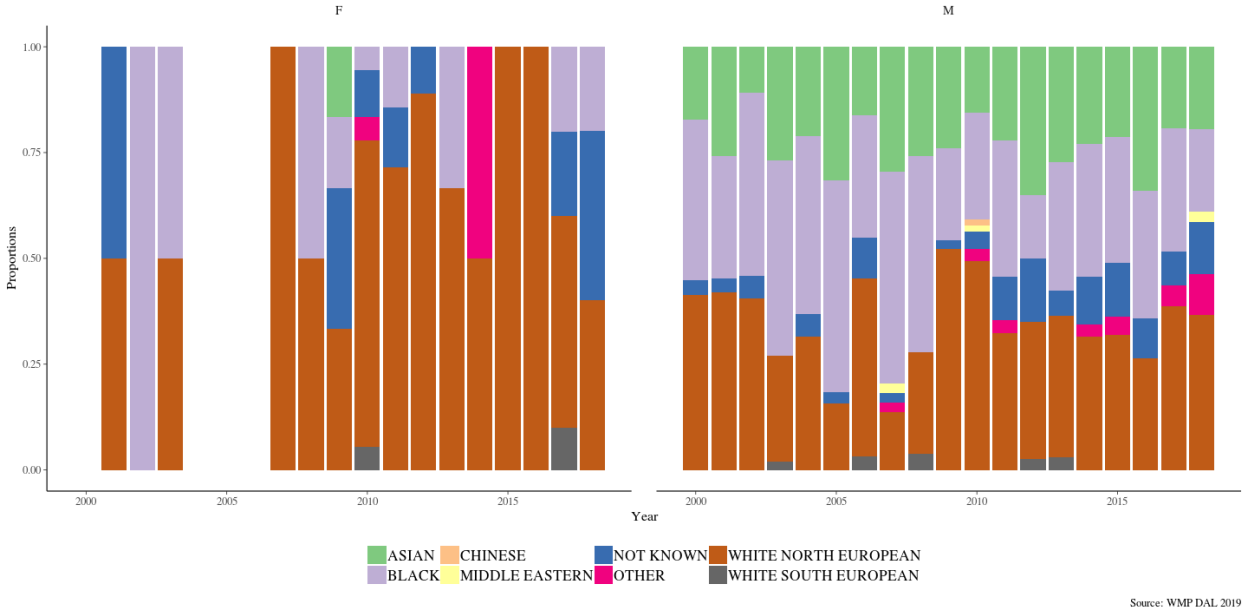
As with knives, the ages of the firearm based MSVs is informative. There are far fewer women involved in firearms offences. The deviation in the ages (measured against the maximum age in the sample rather than the mean) is 37.4698337 for men and 37.1386922 for women. This is due to the relatively small number of women involved in these crimes and their concentration between 19 and 21.

Distribution of Firearms Crimes; Age and Gender



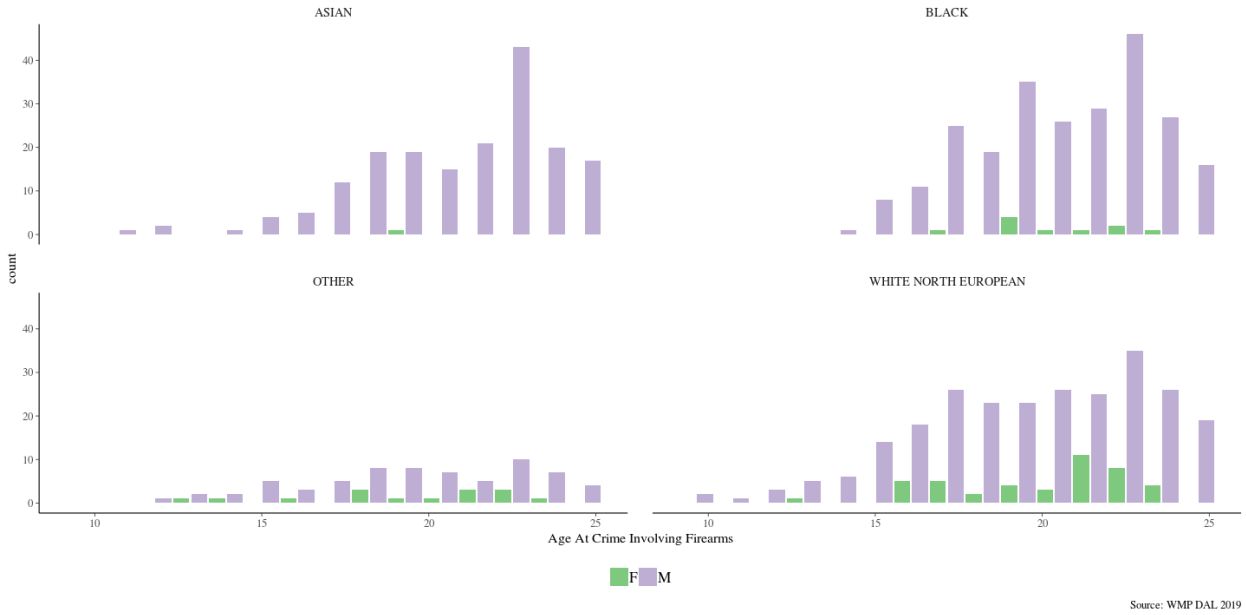
Source: WMP DAL 2019

Firearms Crimes – Proportion by Gender and Ethnicity



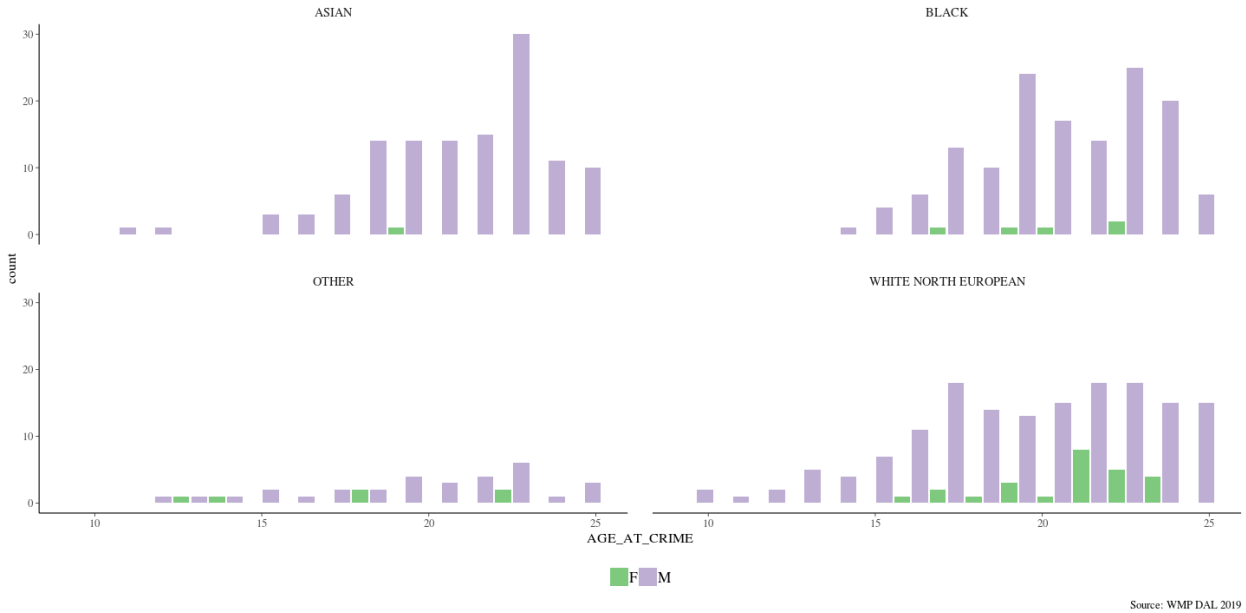
Source: WMP DAL 2019

Firearms Crimes – Age Distributions by Ethnicity and Gender



From the graphs above we can see that the range of ages for first firearm use is relatively broad with Asian, Black and to a lesser extent North European males varying primarily from 16 to 23, with North Europeans having a secondary peak around 17-18. Female firearm use is less frequent but there are peaks more explicitly noticeable in the data (there are very few Asian women involved as nominals in firearms offences). When considering those charged, rather than involved there are a number of properties clear from the data. We can see a more pronounced bi-modality in black men.

Firearms Crimes – Age Distributions by Ethnicity and Gender (Charged)



5.1 Ages between First Crime and First Firearms Crime

There is little correlation directly between the age at which a nominal comes in contact with the force and the age at which they first use a firearm (0.39). The mean time is about 4 years and 2 months. For males it is a little higher at 4 years and 4 months and for females 2 years 8 months. We can see that (ignoring the single Asian Female) that nominals are first known to the force in their mid-teens with the firearm offence in the early twenties.

Table Showing the Mean Time between First Crime and Firearms Crimes

Sex	Ethnicity	Average Time Betweeen Starting & Firearms crime	Average Age of Firearms Crime	Average Age of First Contact	Number of Cases
F	ASIAN	6.000	19.000	13.000	1
M		4.108	20.759	16.651	195
F	BLACK	4.600	20.300	15.700	10
M		4.797	20.304	15.507	276
F	OTHER	0.267	19.400	19.133	15
M		1.747	19.671	17.924	79
F	WHITE NORTH EUROPEAN	2.938	20.104	17.167	48
M		4.683	19.473	14.790	281

The age at which the nominals are known first to the force are not very different for firearms rather than knives, however there is a little more time before a firearms offence is committed. This is expected given the difficulty in obtaining firearms relative to a knife.

Table Showing Average Ages between First Contact and First Offence of Firearms (Years)

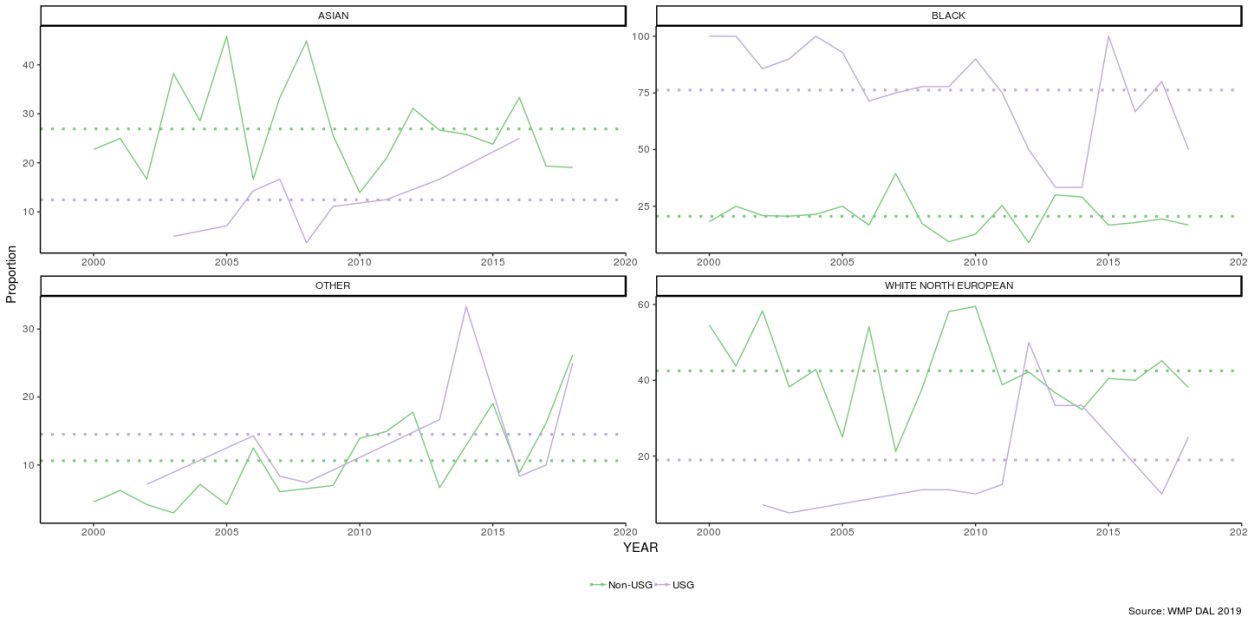
Type	Sex	Difference in Time Between First Contact & Offence	Average Age at Offence	Average Age at First Contact
Knife	M	3.428	19.231	15.838
	F	2.779	19.806	17.066
Firearms	M	4.307	20.070	15.763
	F	2.662	19.973	17.311

5.2 Gangs

Knives have particularly caught the headlines with gangs though of course firearms are an important adjunct to USG crimes. Of the total number of firearms offences in this data set 20.22% are associated with USGs. This data becomes more distinctive when split by ethnicity. Of the 183 cases of offences with firearms, 144 involved black nominals. This constitutes 78.69% of all USG firearms offences, for comparison in non-USG related

firearms crimes 19.67% involved black youngsters. This is lower than both White Northern Europeans and Asians (43.35, 25.48% respectively). We can see that the USG related firearms crimes are consistently above those of the non-USG crimes for Black nominals with some of the *others* also breaching in a similar manner.

Firearms Crimes Over Time – Non-USG v USG by Ethnicity

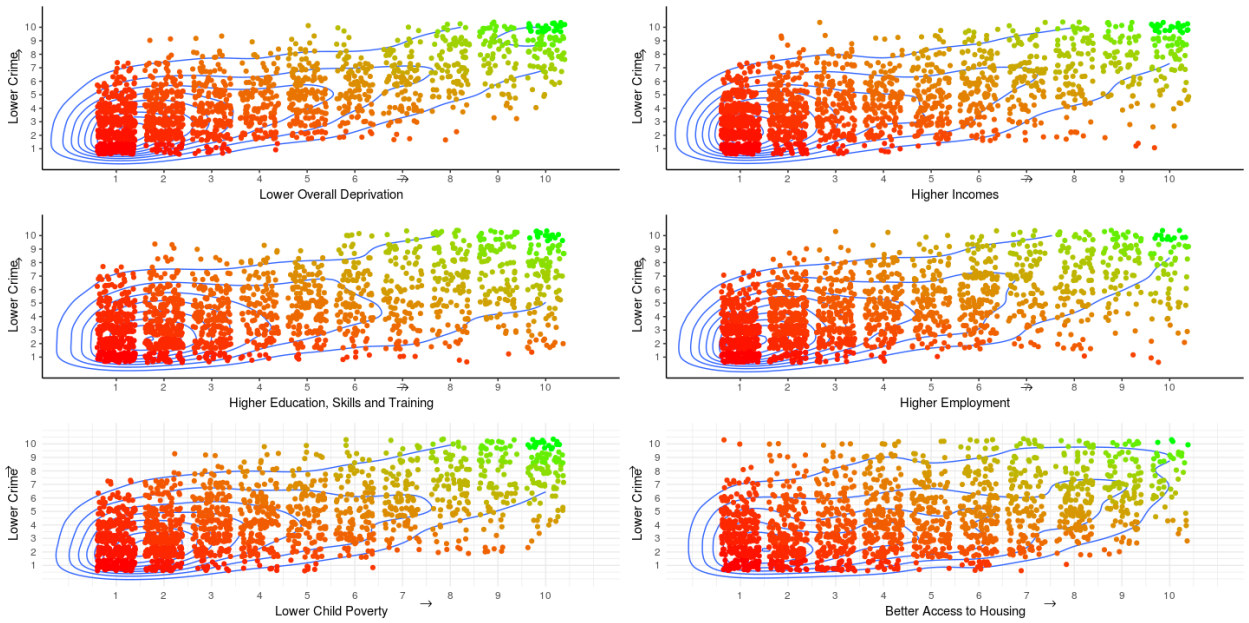


5.3 Socio-Economic Data

Following the thoughts of not only the individual's circumstances but also the environment being important in the consideration of the risks associated with moving towards violent crime, we considered various Deprivation Indices. These are available from a number of sources such as https://www.gov.uk/government/collections/english-indices-of-deprivation. These are available at LSOA level which have been mapped to postcodes. The data is static and based upon 2013 Districts, though not as up-to-date as one would hope for (this data is not published every year, rather a 3 or 5 year cycle is implemented), it should be a close enough proxy for the relative socio-economic position of areas. The decile is used as it is the relative situation that is important not a particular level or ranking; the fact that an area is in the top 10% of areas for crime in England is more useful than it was 11th or 12th.

Examining the crime decile in relation to the income and education, skill and training, we can see that there is a generally positive relationship, which of course would be expected. The graphics below demonstrate the relationship between a number of metrics and the crime index.

Comparison of IMD Subject Area Indices and Crime Index



NOTE: The Axes run from high at the origin to low; for example 10 is less deprived than 1 and 10 has fewer crimes than 1 (because these are deciles). Colours are just for visual purposes to aid delineation.

There is some variation however which are potentially useful features to consider, especially the impact of housing, child poverty (IDACI) and education.

5.4 Geo-Spatial Aspects

LSOAs are normalised to have a population of approximately 1500 people and as such there has been a small increase in the number of LSOAs since 2011; thus comparisons with the 2011 data should be performed with care. The use of deciles will generally mitigate this for all but the boundary areas. From a statistical perspective using all the indices is unnecessary in the explanatory model as the overall index is a weighted sum of the constituents. A number of individual elements are included with a view that at least one, if not more will be removed from the model. The socio economic elements are given most importance in the data description as the Education and Skills gap and especially the Young People Sub domain is important for the impact of the local area on young people's aspirations and opportunities. This is further supplemented by the use of the Employment Deprivation data which gives a measure of the involuntary exclusion from the labour market. This metric includes unemployment as well as other health and social factors. The focus on the younger members of society also requires the selection of income deprivation affecting children. Talking to members of the Service, the role of USGs was highlighted and a part of that was the income levels of the family. It is therefore important to see if the level of child poverty is important, even though this (and other similar explanatory variables) are not directly solvable by WMP, rather requiring a more broad economic and social policy than law enforcement is responsible for.

The data was amalgamated to group the deciles into larger groupings, although the lowest level (representing the most deprived areas) was still kept separate as this was considered as more of a threshold. The data was also used in interaction form in the modelling as it is not only the level of the measure but how it interacts with each of the other variables of interest.

5.5 Data Summary

The data suggests a relationship between a number of demographic and social aspects in the incidence of using a knife, firearm or partaking in MSV. This analysis can inform the inclusion of the variables into an explanatory model that links these and their interactions together in order to give the probability of a nominal being involved in or a defendant in one of the classes of crimes considered. The data's size necessitated a brief overview and description of some of the more interesting relationships. The full variable lists are provided in the Data Dictionary. The data presented suggests that there are a number of potential interactions amongst the variables; the behaviour of Asian females is very different from that of Asian males, relative to the differences between North European males and females. There are additional factors at work that are particular to Asian females compared to other females. These are included in the modelling approaches discussed below. In effect these interactions give us a separate model per group of interaction; the net effect of the nominal being Asian and female will be the sum of the effect of being female, Asian and the interaction of these.

The approaches considered are highlighted in the next section with more technical aspects in the Appendices.

6 Models Considered

The models are looking to extract information from the relevant data in order to look to explain what increases or decreases the chance of a nominal being involved in MSV, firearm or knife crime. In order to do this, we have extracted data that includes nominals who are involved in MSVs before the age of 26 and a sample of those of similar age who have not been involved in MSV by this time. There are a number of approaches to determining the driving factors of the behaviour. The problem becomes one of selecting the most useful variables and then estimating the strength and direction of the relationships with the outcome. The variable selection phase used the three approaches below; these were then used to inform a final logistic regression for the estimation of the relationships between the important variables and the outcome.

- Logistic-type regression (Cox (1958))
- Random Forests (Breiman (2001))
- GBM (Gradient Boosted Models) (Friedman (2001) and Friedman (2002))

The advantage of using a number of approaches is that they each give different insights and can be confirmatory of each other. The variable selection stage used the H2O (https://www.h2o.ai/) variants (@LeDell et al. (2019)) of the relevant algorithms as this was found to be the most efficient method of ascertaining the information necessary. The variables identified were then used in a second logistic Elastic Net regression to estimate the final relationships. A more technical description is given in the Appendices.

Models are presented for those involved in the crime as Defendant/ offender, those suspected and those believed to be responsible. The rationale for this is that the remit of this work was to identify why some youngsters turn to MSV/ knife/ firearm crimes. This can not only be concerned with those who are caught and charged. Whilst this means that the models might not be as 'accurate' it does mean that they look to explain more than the characteristics of those who get caught and charged (i.e. there is more data and information with which to identify the factors which increase or decrease the probability of committing these types of crime). This gives WMP an opportunity to form policies to help those at risk of taking part in such crimes not just those who are likely to be apprehended and charged when it is arguably too late.

6.1 Logistic Modelling

The more standard approach uses a binary dependent variable; whether the nominal was involved in a crime of interest or not (1 or 0). This allows us to consider the probability of the event occurring.

As in many data sets of this kind, we are not able *prima facie* to select the '*correct'* variables to include in the models. This therefore requires a form of variable selection. There are a number of methods of performing this; the most common are variants of the LASSO. A number of different variants were used

- LASSO - this uses a constraint on the coefficients, attempting to shrink them down towards 0 (Tibshirani (1996)).
- Elastic Net - a modification of the LASSO to take account of the issues that LASSOs have with correlated variables (Zou and Hastie (2005)).
- Relaxed LASSO - a second elastic net model or LASSO model which reduces the (statistical) biases introduced into the estimation by the LASSO by using the LASSO for purely selection with a second regression being used for the final fit (Meinshausen (2007)).

The Relaxed LASSO/ Elastic Net was used as this is a generalisation of the subordinate parts. This was modified to allow an elastic net (op. cit.) in both stages. In each step 10 fold cross-validation was used to assess the model fit to ensure that over-fitting biases (in the statistical sense of the term) were minimised.

There is considerable debate about the impact of very rare events on models such as these. After all, if there are very few incidents of say firearms crime, then there might be a potential (statistical) bias in the estimates. Owen (2007) suggests that though the constant element will lead to a base probability of 0 (that would be a good estimate if you know that only 1 out of a million people exhibit a particular behaviour), the coefficient estimates for the other explanatory features can be useful. Indeed in the case in front of us, there is a parallel with Owen's work on fraud detection; the number of positive outcomes is limited and the number of non-positive outcomes grows.

6.2 Random Forests and Gradient Boosted Machines

In order to assist in the variable selection problem, a number of non-regression, tree based models were considered. Random Forests (Breiman (2001)) and Gradient Boosted Machines (Friedman (2001)) (henceforth GBM) are both tree based ensemble methods. They are particularly good at dealing with categorical variables and for determining variable importance. Random Forests work by training many trees with factors selected at random. These trees are grown in parallel and the results are averaged. An analogue would be a form of the wisdom of crowds, where each member of the crowd has potentially different (but independently acquired) information. Individual decisions are often poor, but in combination the decisions improve so long as there is no relationship between the trees. Unfortunately the output of a random forest model is complex to interpret. Given the analysis here, the random forest (and the GBM) are used for measuring explanatory variable importance only. This is derived by the improvements generated by the individual factors as measured by the (squared) errors normalised to allow for comparisons across factors.

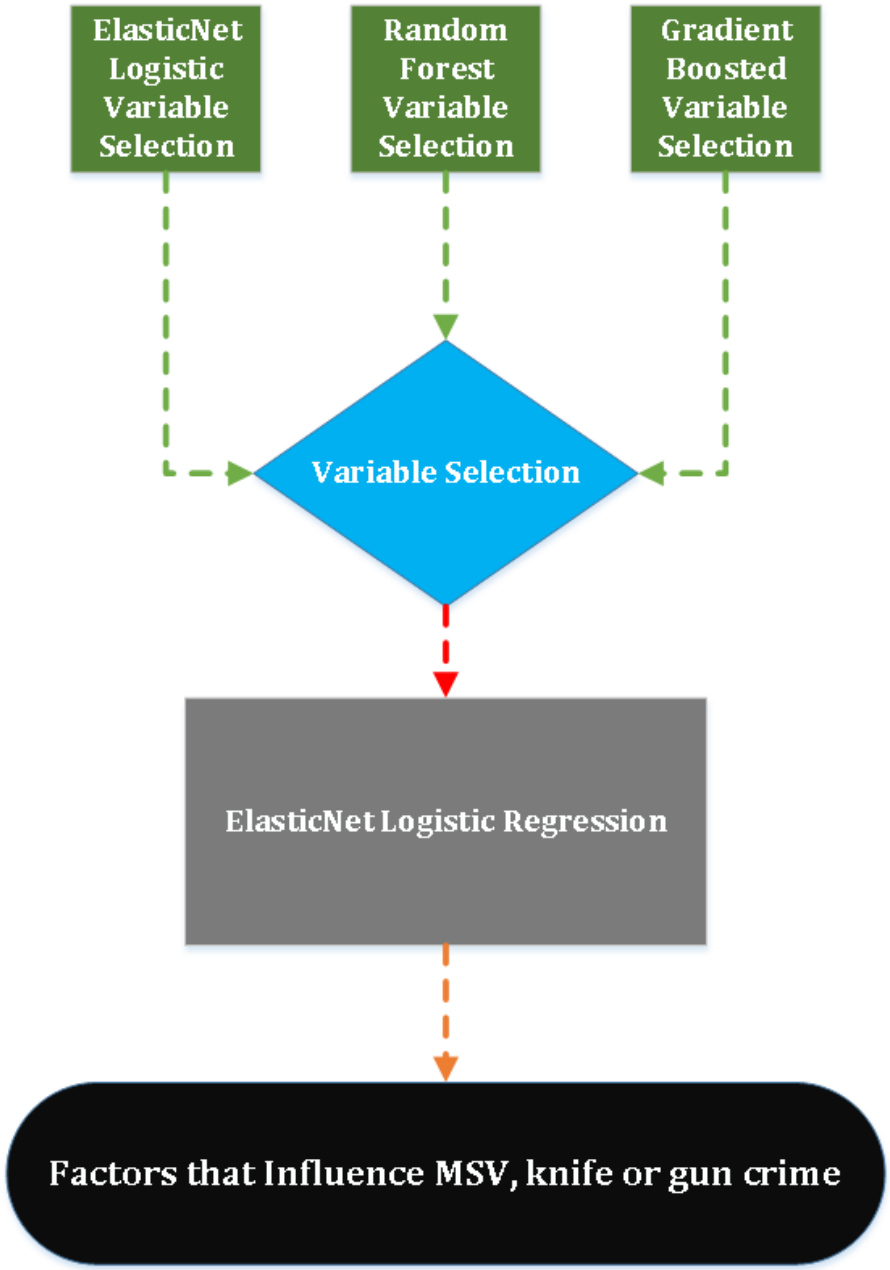
GBMs work sequentially, with later trees improving on the errors of earlier trees. This improvement uses *weak* learning, improving at each step by the use of the error. The trees incrementally use the errors from the previous generations to improve at each step. This approach requires some form of stopping criterion, such as those based on cross-validation otherwise there is the potential for over-fitting, especially where there are outliers in the

data. Rather than the trees being independent of each other as in the case of the random forest, GBM trees are grown sequentially.

It should be noted that these approaches were used only as an aid in variable selection, rather than as a model of explanation *per se*.

The process can be summarised via the flow chart below.

Estimation Process



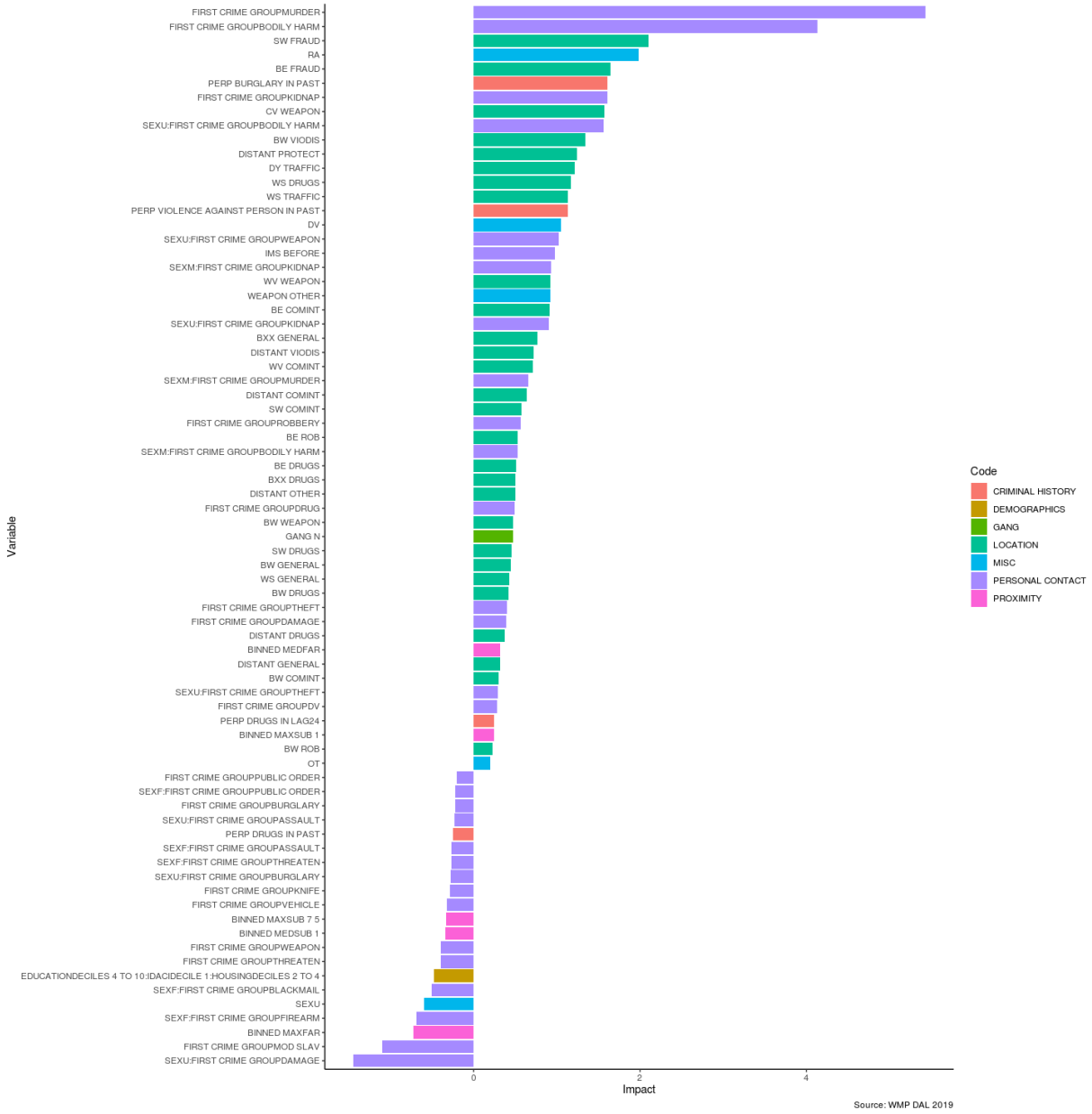
7 Results

The various approaches informed selection of the data for the modelling. The Random Forest and Boosted models were used for variable importance.

7.1 Explanatory Factors for MSVs

The drivers of MSV incidence were specified using the second step of an elastic net regression and a number of interactions as discussed above.

Coefficient Estimates for MSV



Amongst the variables that increase the probability of committing MSV are a number of violence related actions including bodily harm and the use of weapons. More unusual previous crimes include perpetrating burglary, robbery and involvement with drugs. However, if the first crime was burglary, this tends to reduce the probability of committing MSV. Socio-economic factors are also important in that the less deprived an area in terms of housing and education, the less likely they are to enter into MSV, despite IDAC deprivation being high.

There is also evidence for the potential for county lines effects in that intelligence received from a distance away correlates with a higher probability of committing MSV.

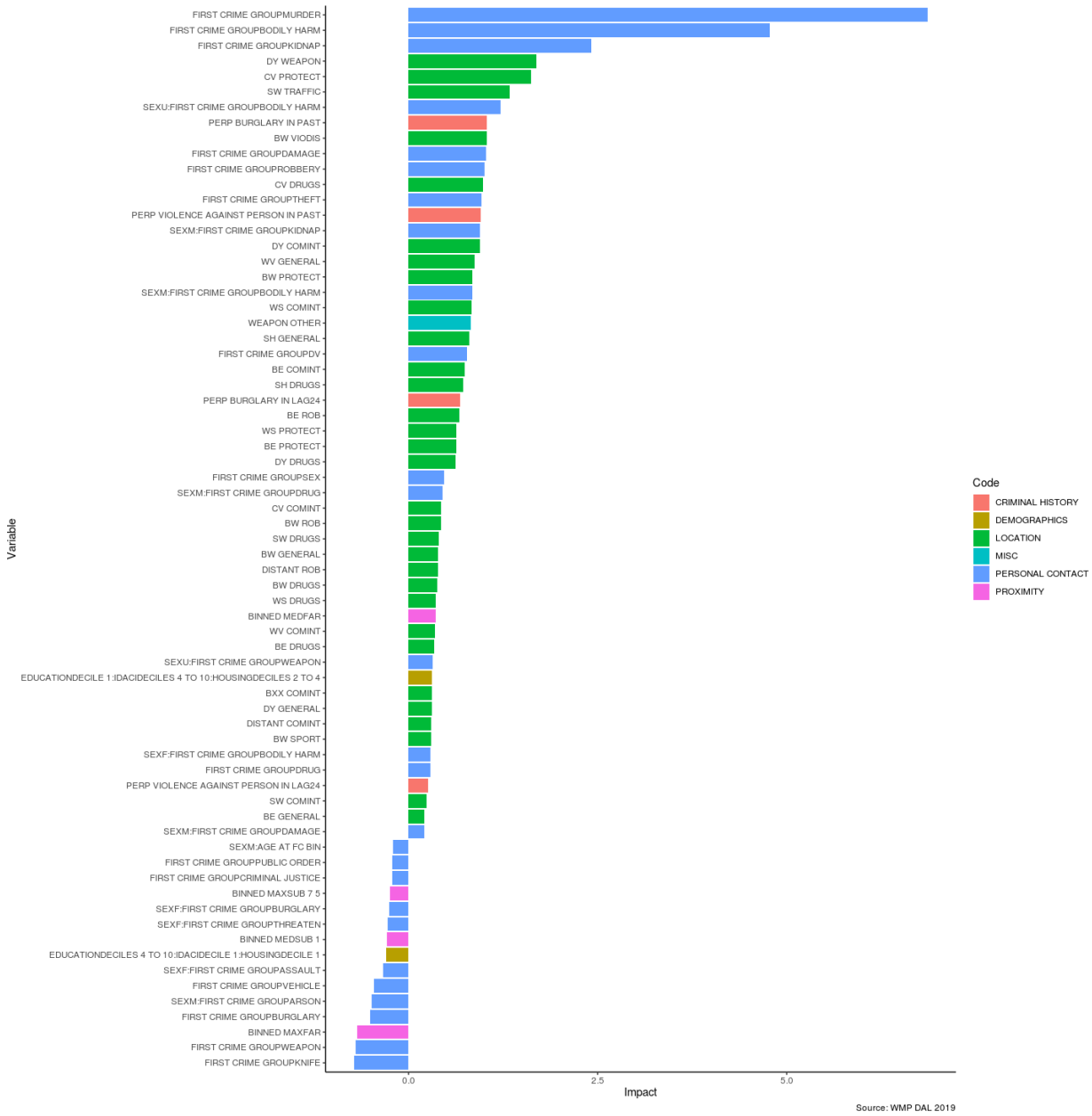
Topic	Freq
CIVIC ENVIRONMENT	0
CRIMINAL HISTORY	4
DEMOGRAPHICS	1
GANG	1
LOCATION	28
MISC	5
NETWORK	0
PERSONAL CONTACT	31
PROXIMITY	5
TIME	0
VICTIM HISTORY	0

Table Showing the Coefficient Groupings for MSV (All Involved)

The drivers of MSVs as a whole are broad and reflect the breadth of the crimes associated with the nominals. The history of the nominals has some impact as does the area and intelligence logs associated with the nominals. The factors that tend to increase the odds of MSV tend to be larger in size than the factors `trying` to reduce the odds on this happening.

Looking at only those that were defendants in MSV crimes shows a similar picture:

Coefficient Estimates for MSV



Nominals who have committed burglary in the last two years are more likely to commit MSV as are those who have committed drugs offences in various locations (most likely due to USG effects). We also see a negative correlation between the probability to commit MSV and age at first crime (for males) – the older the nominal at their first crime, the less likely they are to commit MSV later. There is also the unusual effect of burglary; if they have committed burglary in the past they are more likely to commit MSV, but if their first crime was burglary, they are less likely to go on to commit MSV.

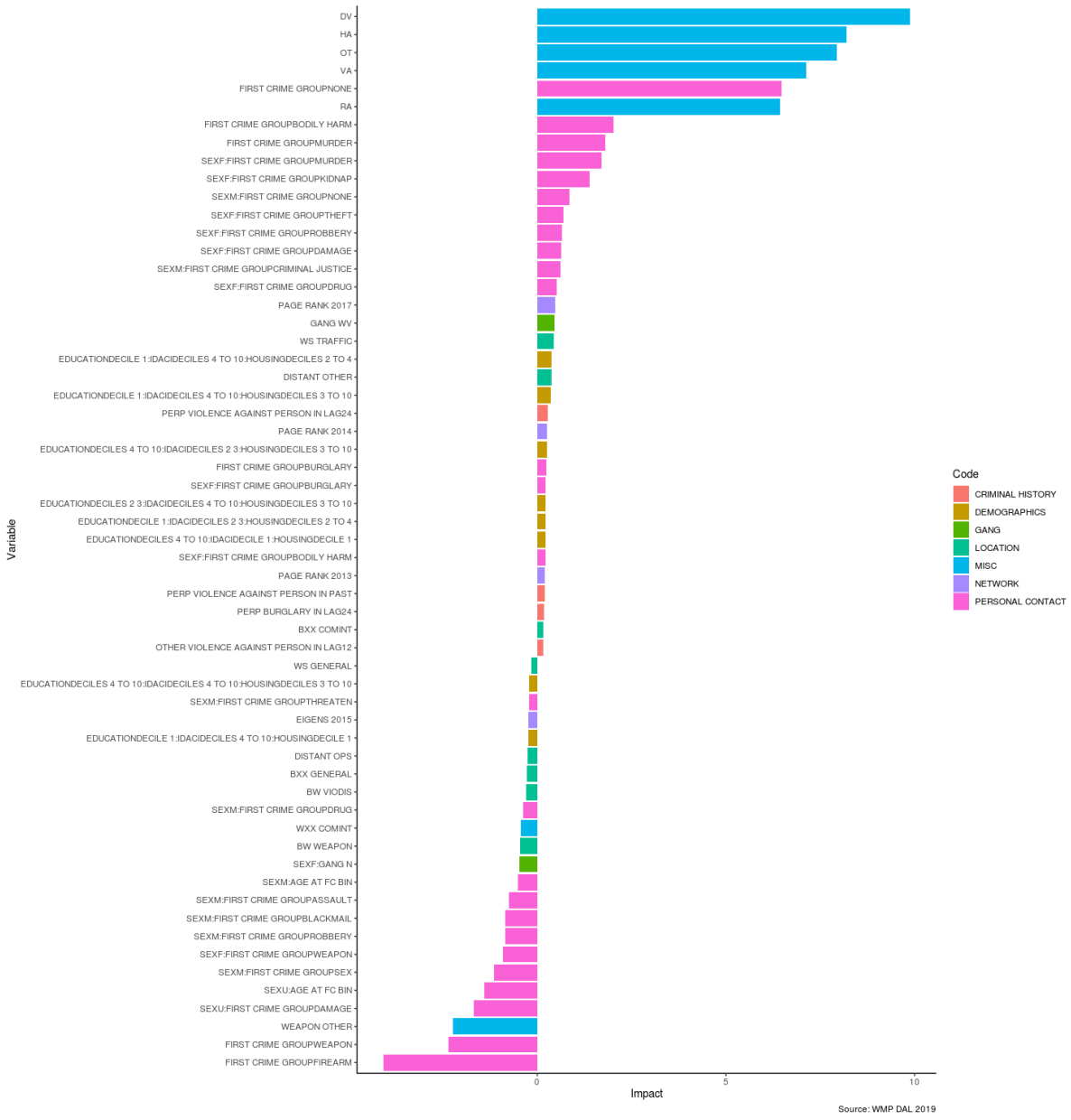
Topic	Freq
CIVIC ENVIRONMENT	0
CRIMINAL HISTORY	4
DEMOGRAPHICS	2
GANG	0
LOCATION	31
MISC	1
PERSONAL CONTACT	27
PROXIMITY	4
TIME	0
VICTIM HISTORY	0

Table Showing the Coefficient Groupings for MSV (Only Defendents)

7.2 Knives

As with the MSV crimes, the magnitudes of the coefficients in the knife models are larger for those greater than zero, suggesting that the drivers towards knifes have a greater impact than those pulling away from knife crime. In contrast with the MSVs in general, the lower education and training deciles in conjunction with child deprivation and poor housing leads to higher odds of being involved with knife crime. Housing in particular when it is in conjunction with other measures seems to be a useful explanatory variable.

Coefficient Estimates for Knife Crime (all)



We can see the factors that have a positive impact on the log-odds of involvement in a knife crime; a number of contemporaneous descriptors, i.e. what sort of crime the nominal is currently involved in but also first crimes of theft, robbery and criminal damage if the nominals are female. We also see network effects increasing the probability of committing knife crime in that a previous year's page rank has a positive coefficient.

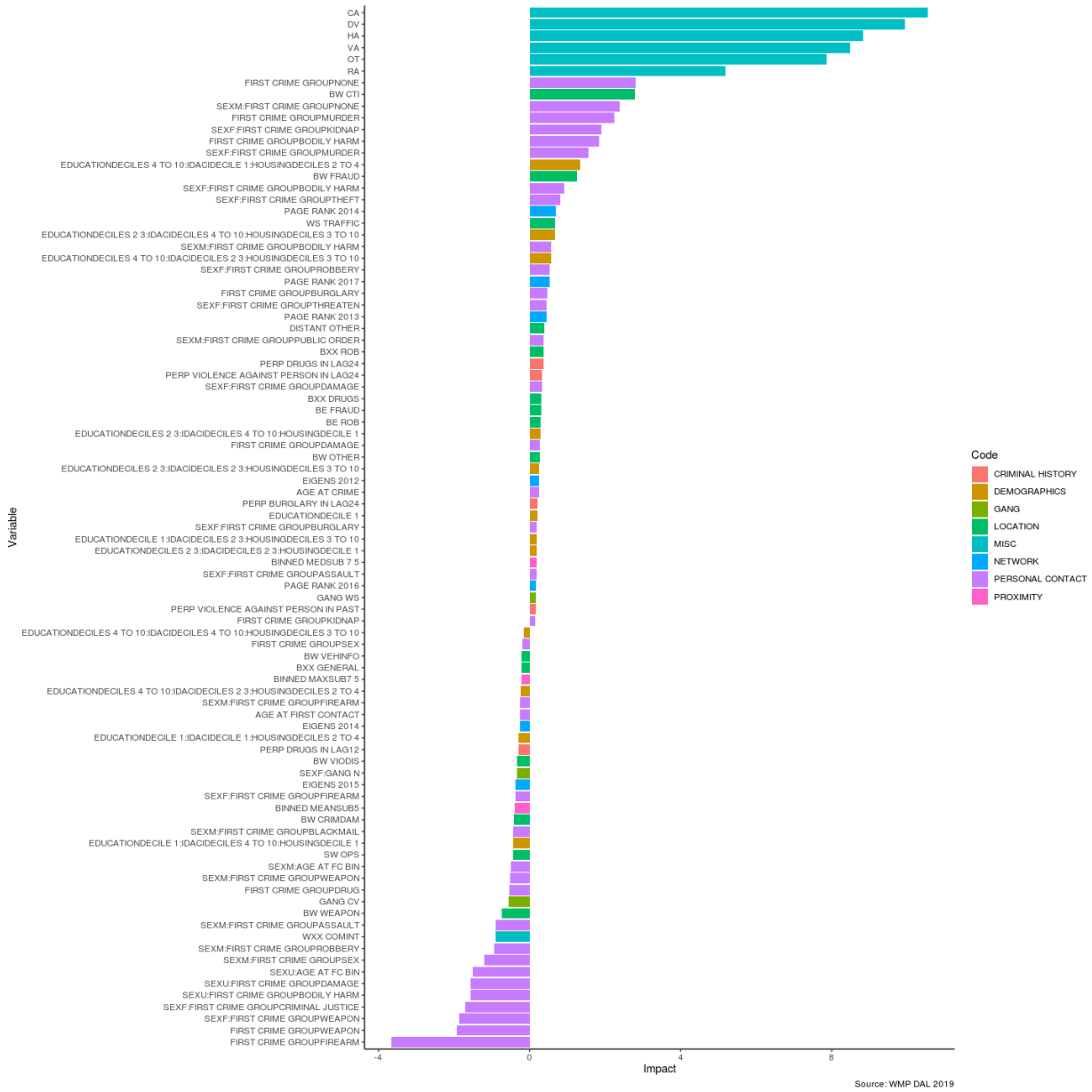
As with MSV (defendants only), age of males at their first crime correlates negatively with the probability of committing knife crime.

Topic	Freq
CIVIC ENVIRONMENT	0
CRIMINAL HISTORY	4
DEMOGRAPHICS	8
GANG	2
LOCATION	8
MISC	7
NETWORK	4
PERSONAL CONTACT	26
PROXIMITY	0
TIME	0
VICTIM HISTORY	0

Table Showing the Coefficient Groupings for MSV (All Involved)

The coefficient estimates for defendants only are on the next page.

Coefficient Estimates for Knife Crime (defendants only)



A similar picture as for all roles is found with contemporaneous crimes having the largest coefficient estimates. Higher levels of locational deprivation also contribute towards an increased probability of committing knife crime. Network effects are also prevalent and the older a nominal is at first contact with WMP, the less likely they are to commit knife crime.

Topic	Freq
CIVIC ENVIRONMENT	0
CRIMINAL HISTORY	5
DEMOGRAPHICS	12
GANG	3
LOCATION	15
MISC	7
NETWORK	7
PERSONAL CONTACT	37
PROXIMITY	3
TIME	0
VICTIM HISTORY	0

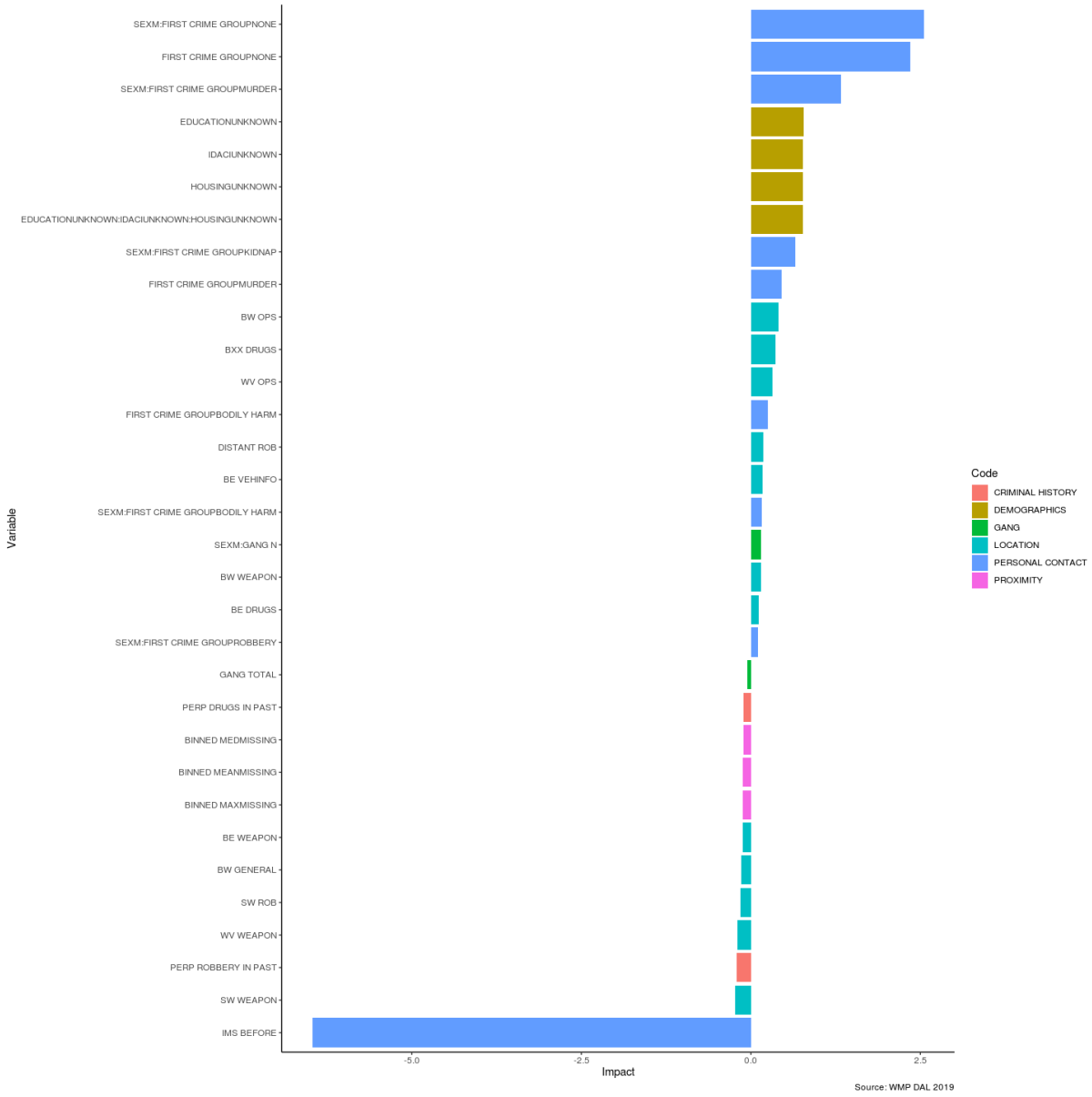
Table Showing the Coefficient Groupings for Knife Based MSV (Defendents Only)

7.3 Firearms

As with knives, the second modelling stage is used to identify the important explanatory factors. More socio-economic factors appear to come into play in firearms crimes than for either knife crime or MSV in terms of a higher probability of committing forearms offences.

Interestingly, being a perpetrator of robbery in the past reduces the probability of committing firearms offences.

Coefficient Estimates for Firearms Crimes (all)

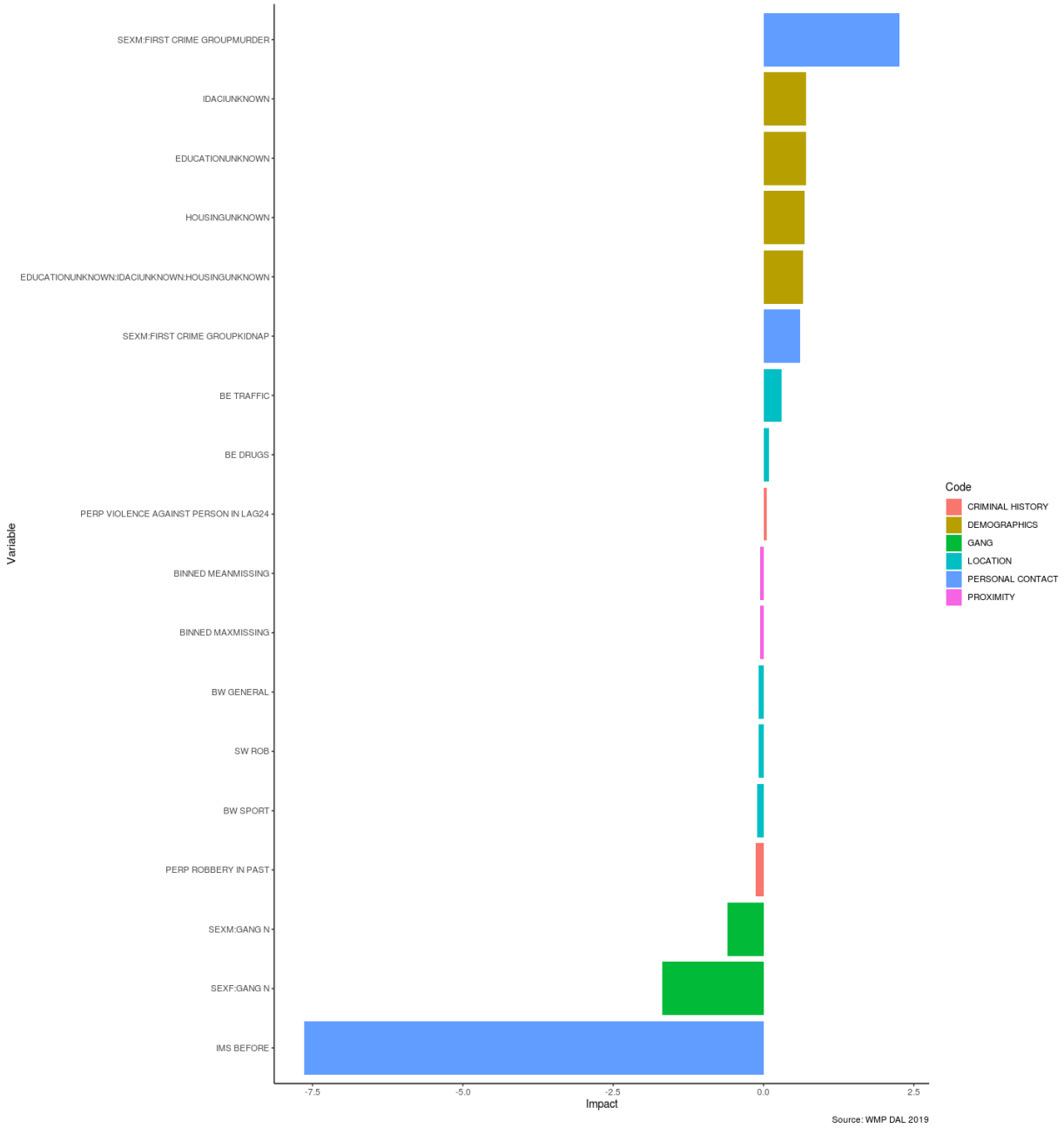


Topic	Freq
CRIMINAL HISTORY	2
DEMOGRAPHICS	4
GANG	2
LOCATION	12
MISC	0
NETWORK	0
PERSONAL CONTACT	9
PROXIMITY	3
TIME	0

Table Showing the Coefficient Groupings for Knife Based MSV (Defendents Only)

In the case of defendants only, a similar picture emerges:

Coefficient Estimates for Firearms Crimes (defendants only)



Topic	Freq
CRIMINAL HISTORY	2
DEMOGRAPHICS	4
GANG	2
LOCATION	5
MISC	0
PERSONAL CONTACT	3
PROXIMITY	2

Table Showing the Coefficient Groupings for Knife Based MSV (Defendents Only)

8 Conclusions

Seriously violent crime is a complex phenomenon with a multitude of factors explaining how younger offenders move into this type of crime. The sub-classifications of these crimes determine the relevant drivers. Common factors in all three classifications show that the age at which the nominal comes into contact with the force and whether there is a history of non-MSV violence against individuals over the last 24 months and intelligence that there is some involvement in the Drugs trade increase the probability of MSV / knife / firearms crime. In addition to these, various intelligence information based in Birmingham is also of importance. The offence flag of Domestic Violence is also a potential factor in moving a nominal towards committing MSV. This is one of the contemporaneous factors, though it should be noted that it would be one that is potentially noticeable before the crimes in question occur. If the nominal is known in the intelligence logs this tends to increase the odds of them being involved in an MSV incident. Nominals coming into contact with the Police at earlier ages are more likely to be involved in all three types of crime than a nominal who is known later. This would suggest a potential intervention for the youngest nominals especially males whose first crimes tend to involve non-MSV violence, harm or damage.

The intersection of knife crime factors and those of firearms also include the socio-economic factors at a more macro level but also a number of factors concerned with the connectedness with other known nominals and appearances in intelligence and also Gangs (especially in Birmingham West). Again there are signals based on being involved in violence against the person over at least 2 years again as might be expected males are most at risk. The level of connectedness as measured by the nominals' page rank in 2017 does demonstrate a positive impact on the odds of being involved in both types of crimes.

Though the county lines information from COMPACT is not important, there is evidence that if intelligence comes from outside the West Midlands (and beyond), then there is an increase in the chance that the nominal will be involved in MSV and knife crime. The housing variable demonstrates similar properties; that poor access to the housing stock will increase the odds of a nominal being involved in knife crime. As before, being involved in violence against people in the last 2 years leads to increased odds of being involved in all types of crime considered here. The greatest increases are those of being involved in knife crime, with the smallest increase being for MSV in general. This is because of the breadth of the MSV model where more factors come into play and therefore can reduce the influences of any particular factor.

9 Technical Details

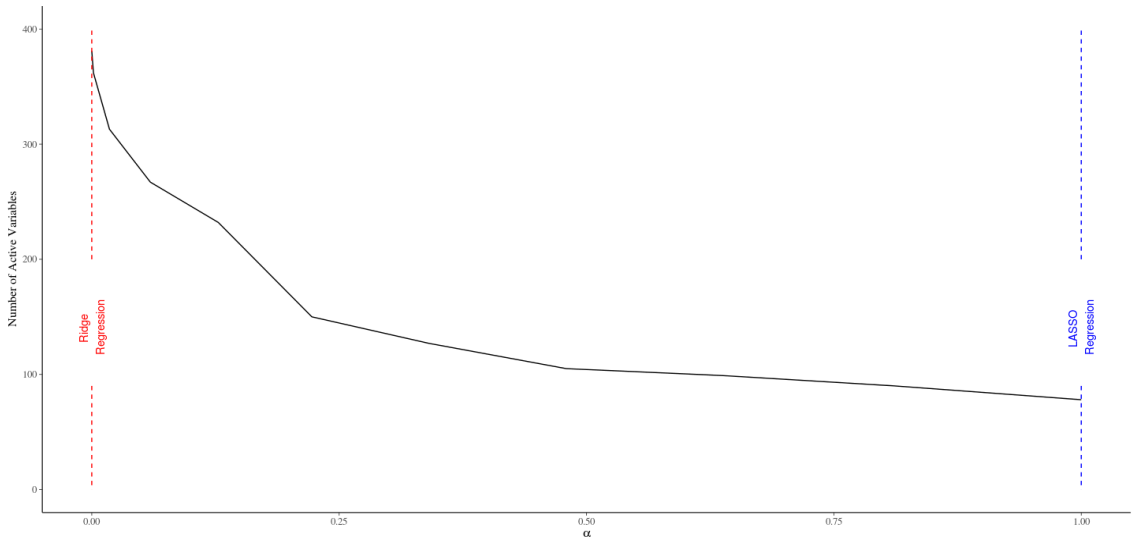
9.1 Variable Selection Step

As ever in the situation where we have a plethora of data, the selection of the variables is important. This section considers the selection of the data for the final steps of the explanatory model. These are discussed for each of the methods with the final model being generated after the selection. The aim here is to reduce the number of variables as 380+ is too large to interpret in a meaningful manner. This section looks at the variable selection criteria before moving on to discuss the final explanatory model for each of the cases considered. The selection of variables is discussed by technique as this allows a more direct comparison of the variables selected across the crime types.

9.2 Relaxed LASSO; MSV in general

The first step of this method is variable selection. The number of variables depends upon the regularisation strength; the firmer the constraint, the more variables are disposed of. Because the first step is set to reduce the number of variables, the Ridge approach was discounted. We can use the LASSO element to select the number of variables and consider the variables in this and other approaches outlined in coming sections.

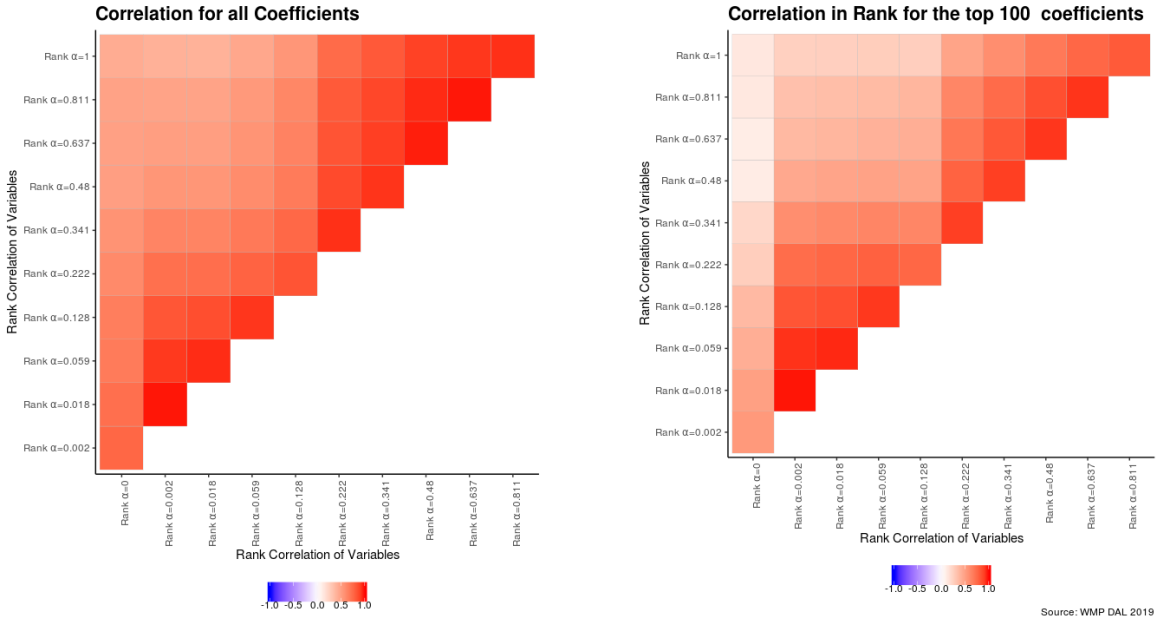
Number of Variables Selected By Choice of a for MSV crimes



The variables selected are generally the same with strong correlations between the ranks of their importances. There is clearly a stronger correlation between those variables ranked in the top 100 for models *close* to each other; but those further apart have less of a relationship. These are shown in the diagrams below, with all variables' ranks being strongly correlated (in parameter space). As the selection becomes more liberal the correlation of course becomes stronger. In the case of the first LASSO step there are at least 78 variables. Using a measure of the variable being at least once in the top 100 most

important variables, we have a list of 116 variables. These include some of the variables that SMEs would recognise as important, including gangs, knife prevalence as measured by growth and previous levels. The social factors are also important with IDACI, HOUSING and EDUCATION all appearing (though not all in a positive manner and small). The history of the nominal is also important with a history of Violence against the person, Drugs, Criminal Damage, Robbery and Theft all being important. The role of the near family (husbands, wives, fathers /mothers /sons/ daughters/ partners) appears to also have a positive effect.

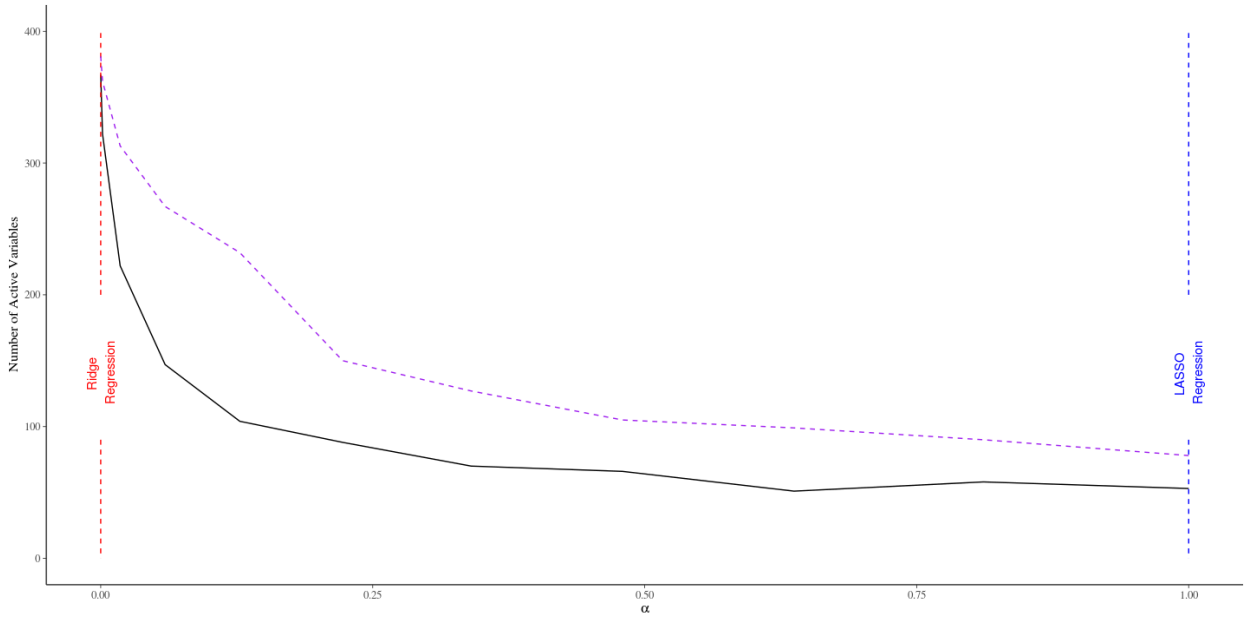
Rank Correlations between Variables Given Estimation Parameterisation for MSVs in General



9.3 Relaxed LASSO- Firearms

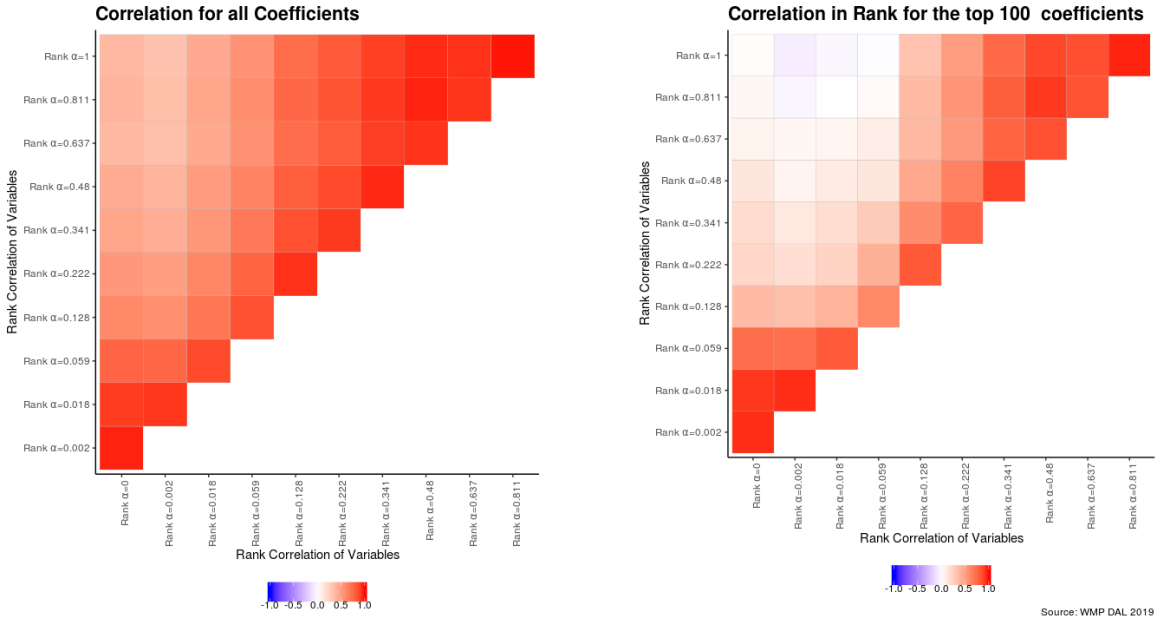
The firearms data is more restricted than the MSVs in general. There are far fewer positive observations and fewer variables are selected as can be seen below with the purple line representing the MSVs as a whole. As with the MSVs as a whole, gangs are an important factor though it is noticeable that the only *positive* effects are those gangs active in Walsall and non-WMP locations. The main USGs involved are The Burger Bar Crew, Johnsons Crew and The Raiders.

Number of Variables Selected By Choice of for Firearm crimes



The correlations between the ranks are also reduced (to all intents and purposes to 0) with the correlations of the α >0.48 with $= 0$.

Rank Correlations between Variables Given Estimation Parameterisation for Firearms Crime

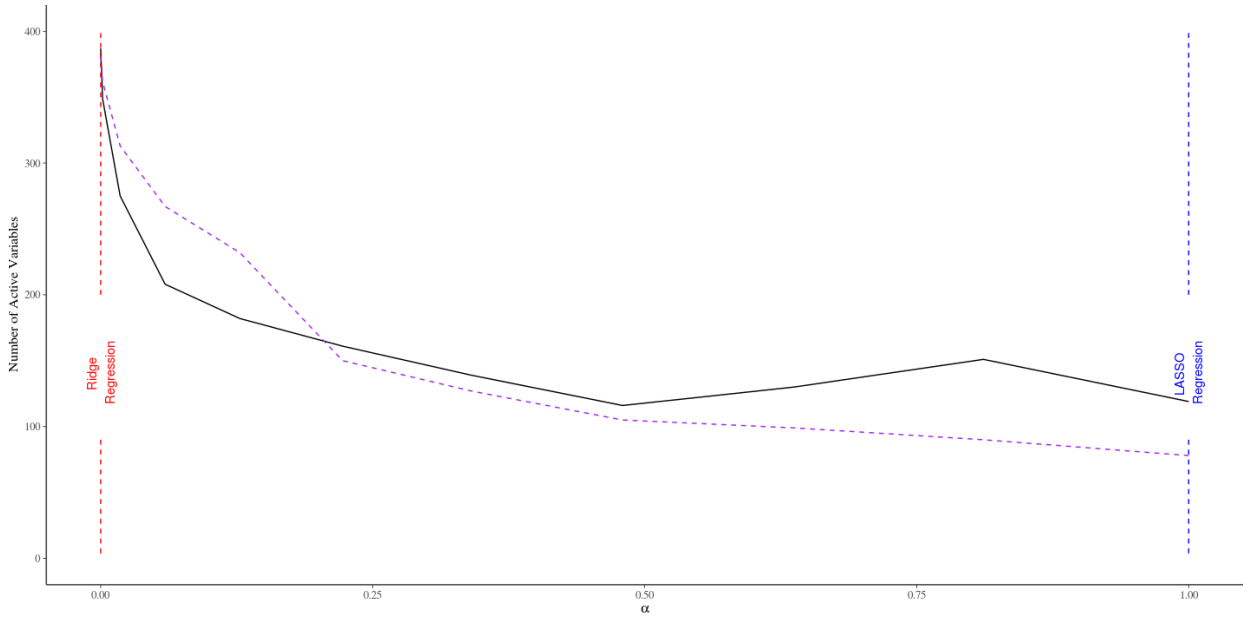


9.4 Relaxed LASSO- Knives

As before there is a reduction in the number of variables included in the explanatory model for knives. The decline mirrors that of the firearms model. Again there are approximately 100 variables that are in the top 75 at least once; using an appearance in the top 50 limits

this to 79 for both models of involvement and defendants. There are a number of locational variables across the region including a metric for distance from the nominals. As before, the deprivation indices are included. The SME's belief that at least some of the knife crime was caused by nominals holding knives as a side effect of knife crime in general was proxied by the growth variables in the data and the monthly (lagged) growth rate was found to be one of the important variables.

Number of Variables Selected by Choice of a for Knife crimes

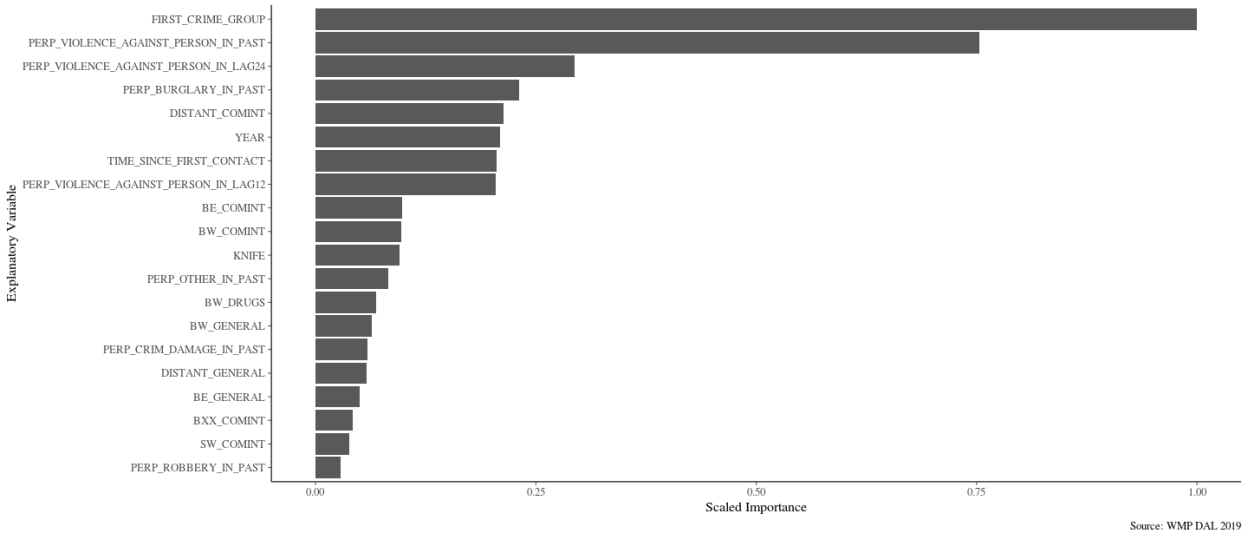


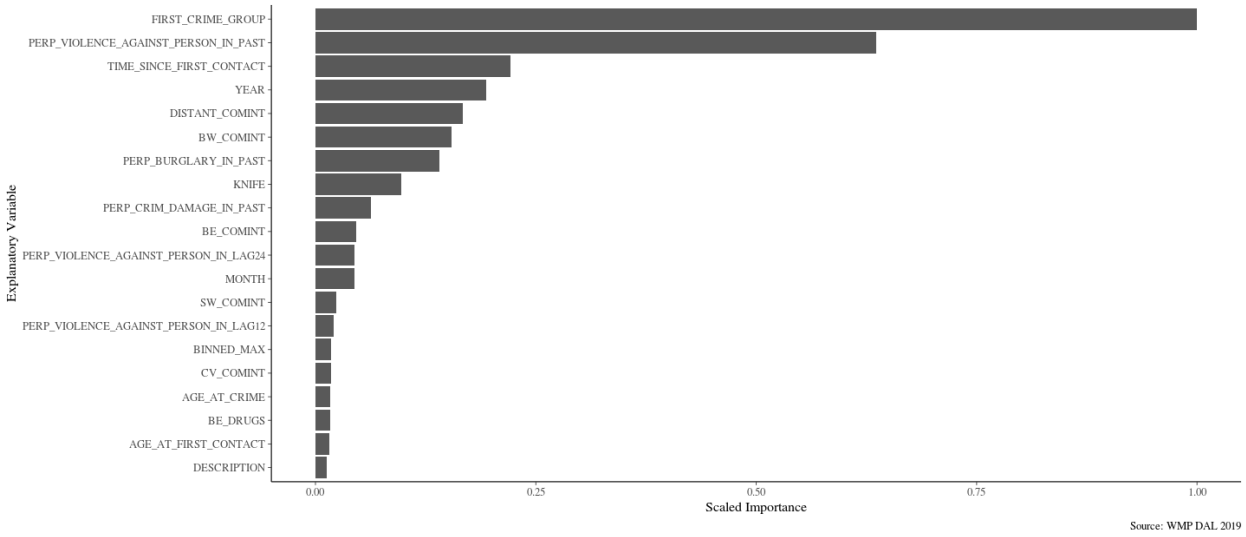
9.5 Random Forests & Gradient Boosted Approaches

In general, these models were confirmatory of the previous analyses.

The top 20 features are presented for each of the crimes for each model using the scaled importance. As with the previous section violence against people is a major factor, though the role of burglary is interesting. It too is included in the LASSO based selection but not in such a prominent manner.

Feature Importance Plot Random Forest Model MSV crime

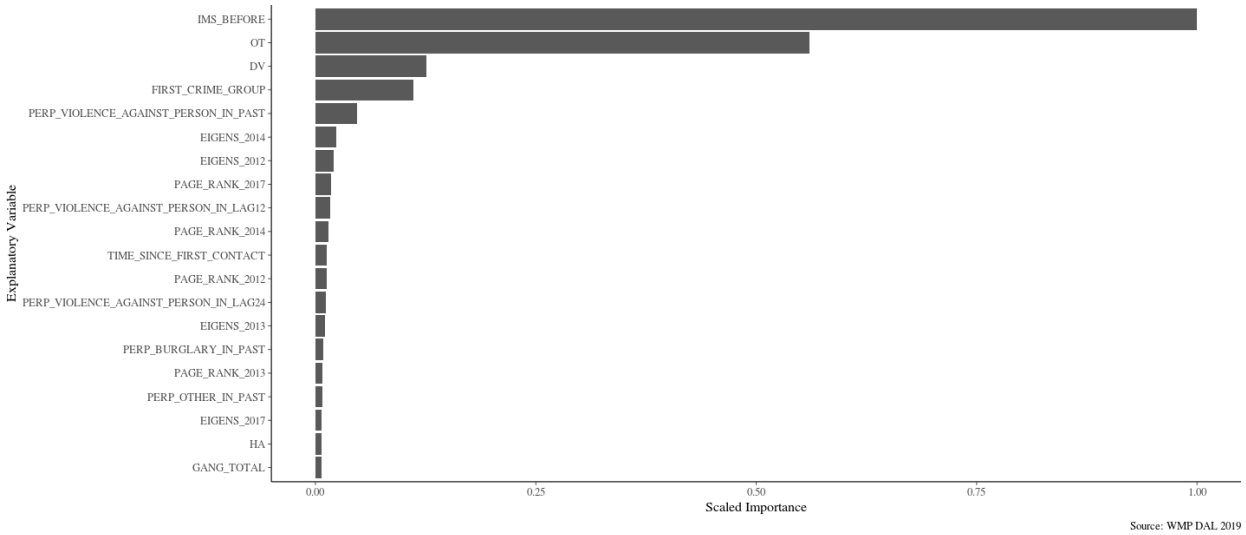




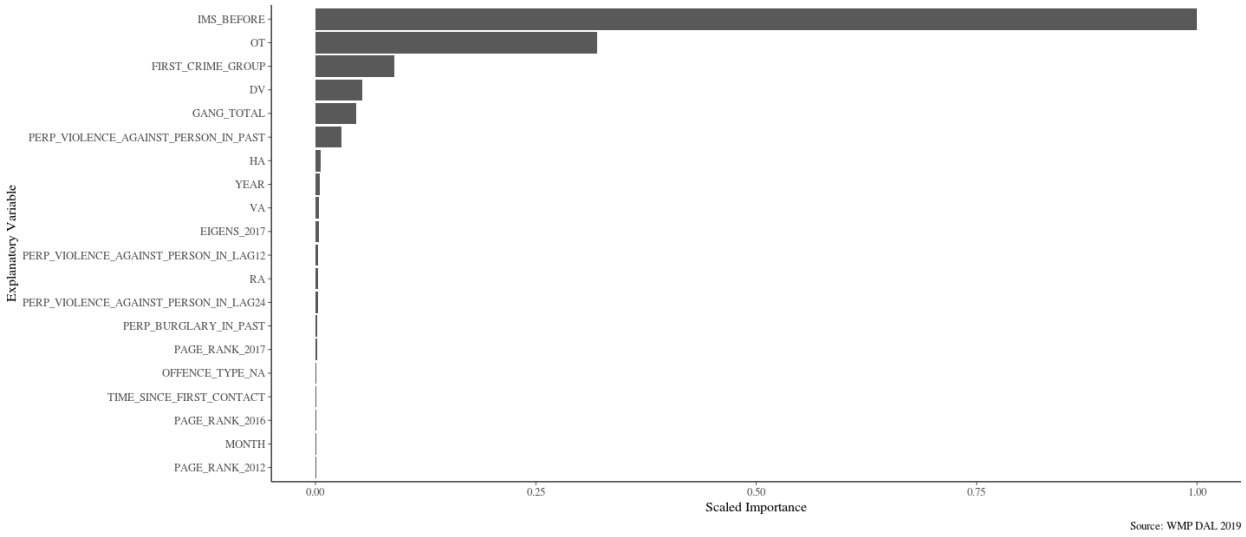
These both highlight the first crime grouping as an important factor. Similarly the age at the first contact is found to be important although perhaps not as high as the other approach found.

In the case of knife crime, we can see some similarities with the MSV selection. However the gang involvement and the connectedness of the individual is also more important than in the overall (MSV) model, though these have relatively low values.

Feature Importance Plot Random Forest Model Knife Crime

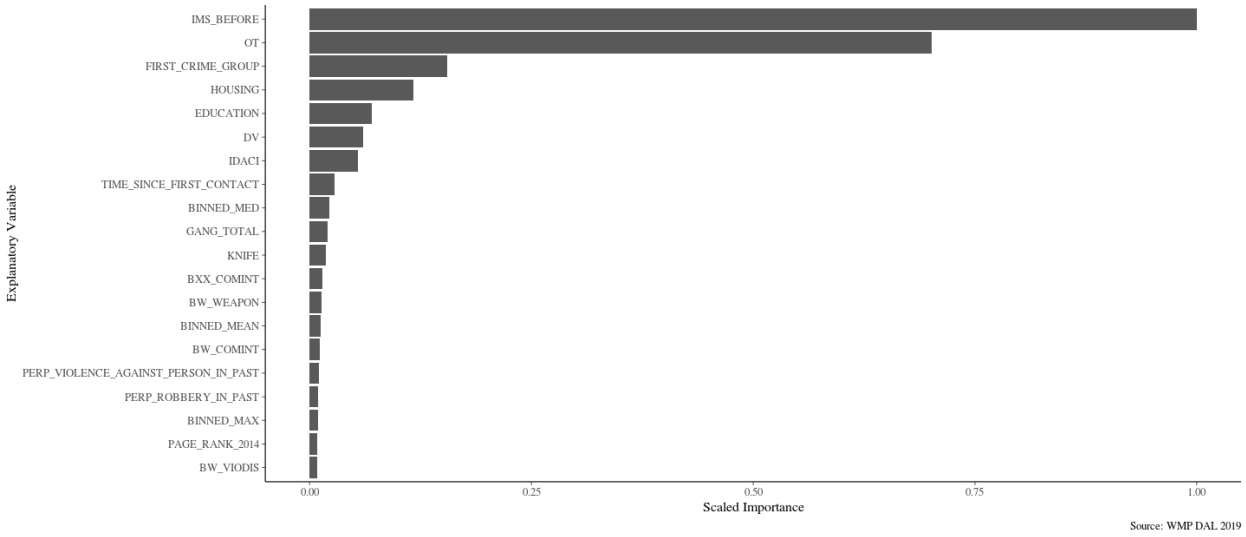


Feature Importance Plot Gradient Boosted Model Knife Crime

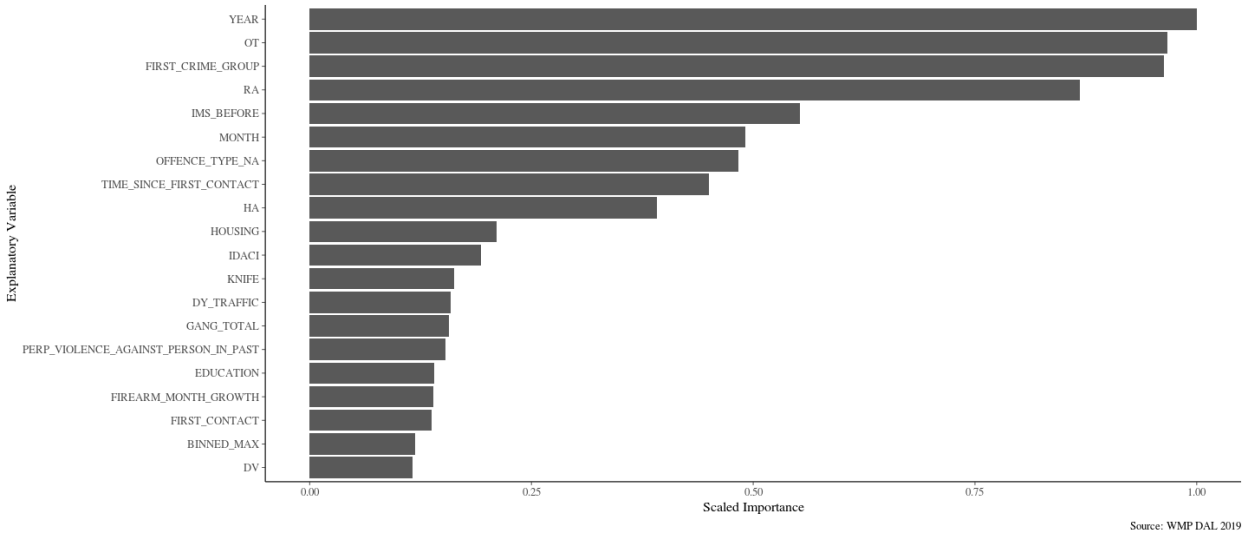


The knife crime models confirm the influence of gangs and connectedness noted previously.

Feature Importance Plot Random Forest Model Firearms



Feature Importance Plot Gradient Boosted Model Firearms



The firearms models suggest that the socio-economic factors are more important in the explanation of the change towards firearms crimes.

9.6 Relaxing the LASSO, Estimation & Model Metrics

Following the variable selection phase, the estimated variable coefficients are (statistically) biased (this is one of the side effects of the variable selection). In order to *correct* this, a second step using a second modelling approach is used. This will be limited in the variables selected and the method is no longer as biased (in the statistical sense of the term) as the first step. Some statistical bias will remain as there is an amalgamation of the variables across the approaches and addition of a number of interaction terms. This uses a second

elastic net regression that directly replicates the approach of Meinshausen (2007). The Elastic Net was estimated with a near ridge constraint in order to reduce the variable selection tendencies. It should be noted there are no standard significance tests or associated p −values with these approaches. Ridge regression and other methods such as the LASSO will bias the coefficients downwards for the selection and upwards for those variables selected and so their use in inference should be treated with care (Cule and De Iorio (2012) do look at a specific approach to the ridge).

The testing of the LASSO is still in its infancy (Tibshirani et al. (2014), van de Geer et al. (2013) and Berk et al. (2013) for example) and further these approaches are only for the ℓ_1 penalty rather than the combination of ℓ_1 and ℓ_2 penalties of the elastic net. The study of {Reid, Tibshirani, and Friedman (2016)} applies only to *linear* problems rather than the non-linear problem here and then only for LASSO rather than other forms of regularisation. Though a bootstrap is possible, Chatterjee and Lahiri (2011) show that the approach is inconsistent for the LASSO (and again no explicit discussion of the more complex regularisation is made).

Results are in the main presented in terms of the effect on the log odds of the event with the Year factor removed as this is an idiosyncratic trend and of little help beyond the observation that some years were higher than others. Months are included where appropriate as these recur.

The results in this form can be interpreted as follows - a positive coefficient reflects an increased chance in the log odds of the event (moving into MSV crime) occurring. These are sometimes reported as odds (as in the bookmakers) where odds more than 1 mean that there is an increased chance of an occurrence; 2 means that the odds double, 3 triples etc. Thus a coefficient of 1.2 in the results translates to an increase in the odds of 3.32 *ceteris paribus*; i.e. a positive coefficient increases the odds / probability of committing MSV / knife / gun crime whilst a negative coefficient decreases the odds / probability of committing such a crime.

There is a caveat. With these models, using a single odds calculation is misleading as one needs to take into account all the other variables in order to think about the probability. Often this is taken at the average levels, say the average age of offenders, however this does not work in the case of (dichotomous) factor variables; has this nominal been involved in a drugs offence yes or no? In order to assess these, all the possible combinations of these factor variables with the other continuous variables need to be calculated. This becomes difficult to isolate the important information; thus it is often easier to look at the (log) odds to identify the relevant variables and to remember that the effect of changing the variables will vary as they themselves vary and as the other variables take other values.

The main performance metrics are reported below with sensitivity, specificity and f1 (the average of the sensitivity and specificity) accompanied by the Youden Index (or J score) (Youden (1950)). This measures the probability of an informed decision. The H-measure (Hand (2009))[^3] looks to deal with some of the issues associated with the AUC that arise from its incoherence with respect to the costs of mis-classification (which must be of importance in the type of situation that is considered here). This measure is included here

with higher values representing better performance. These measures are produced in relation to both training and test sets of data. One would expect the test metrics to be lower than those of the training set, but of the same order. This is the case.

All estimations were performed in R(R Core Team (2018)) or H2O linked from R (LeDell et al. (2019)).

9.7 Logistic Regression

Ordinary Least Squares (OLS) is not valid in the case where there is a limited dependent variable. The classic case is the binary choice case, where the outcome is a *YES* or *NO*. In the case of the logit regression, the estimation is based on the log odds ratio of a success relative to failure estimated via Maximum Likelihood. Other forms of binary choice regression are available, however these are often indistinguishable (Gunduz and Fokoué (2015)) with the relationship between logits and probits particularly well known. The form of the regression is given by:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta^T X + \epsilon$$

This implies that the outcome variable has a non-linear relationship with the explanatory variables (the X's). This therefore implies that the coefficients (β) cannot be interpreted as the impact of changing any one particular explanatory variable. The marginal effects (which do represent the impact of changing one of the explanatory variables by one unit) are dependent on the values of all the explanatory variables. These are usually assumed to be at their mean value, though this is problematic for dummy or factor variables. The intuition behind this is that because probability is bounded between 0 and 1, increasing the value of one variable (given a positive relationship) cannot keep having a positive impact on the probability of the event occurring, otherwise we would achieve a probability of more than 1. Thus the effect of increasing the variable must reduce as we approach the probability of 1. This impact is the marginal effect and it *has* to change. In light of that the coefficients for the log-odds are given.

9.8 Regularized Regression

Regularised regression has a considerable history. This approach constrains the coefficients, having the effect of selecting variables and/ or dealing with multicollinearity. If selection is the sole consideration, the LASSO is used; dealing with multicollinearity is the realm of ridge regression. Using a combination of both of them is known as the Elastic Net (Zou and Hastie (2005)). The estimates are calculated using a set of constraints that reduce the variance of the estimates but bias (in the statistical sense of the term) them relative to ordinary least squares.; i.e. coefficient estimates are biased in order to reduce variance. The form of the problem solved is the Lagrangean:

$$\min_{\beta}(y - \beta^T X)^2 + \lambda\sum(\alpha_1 \mid \beta \mid + (1-\alpha_1)\beta^2)$$

.

The values of λ are selected by cross-validation and the value of α adjusts the relative importance of ridge regression; $\alpha = 1$ is the LASSO whereas $\alpha = 0$ is ridge regression. This too can be estimated by cross-validation as there is no direct analytical solution as there is in the case of β. The equation listed above deals with standard regression (OLS) however this approach is equally applicable to other forms of regression such as logit or Cox survival models and thus applicable in the scenarios considered here. The first step of the variable selection was estimated using a 10 fold cross-validation for both the λ and the α variables. In the case of the λ variable, the $\min + 1SE$ rule was used.

The relaxed LASSO (Meinshausen (2007)) seeks to reduce the bias caused by the initial variable selection. A number of approaches such as LARS-OLS hybrid (Efron et al. (2004)) address this issue, but these are not as effective as the relaxed LASSO that applies the algorithm a second time to reduce the impact of the first regression.

$$\min_{\beta}(y - (\beta^T.1)X)^2 + \phi\lambda \mid \beta \mid$$

$$(\beta^T.1) = \begin{cases} 0 & \text{if variable not included} \\ \beta & \text{if variable is included} \end{cases}$$

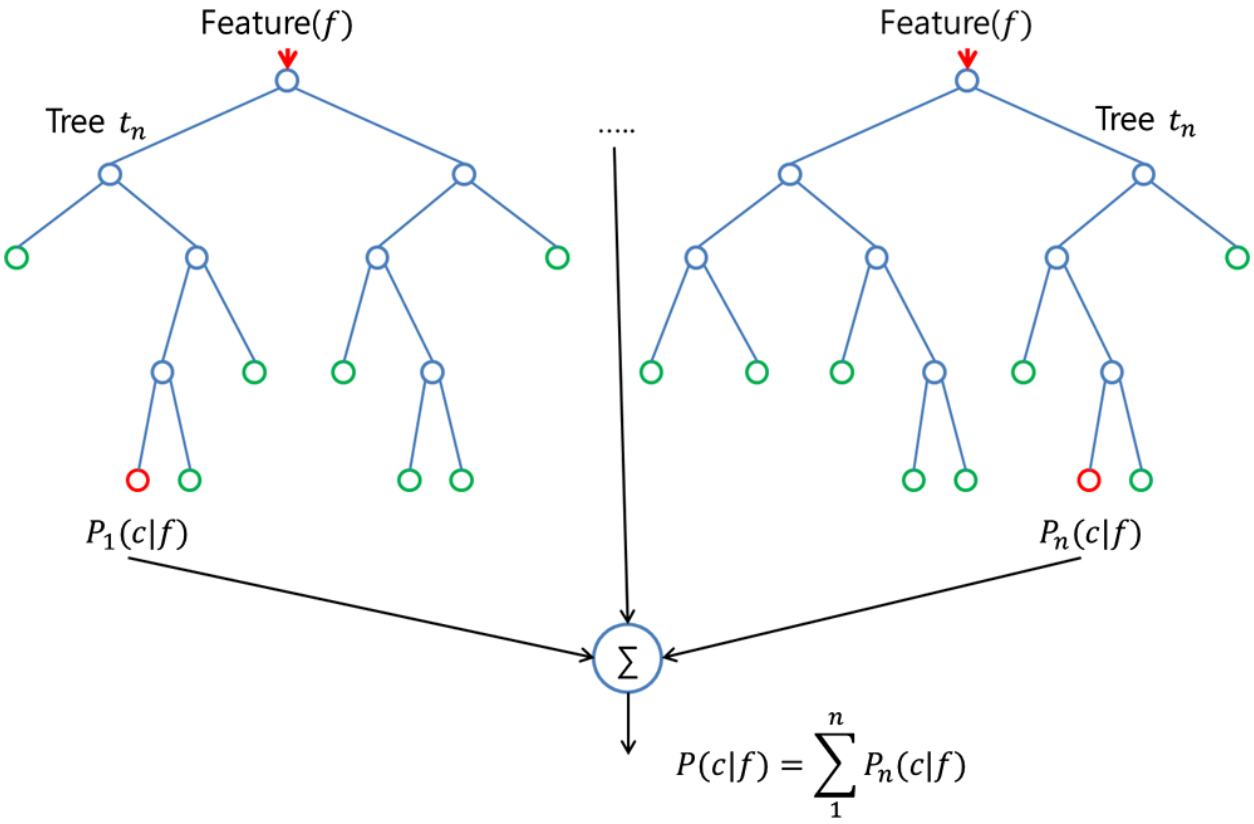
where ϕ is the relaxation parameter and λ is as before.

The relaxation parameter is the equivalent of λ in the second stage, which again used 10-fold cross-validation. The algorithm used in the estimation of the various relationships here are flexible so allowing either an elastic net, LASSO or standard or Bayesian regression to create the second *relaxed* version.

9.9 Random Forests

Random forests use an average of many individual trees to make predictions. The bagging (bootstrap aggregation[^4]) of trees reduces the variance of the estimated prediction function. Random Forest builds a large collection of de-correlated trees, based on the bootstrapped samples and then averages them. It is simpler to train than gradient boosting, often with similar performance. The trees are generated in parallel (thus making the algorithm generally faster than GBM).

Random Forest is generally more robust to over-fitting than other methods (Hastie, Tibshirani, and Friedman (2009)). The diagram below provides a graphical summary of a random forest model:



Random Forest Algorithm

The use of random forests in this project is in variable selection rather than the final modelling step. The random nature of the splits can assist in the discovery of the underlying variables of importance. Further, due to the nature of the tree splits, the variable importance measurement will allow for not just a direct importance to be included, but also the variable interactions to be measured down the tree and across the forest. Variable importance measures the importance of the j^th feature after training, the values of the j^th feature are permuted among the training data and the out of bag (OOB) error is computed on this perturbed dataset. The importance score for that j^th feature is computed by averaging the difference in OOB error before and after the permutation over all trees. The score is normalised by the standard deviation of these differences.

This is a different approach to that of the regressions where a lack of coefficient shrinkage is used to assess importance and interactions are not included as a matter of course.

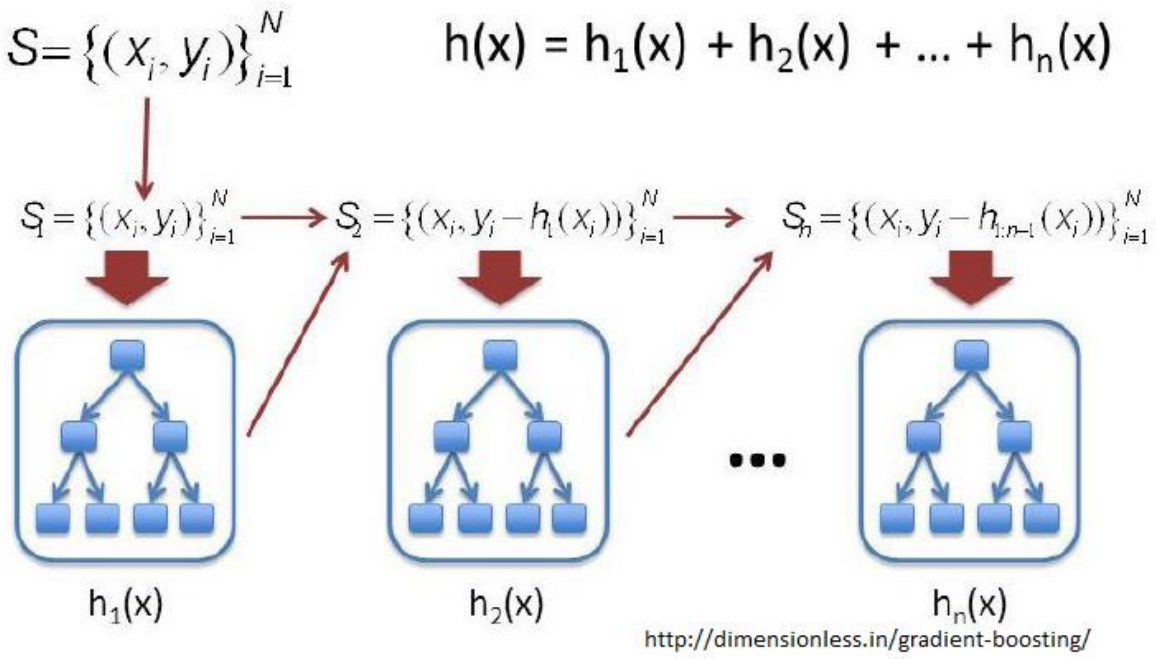
9.10 Gradient Boosted Models

In general terms, the boosting approach averages many trees, each grown to reweighted versions of the training data. The final classifier is a weighted average of the classifiers.

Gradient boosting generalises boosting by allowing optimisation of an arbitrary differentiable loss function.

Gradient boosting inherits all the good features of trees (variable selection, dealing with missing data and mixed predictors), and improves on the weak features, such as prediction performance. The latter is important for the model even when only looking at feature importance.

This can be represented graphically:

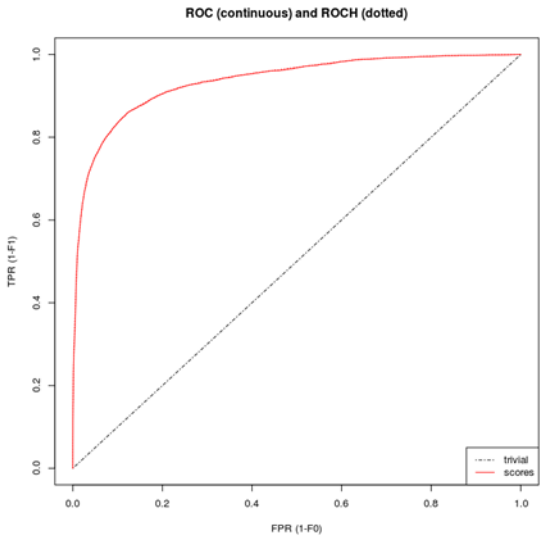


Gradient Boosted Model Algorithm

9.11Performance measures of the main models

Whilst the main models are explanatory rather than predictive, their predictive capacity (with the data split between train and test) can provide a good indication as to their veracity:

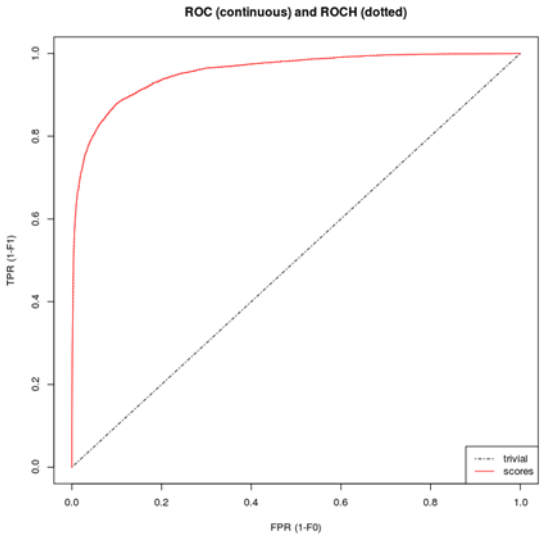
9.11.1 MSV (all)



	30% Cutoff	40% Cutoff	50% Cutoff	60% Cutoff	70% Cutoff
Sensitivity	0.752	0.714	0.675	0.643	0.594
Specificity	0.951	0.964	0.973	0.978	0.983
F1	0.775	0.771	0.758	0.745	0.716
Youden	0.702	0.678	0.648	0.621	0.577
H	0.651	0.651	0.651	0.651	0.651

Table Showing the Accuracy Statistics for MSV (All Involved)

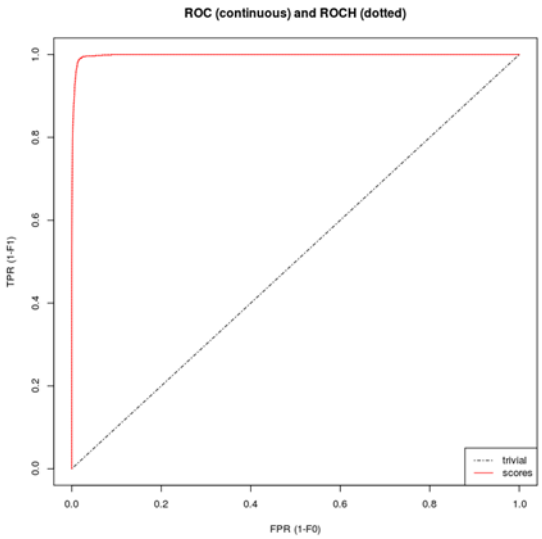
9.11.2 MSV (defendants only)



	30% Cutoff	40% Cutoff	50% Cutoff	60% Cutoff	70% Cutoff
Sensitivity	0.783	0.745	0.711	0.689	0.659
Specificity	0.961	0.972	0.979	0.983	0.988
F1	0.799	0.797	0.788	0.781	0.769
Youden	0.743	0.718	0.691	0.672	0.647
H	0.706	0.706	0.706	0.706	0.706

Table Showing the Accuracy Statistics for MSV (defendants only)

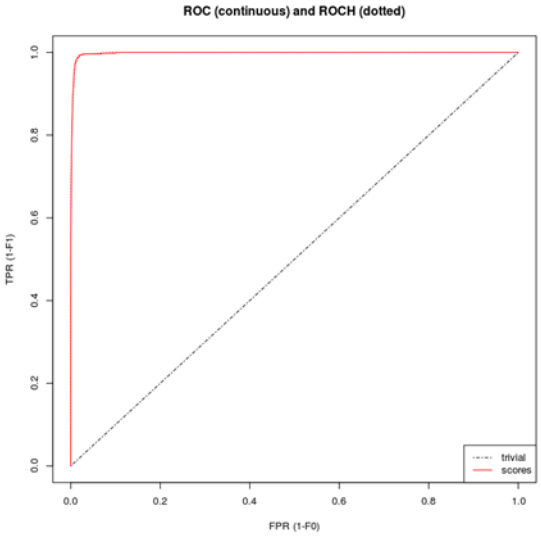
9.11.3 Knife Crime (all)



	30% Cutoff	40% Cutoff	50% Cutoff	60% Cutoff	70% Cutoff
Sensitivity	0.959	0.926	0.875	0.811	0.755
Specificity	0.990	0.993	0.996	0.998	0.998
F1	0.858	0.874	0.878	0.866	0.839
Youden	0.950	0.919	0.871	0.809	0.753
H	0.954	0.954	0.954	0.954	0.954

Table Showing the Accuracy Statistics for Knife Based MSV (All Involved)

9.11.4 Knife Crime (defendants only)

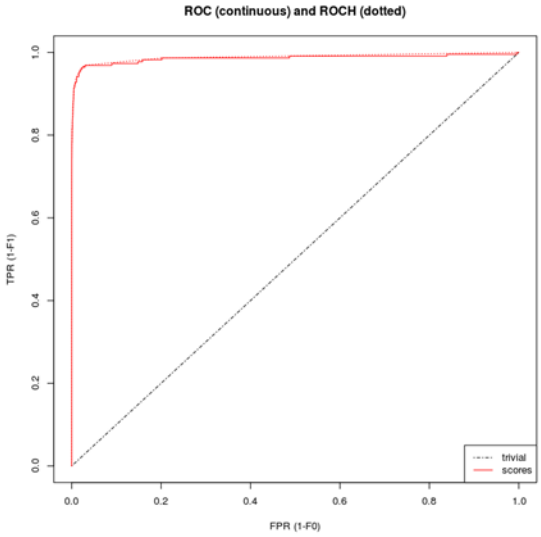


	30% Cutoff	40% Cutoff	50% Cutoff	60% Cutoff	70% Cutoff
Sensitivity	0.957	0.912	0.868	0.814	0.773
Specificity	0.992	0.994	0.996	0.997	0.998
F1	0.862	0.868	0.867	0.854	0.841
Youden	0.949	0.907	0.864	0.811	0.771
H	0.956	0.956	0.956	0.956	0.956

Table Showing the Accuracy Statistics for Knife Based MSV (defendants only)

A test for heteroskedasticity was also performed using a form of Breusch-Pagan test (Breusch and Pagan (1979)) based on the squared residuals[^2] and the explanatory variables. All the variables had coefficients that were of the order $10^{-}2$ at most. There is some explanatory power to these variables however these variables are those which are generally rare (mostly 0) and so the variation in the squared residuals in conjunction with the explanatory terms is very limited.

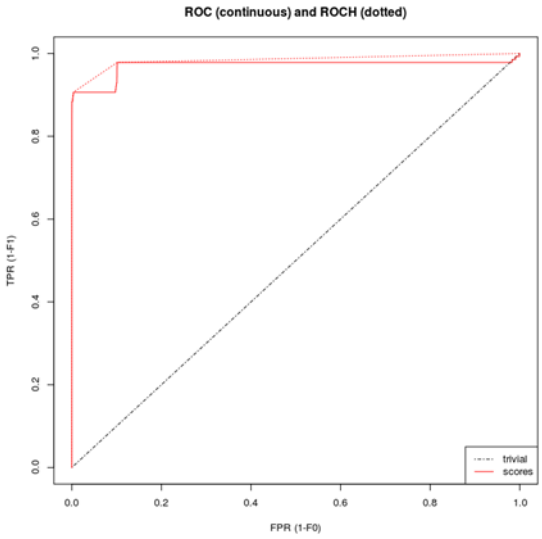
9.11.5 Firearms Crime (all)



	30% Cutoff	40% Cutoff	50% Cutoff	60% Cutoff	70% Cutoff
Sensitivity	0.786	0.759	0.750	0.741	0.732
Specificity	0.999	1.000	1.000	1.000	1.000
F1	0.840	0.842	0.838	0.836	0.832
Youden	0.785	0.758	0.750	0.741	0.732
H	0.917	0.917	0.917	0.917	0.917

Table Showing the Accuracy Statistics for Firearms Based MSV (All Involved)

9.11.6 Firearms Crimes (defendants only)



	30% Cutoff	40% Cutoff	50% Cutoff	60% Cutoff	70% Cutoff
Sensitivity	0.871	0.820	0.770	0.763	0.755
Specificity	1.000	1.000	1.000	1.000	1.000
F1	0.903	0.880	0.849	0.848	0.847
Youden	0.870	0.820	0.770	0.762	0.755
H	0.893	0.893	0.893	0.893	0.893

Table Showing the Accuracy Statistics for Firearms Based MSV (only defendants)

10 Notes

[^1] In the case of levels & year on year differences, a regression was run to ascertain the importance of the break. The Bayes Factor in the levels is extremely large (10^41), whereas in the logged differences (i.e. growth rates) the Bayes Factors were extremely small 10^-1). This is suggestive of non-stationarity in the data. Looking at Knife Crime, comparison with an intercept model for the counts and year on year differences give similar outcomes, though with a little more (but still weak) evidence in support of the break in growth rates. Comparing the knife crime counts with a single break, rather than the full model with the 2 breaks gives weak evidence in both levels and log differences.

[^2] It is noted that these are not *true* residuals in the OLS sense. Rather they are used to proxy the systematic discrepancy based on the variables in point. Logits are always heteroskedastic (the outcomes are 0,1 and thus this should be considered as a base line)

[^3] The H-measure (Hand (2009)) allows for the weighting of the costs differently across the outcomes. This allows misclassifying a positive outcome (i.e. involvement in a crime) to be weighed by more than a negative outcome. Hand suggests a $beta(x; 2,2)$ distribution in order to standardise the measure allowing comparisons across classifiers.

[^4] Bagging involves the use of bootstrap sampling on the data set to generate m bootstrap samples. Models are fitted across these m data sets. With the outcome involving aggregation, either by averaging in the case of regressions or voting for classification algorithms.

11 Data Dictionary

In the context of these data definitions, first MSV includes first knife crime, first firearms crime, and first control-type crime.

11.1.1 Nominal history

A group of variables have the format ROLE_CRIMEGROUP_LAG, for example PERP_THEFT_AND_HANDLING_IN_LAG12. The values for these variables are counts.

The roles are:

Role	Description
PERP	defendant/offender, person probably responsible, person thought responsible for the offence, suspect
VICTIM	victim or additional victim
OTHER	other role e.g. witness, sibling

The crime groups are:

Crime Group
BURGLARY
CRIM_DAMAGE
DRUGS
FRAUD_FORGERY
OTHER
ROBBERY
SEXUAL_OFFENCES
THEFT_AND_HANDLING
VIOLENCE_AGAINST_PERSON

The lags are:

Lag	Description
IN_LAG12	in the 12 months prior to the first MSV
IN_LAG24	in the 24 months prior to the first MSV
IN_PAST	at any time prior to the first MSV

11.1.2 Intelligence by area

Another large group of variables relate to intelligence, and have been split into different areas. These take the form AREA_CRIMECODE, for example BE_CRIMDAM showing the area where the intelligence originates and the type of intelligence . The values in the dataset are counts.

Not every area had every type of intelligence; the following tables show all the possible areas and types.

11.1.3 Areas:

Code	Area
BE	Birmingham East
BW	Birmingham West
BXX	Birmingham - general
CV	Coventry
DISTANT	more than 2 counties from WMP area
DY	Dudley
NEXT_COUNTY	County next to WMP area
SH	Solihull
SW	Sandwell
TWO_COUNTY	two counties away from WMP area
WS	Walsall
WV	Wolverhampton
WXX	Unknown location in Walsall or Wolverhampton

11.1.4 Intelligence types

Code	Intelligence type
ANTISOC	antisocial behaviour
BRGLRY	burglary
COMINT	community
CRIMDAM	criminal damage
CTI	counter terrorism
DRUGS	drugs
FRAUD	fraud
GENERAL	general intelligence
OPS	operations
OTHER	other intelligence
OUTSIDE	outside intelligence
PROTECT	protection
ROB	robbery
SERCOBRCH	Breach of SERCO conditions eg tag, curfew, bail
SPORT	sport
THEFT	theft
TRAFFIC	traffic
VEHINFO	vehicle information
VIODIS	violent disturbance
WEAPON	weapon

11.1.5 Other Variables

Variable name	Description	Extra info	Levels/range/type
AGE_AT_CRIME	age at time of first MSV or other target crime type		0 to 25
AGE_AT_FIRST_CONTACT	age at time of first offence		0 to 25
DESCRIPTION	ethnic appearance		Asian, Black, Chinese, Middle Eastern, Not Known, Other, White North European, White South European
SEX	sex of nominal		F, M, U
ROLE_DESC	type of perpetrator role in first relevant crime (first MSV/knife/firearm or control)		defendant/offender, person probably responsible, person thought responsible for the offence, suspect

NEARFAMILY	occurrences of spouse, child, stepchild, stepparent, spouse (inc common law), partner (inc boyfriend, girlfriend)		count
FARFAMILY	occurrences of family not included in NEARFAMILY		count
FIRST_CRIME_GROUP	grouping for first crime implicating nominal	Arson, assault, blackmail, bodily harm, burglary, child abuse, criminal justice, damage, drug, DV, firearm, fraud, kidnap, knife, modern slavery, murder, other, public order, robbery, sex, theft, threaten, vehicle, weapon (These are inclusive of conspiracy or attempted variants)	
IMS_BEFORE	whether nominal has IMS mention before first MSV or equivalent		0,1
TIME_KNOWN_WKS	weeks between first IMS mention and first relevant crime	negative if crime came first, 0 otherwise	numeric
TIME_SINCE_FIRST_CONTACT	weeks from first crime to first relevant crime	0 if both are same incident	rounded
TOTAL_FIREARMS_LAG1 TOTAL_FIREARMS_LAG2 TOTAL_FIREARMS_LAG12 TOTAL_FIREARMS_LAG13	crimes involving firearms by nominal	in past month in past 2 months in past 12 months in past 13 months	count
TOTAL_KNIFE_LAG1 TOTAL_KNIFE_LAG2 TOTAL_KNIFE_LAG12 TOTAL_KNIFE_LAG13	crimes involving knives by nominal	in past month in past 2 months in past 12 months in past 13 months	count
CSE	victim of child sexual exploitation	from Compact data	FALSE, TRUE
DV_COMPACT	victim of domestic violence	from Compact data	FALSE, TRUE
MENTAL_ILLNESS	has or had a mental	from Compact	FALSE, TRUE

	illness	data	
MOD_SLAV	victim of modern slavery	from Compact data	FALSE, TRUE
EIGENS_2012 EIGENS_2013 EIGENS_2014 EIGENS_2015 EIGENS_2016 EIGENS_2017	Eigen measure of connectedness to other criminals as specified in IOMS model by year		higher absolute value = more connected
PAGE_RANK_2012 PAGE_RANK_2013 PAGE_RANK_2014 PAGE_RANK_2015 PAGE_RANK_2016 PAGE_RANK_2017	pagerank measure of connectedness to other criminals as specified in IOMS model by year		higher absolute value = more connected
gang_total	total mentions of gang membership		count
GANG_N	Nominal involved in gang	gang_total>0	0,1
CA	child abuse	offence type	0, 1
CLINES	county lines	offence type	FALSE, TRUE
DV	domestic violence	offence type	0, 1
HA	hate crime	offence type	0, 1
HATE_FLAG	whether a hate strand has been recorded		0, 1
HO	homophobic	offence type	0, 1
OFFENCE_TYPE_NA	no offence type details	offence type	0, 1
OT	other	offence type	0, 1
RA	racist	offence type	0, 1
YP VA	young person Vulnerable adult	offence type	0, 1
DRUG	drug involved		FALSE, TRUE
FIREARM	whether firearm involved in crime		0, 1
KNIFE	whether knife involved in crime		0, 1
WEAPON_OTHER	whether weapon not knife or firearm involved in crime		0,1
YEAR	Year first relevant crime occurred		2000 to 2018
BINNED_MAX BINNED_MEAN	distance between nominal address	maximum mean	within 1km, 3km, 5km, 7.5km, more than

BINNED_MED	and crime address	median	7.5km, missing (no data)
EDUCATION	from Education sub-domain of IMD	Decile 1, deciles 2 & 3, deciles 4 to 10, unknown	
HOUSING	from Housing sub_domain of IMD	Decile 1, deciles 2 to 4, deciles 5 to 10, unknown	
IDACI	from Income Deprivation affecting children index	Decile 1, deciles 2 & 3, deciles 4 to 10, unknown	
FIREARM_MONTH_GROWTH	percentage increase in firearms in past month		percentage
FIREARM_YEAR_GROWTH	percentage increase in firearms in past year		percentage
KNIFE_MONTH_GROWTH	percentage increase in knives in past month		percentage
KNIFE_YEAR_GROWTH	percentage increase in knives in past year		percentage
MONTH	month offence occurred		January to December

12 References

Berk, Richard, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. 2013. "Valid Post-Selection Inference." *Ann. Statist.* 41 (2): 802–37. https://doi.org/10.1214/12-AOS1077.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Breusch, T. S., and A. R. Pagan. 1979. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica* 47 (5): 1287–94. http://www.jstor.org/stable/1911963.

Chatterjee, A., and S. N. Lahiri. 2011. "Bootstrapping Lasso Estimators." *Journal of the American Statistical Association* 106 (494): 608–25. https://doi.org/10.1198/jasa.2011.tm10159.

Cox, D. R. 1958. "The Regression Analysis of Binary Sequences." *Journal of the Royal Statistical Society. Series B (Methodological)* 20 (2): 215–42. http://www.jstor.org/stable/2983890.

Cule, Erika, and Maria De Iorio. 2012. "A semi-automatic method to guide the choice of ridge parameter in ridge regression." *arXiv E-Prints*, May, arXiv:1205.0686. http://arxiv.org/abs/1205.0686.

Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. 2004. "Least Angle Regression." *Ann. Statist.* 32 (2): 407–99. https://doi.org/10.1214/009053604000000067.

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Ann. Statist.* 29 (5): 1189–1232. https://doi.org/10.1214/aos/1013203451.

Friedman, Jerome H. 2002. "Stochastic Gradient Boosting." *Computational Statistics & Data Analysis* 38 (4): 367–78.

Gunduz, Necla, and Ernest Fokoué. 2015. "On the Predictive Properties of Binary Link Functions." *arXiv Preprint arXiv:1502.04742*.

Hand, David J. 2009. "Measuring Classifier Performance: A Coherent Alternative to the Area Under the Roc Curve." *Machine Learning* 77 (1): 103–23. https://doi.org/10.1007/s10994-009-5119-5.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. 2nd ed. Springer. http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

LeDell, Erin, Navdeep Gill, Spencer Aiello, Anqi Fu, Arno Candel, Cliff Click, Tom Kraljevic, et al. 2019. *H2o: R Interface for 'H2o'*. https://CRAN.R-project.org/package=h2o.

Meinshausen, Nicolai. 2007. "Relaxed Lasso." *Computational Statistics & Data Analysis* 52 (1): 374–93.

Owen, Art B. 2007. "Infinitely Imbalanced Logistic Regression." *Journal of Machine Learning Research* 8 (Apr): 761–73.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reid, Stephen, Robert Tibshirani, and Jerome Friedman. 2016. "A Study of Error Variance Estimation in Lasso Regression." *Statistica Sinica*, 35–67.

Sherman, Lawrence, Peter William Neyroud, and Eleanor Neyroud. 2016. "The Cambridge Crime Harm Index: Measuring Total Harm from Crime Based on Sentencing Guidelines." *Policing: A Journal of Policy and Practice* 10 (3): 171–83. https://doi.org/10.1093/police/paw003.

Theil, Henri. 1971. “Applied Economic Forecasting.”

Tibshirani, Robert. 1996. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–88.

Tibshirani, Ryan J., Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. 2014. “Exact Post-Selection Inference for Sequential Regression Procedures.” *arXiv E-Prints*, January, arXiv:1401.3889. http://arxiv.org/abs/1401.3889.

van de Geer, Sara, Peter Bühlmann, Ya’acov Ritov, and Ruben Dezeure. 2013. “On asymptotically optimal confidence regions and tests for high-dimensional models.” *arXiv E-Prints*, March, arXiv:1303.0518. http://arxiv.org/abs/1303.0518.

Youden, W. J. 1950. “Index for Rating Diagnostic Tests.” *Cancer* 3 (1): 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

Zou, Hui, and Trevor Hastie. 2005. “Regularization and Variable Selection via the Elastic Net.” *Journal of the Royal Statistical Society, Series B* 67: 301–20.