MSV Crime Prediction

Data Analytics Lab

November 2021

This project aims to make predictions of most serious violence (MSV) crime in terms of the likely number and location of such crimes over a 4 weekly period.

1 Contents

2 Introduction ... 3

3 Exploratory (Spatial) Data Analysis ... 5

4 Modelling Approach ... 8

Appendix – Definition of MSV ... 9

Appendix – Modelling Details ... 11

Identifying the best Model ... 15

References ... 20

Figure 1: monthly MSV from 2009 - 2021 ... 3
Figure 2: incidents of MSV ... 5
Figure 3: Density of MSV Crime 2019 ... 6
Figure 4: distribution of monthly count of MSV crimes ... 6
Figure 5: MSV crimes by day of week, 2019 (total and distribution) ... 7
Figure 6: count of crimes by hour of day, 2019 ... 7
Figure 7: example of predictions for a 4 week period (note: accuracy measures, etc. are noted in the appendix) ... 8
Figure 8: testing for CSR. ... 11
Figure 9: global spatial autocorrelation ... 12
Figure 10: cluster map of local indicator of spatial association (2020 – 2021) ... 12
Figure 11: time series of MSV crimes for 20 years ... 13
Figure 12: decomposition of MSV crime time series ... 13
Figure 13: autocorrelation and partial autocorrelation functions of MSV crime (lags on the x axis, correlation on the y axis) ... 14
Figure 14: map of relative score of an empirical Bayes scan statistic ... 14
Figure 15: initial models (RMSE calculated over a single prediction period) ... 16
Figure 16: RMSE of final models ... 16
Figure 17: Mean absolute error (MAE) of models ... 17
Figure 18: accuracy measures for moving window predictions ... 17
Figure 19: Benefit - Cost of the different models ... 18
Figure 20: actual MSV crimes month 1 ... 19
Figure 21: predictions from the GAM (left) and residuals (right) ... 19
Figure 22: using the naive (mean-based) approach (left) and residuals (right) ... 19

2 Introduction

The overall aim of this project is to develop a statistical model to predict the likely levels of most serious violence (MSV) over time and space within the WMP area.

There has been a notable increase in violence within several of the UK's urban areas, and within the West Midlands in particular. In this region, gun crime has increased by 33%, and instances of MSV crime have increased by 85% since 2012 and violent crime against the person is up 32% in the last year.[1]

In addition, the rate per 1000 residents of Violence with Injury offences in the West Midlands is above average when compared to the average for England and Wales and compared to our most similar force of Merseyside.[2]

The project was requested by Project Guardian, the aim of which is to reduce serious violence, in particular between young people in public spaces.

Increases in MSV in the West Midlands reflect the national trend; the Force has seen an increase in reporting from a low in 2012/13. Indeed, the mean number of monthly MSV in 2019 represents an increase of 79% over the mean number of monthly crimes in 2012.

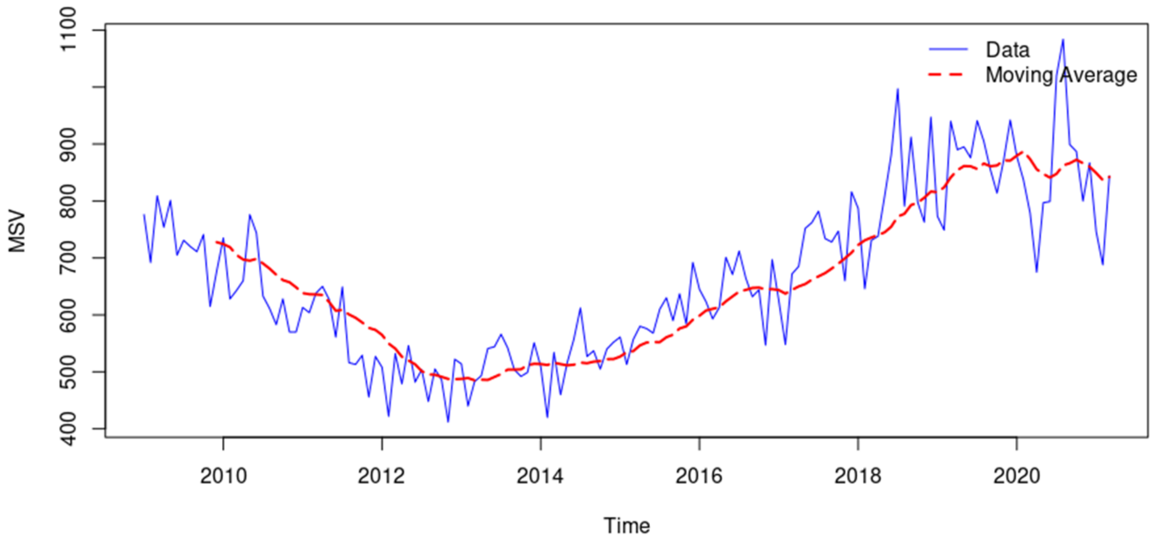

Figure 1: monthly MSV from 2009 - 2021

[1] WMP OPCC: SPCB Violence Reduction Unit paper 19/11/2019 Item 8b https://www.westmidlands-pcc.gov.uk/strategic-policing-crime-board/agendas-minutes-reports/

[2] *Office for National Statistics Crime in England and Wales: Police Force Area data tables year ending Dec 2019: E&W 9.2; WMP 11.5; Merseyside 10.5; W. Yorkshire 11.8; no data for Greater Manchester.* https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/datasets/policeforceareadatatables

In April 2019 the Home Office gave West Midlands Police (WMP) £7.62million in police surge funding with the mandate that it is to be used to reduce serious violence in public spaces, with a focus on reducing MSV and MSV crimes among young people. The force's response has been to create a team, known as Project Guardian. The team works closely with the Violence Reduction Unit (VRU) launched the same year by the Police and Crime Commissioner (PCC) and other local agencies. In this third year of funding (2021/22), the Home Office awarded a grant of £2,940,000 to augment WMP's operational response to serious violence. The specified areas of focus include identifying and monitoring hotspot activity and crime within them and the use of visible patrols to reduce violent crime in hotspots.[3]

The types of activity envisaged by the Home Office include uniformed / visible patrols, problem oriented policing (including intelligence led operations), offender-targeted activity, weapons sweeps and drug enforcement. These have been agreed by the National Police Chiefs Council (NPCC).[4]

The output of this analysis will be a dashboard for the use of the intelligence team which supports Project Guardian. In conjunction with other information, such as intelligence reports, the team will use the prediction to make recommendations to the monthly Force Tasking and Delivery Board (FTDB) about where to allocate resources. The nature of the policing activity undertaken will depend on the type of issues presenting in the predicted hotspots at the time and on any current activity by WMP and partners.

[3] Letter from Home Office outlining grant agreement for Police Grip Funding 2021/22

[4] Hotspots Grip 2021_22 reporting sheet supplied by Home Office.

3 Exploratory (Spatial) Data Analysis

The data used in this project relate to MSV (see the appendix for a definition). For the exploratory (spatial) data analysis data going back to the year 2009 have been used (see the appendix as to why the year 2009).

In a spatial sense, the incidents of MSV tend to exhibit clustering, particularly around Coventry, Birmingham and Wolverhampton (and to a lesser extent Walsall) city centres.

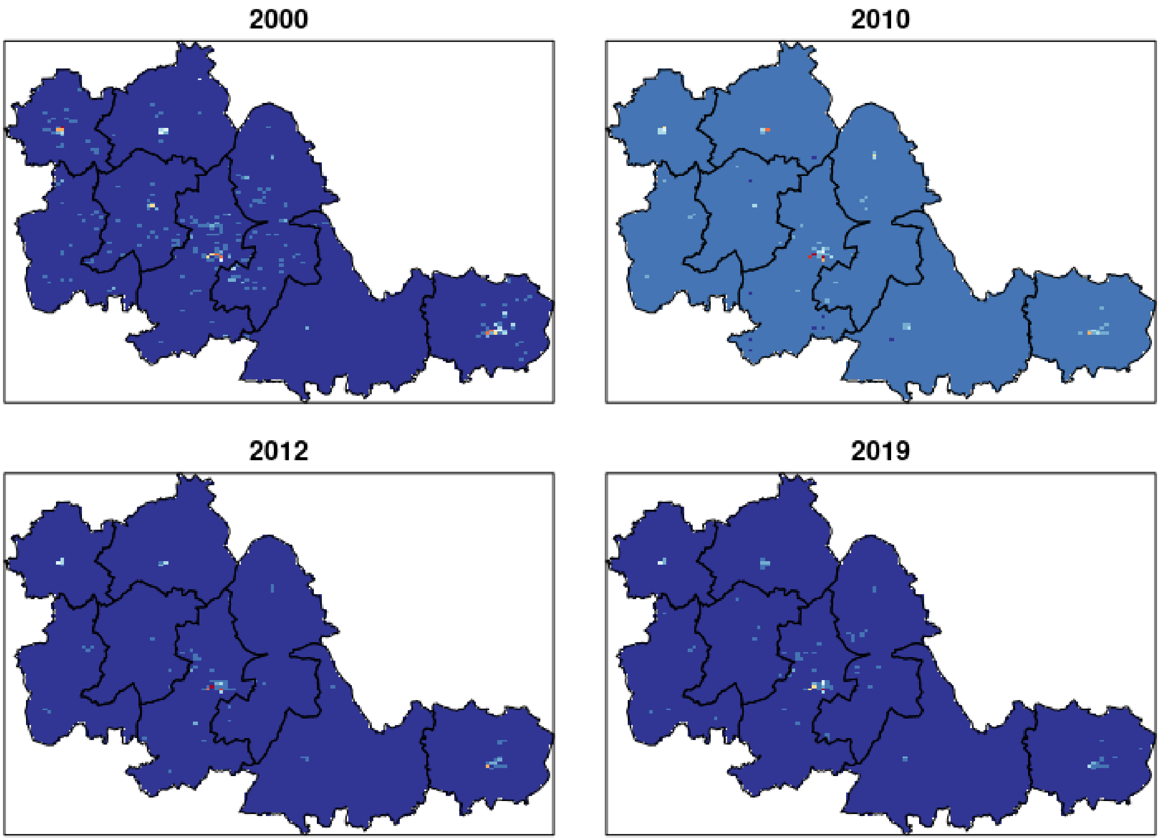

Figure 2: incidents of MSV

The figure below highlights the spatial distribution of MSV crimes over the WMP area for the year 2019. The numbers occurring in and around the centres is apparent:

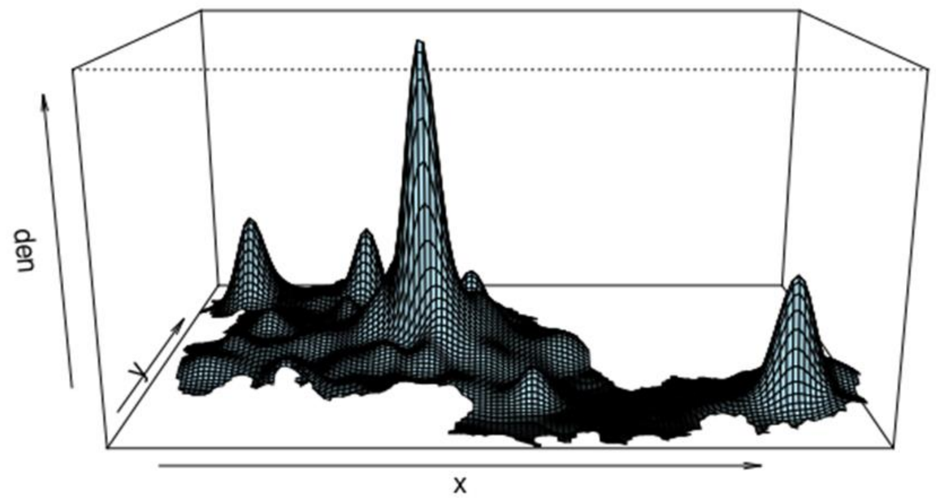

Figure 3: Density of MSV Crime 2019

It can also be seen from the chart below that over recent years (2018 – 2020) there have been between 60 – 100 crimes per month:

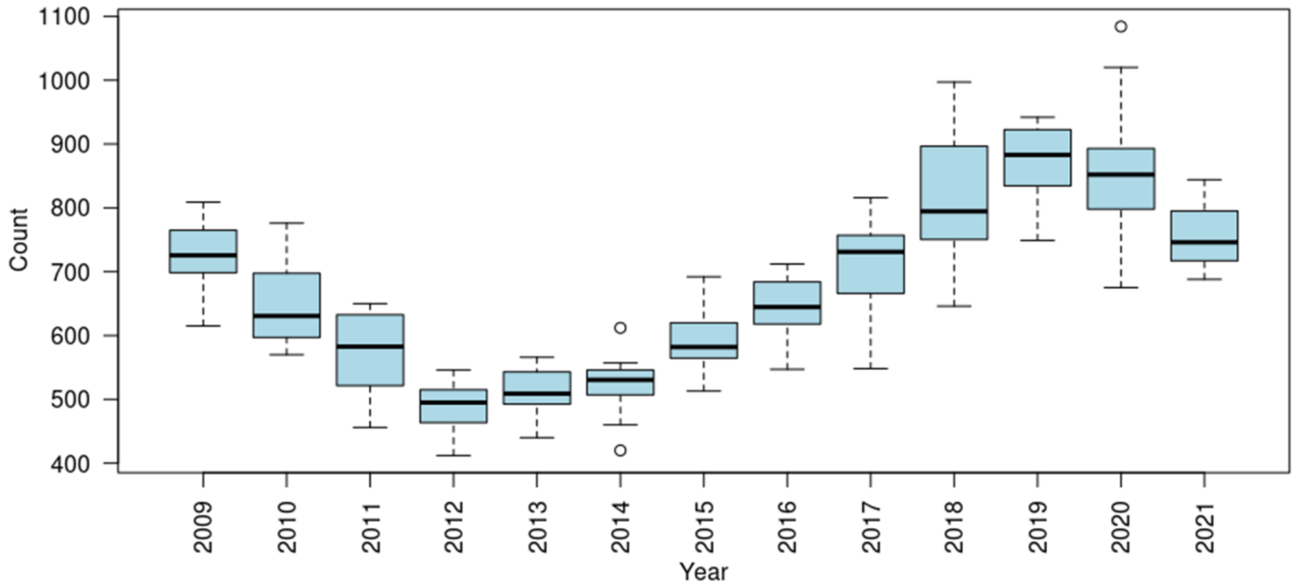

Figure 4: distribution of monthly count of MSV crimes

Over the course of 2019, there have tended to be more MSV crimes at weekends:

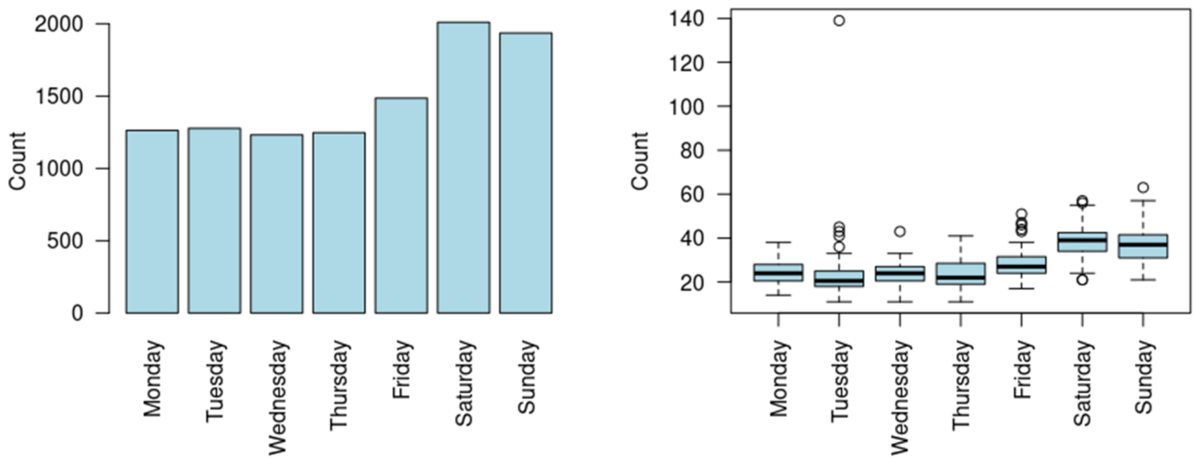

Figure 5: MSV crimes by day of week, 2019 (total and distribution)

As can be seen from the charts below, depending on the day of the week, the majority of crimes occur either in the late afternoon / early evening or between circa 22:00 – 01:00.

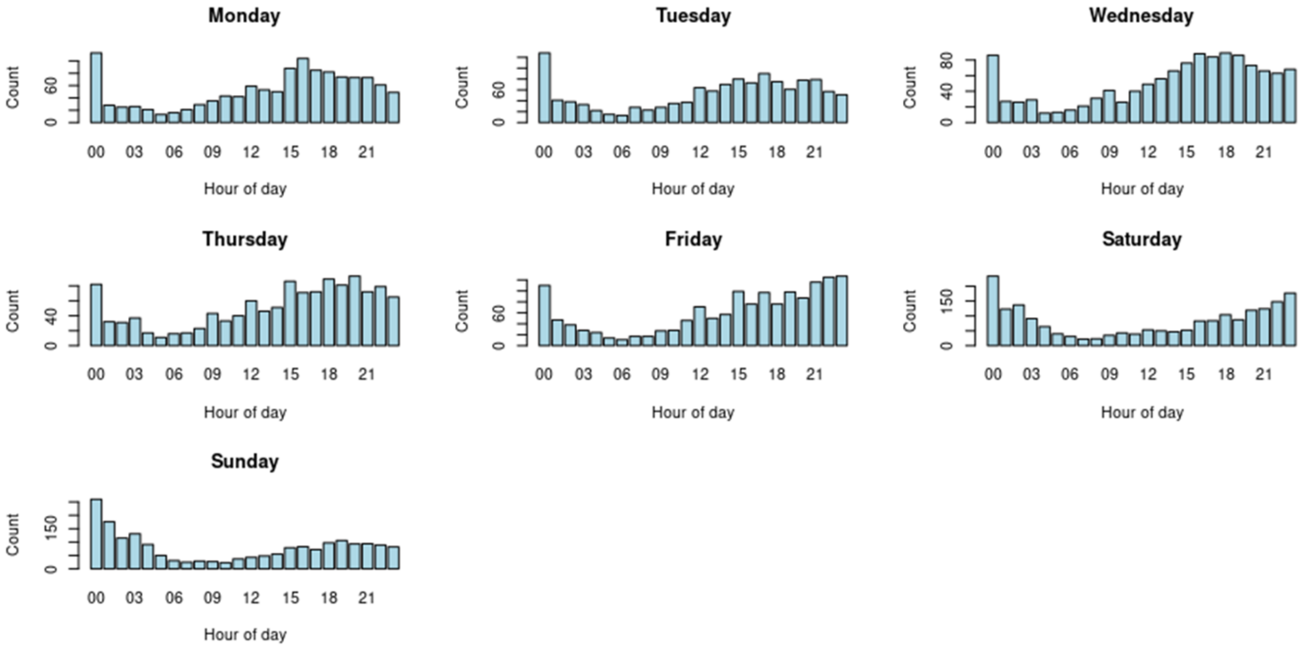

Figure 6: count of crimes by hour of day, 2019

4 Modelling Approach

The preceding analyses and those contained in the appendix show that a spatio-temporal model using past levels of crimes as an additional feature is appropriate. Details of both these analyses and the final models are in the appendix (where it is also shown that using the model performs better than using a general hotspot approach).

Due to identifiable areas being most useful operationally, for the purposes of this project the WMP area has been split into a grid with each grid being circa 0.5 square km.

An example of the resultant output is shown in the figure below (for actual use, it is intended that this will be shown on a dashboard):

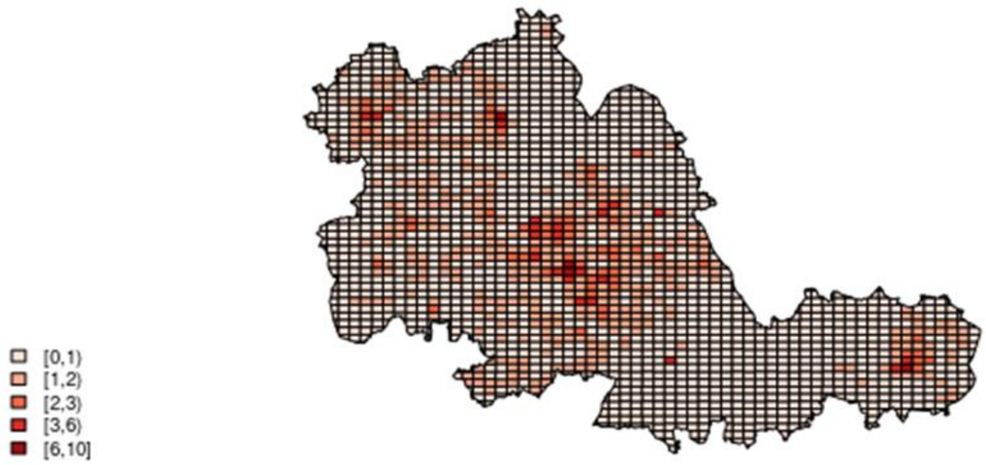

Figure 7: example of predictions for a 4 week period (note: accuracy measures, etc. are noted in the appendix)

Appendix – Definition of MSV

A list of offence codes from previous projects have been agreed with subject matter experts to include only 'most violent crime' (not including domestic abuse). The same definition is used here. The current list is as follows.

Offences	
RACIALLY AGGRAVATED MALICIOUS WOUNDING	ATTEMPT TO INFLICT GBH
WOUND W/I TO RESIST/PREVENT ARREST	CAUSE NOXIOUS THING TAKEN-ENDANGER LIFE
RACIALLY AGGRAVATED WOUNDING S.20	CAUSE POISON TO BE TAKEN-ENDANGER LIFE
POLICE - ATTEMPT TO CAUSE S.18 GBH WITH INTENT TO DO GBH	CAUSE GBH WITH INTENT
RELIGIOUSLY AGGRAVATED INFLICTING GBH, WITHOUT INTENT	CAUSE GBH W/I TO RESIST/PREVENT ARREST
POLICE - INFLICTING GBH WITHOUT INTENT	ADMINISTER POISON TO ENDANGER LIFE
POLICE - MALICIOUS WOUNDING	CAUSE POISON ADMINISTERED-ENDANGER LIFE
WOUNDING	ADMINISTER POISON SO AS TO INFLICT GBH
MANSLAUGHTER	ATTEMPT TO CAUSE GBH W/I TO DO GBH
INFLICTING GBH WITHOUT INTENT	CONSPIRACY TO WOUND W/I TO DO GBH
POLICE - S.18 CAUSE GREVIOUS BODILY HARM WITH INTENT TO DO GBH	ATTEMPT TO CHOKE/SUFFOCATE/STRANGLE W/I
MALICIOUS WOUNDING	ATT CAUSE GBH W/I RESIST/PREVENT ARREST
RACIALLY AGGRAVATED INFLICTING GBH WITHOUT INTENT	CONSPIRE MURDER VICTIM 1 YR OLD OR OVER
THROW EXPLOSIVE SUBSTANCE W/I	ATTEMPT TO INFLICT GBH WITHOUT INTENT
RACIALLY/RELIGIOUSLY AGGRAVATED INFLICTING GBH WITHOUT INTENT	ADMINISTER NOXIOUS THING-ENDANGER LIFE
SOLICITE TO MURDER	ATTEMPT MALICIOUS WOUNDING
WOUND WITH INTENT TO COMMIT GBH	ATTEMPT MURDER VICTIM UNDER 1 YR OLD
RELIGIOUSLY AGGRAVATED WOUNDING/GBH	CONSPIRACY TO CAUSE GBH W/I TO DO GBH
COUNSEL/PROCURE ACT OF	APPLY CORROSIVE FLUID W/I

FEMALE GENITAL MUTILATION OUTSIDE UK	
MURDER-VICTIM 1 YR OLD OR OVER	ATTEMPT MURDER-VICTIM 1 YR OLD OR OVER
RACIALLY/RELIGIOUSLY AGGRAVATED S47 ASSAULT AND MALICIOUS WOUNDING	ADMINISTER NOXIOUS THING TO INFLICT GBH
THROW CORROSIVE FLUID W/I	ADMINISTER POISON W/I
KIDNAPPING	ATTEMPTED MALICIOUS OR UNLAWFUL WOUNDING
EXCISE/INFIBULATE/OTHERWISE MUTILATE FEMALE GENITALIA	ADMINISTER NOXIOUS THING W/I
MURDER VICTIM UNDER 1 YR OLD	BURGLARY W/I INFLICT GBH DWELLING
INFANTICIDE	ATTEMPT TO WOUND W/I TO DO GBH
DO ACT W/I CAUSE EXPLOSION ENDANGER LIFE	CAUSE NOXIOUS THING TO BE TAKEN W/I
RELIGIOUSLY AGGRAVATED MALICIOUS WOUNDING	CAUSE NOXIOUS THING ADMINIST INFLICT GBH
DO ACT W/I TO CAUSE EXPLOSION - ENDANGER LIFE OTH BUILDING	CONSPIRACY TO KIDNAP
INFLICT GBH	ATTEMPT WOUND W/I RESIST/PREVENT ARREST
POLICE - WOUNDING WITH INTENT TO RESIST/PREVENT ARREST	CAUSE/ALLOW CHILD/VULNERABLE PERSON TO SUFFER SERIOUS PHYSICAL HARM

Appendix – Modelling Details

Exploratory (spatial) data analysis shows that there is correlation between MSV crimes both over space and time. This appendix analyses this correlation and outlines the approach taken to the predictive modelling.

Whilst the majority of incidents occur within and around the main 4 city centres, there are incidents occurring outside the central areas and their surrounds. Despite this, when examined as a spatial point pattern (on a monthly basis), the MSV crime incidents do not follow complete spatial randomness (CSR)[5]:

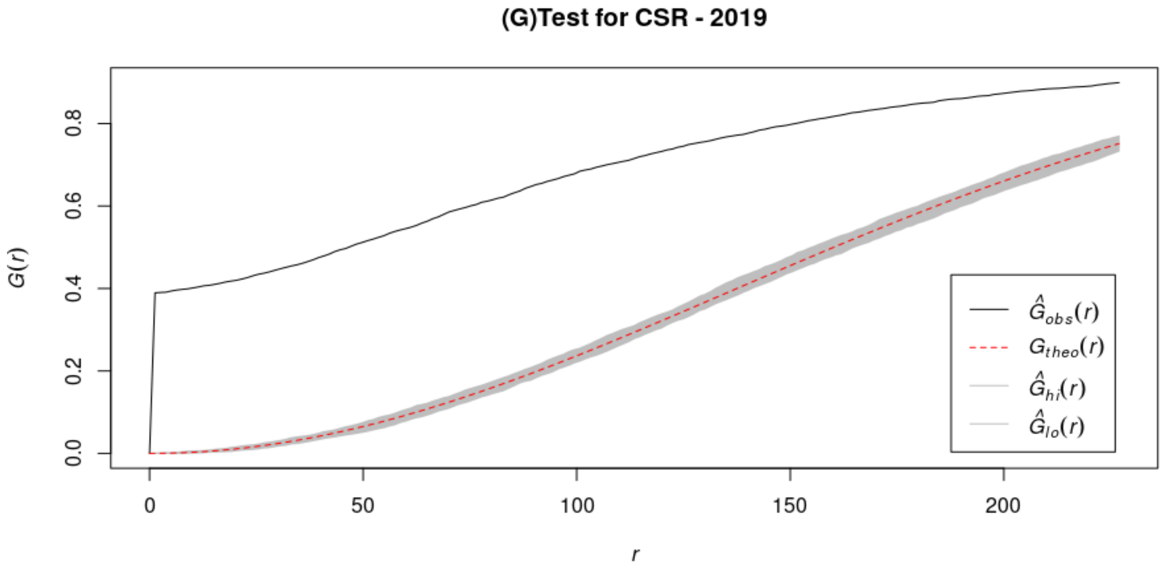

Figure 8: testing for CSR.

NOTE: this chart shows the result of testing of the locational patterns of MSV crimes against a hypothesis of CSR; the black line being outside of the grey boundaries shows that there is clustering and the patterns of MSV crime incidents do not follow CSR.

Due to identifiable areas being most useful operationally, for the purposes of this project the WMP area has been split into a grid with each grid being circa 0.5 square km. It is on this gridded pattern (and now therefore areal, as opposed to point pattern, data) that the modelling has been undertaken.

The figure below shows that when MSV crimes are examined for the period March 2020 – March 2021, there is a small degree of global spatial autocorrelation:

[5] This is the case whether incidents inside the city centres are excluded or not and whether or not it is assumed that there is spatial non-stationarity.

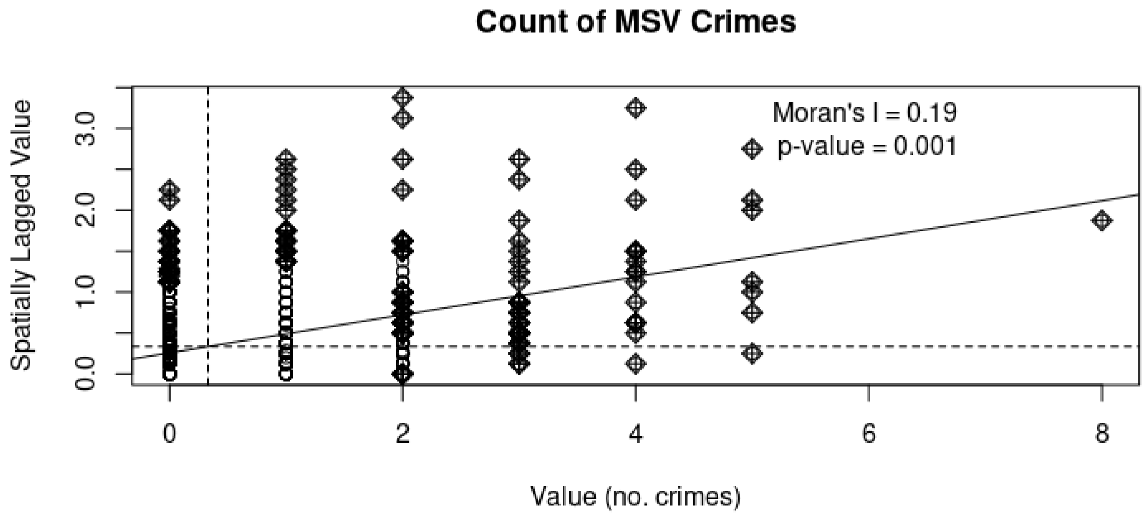

Figure 9: global spatial autocorrelation

A more localised analysis shows the presence of spatial clusters[6]:

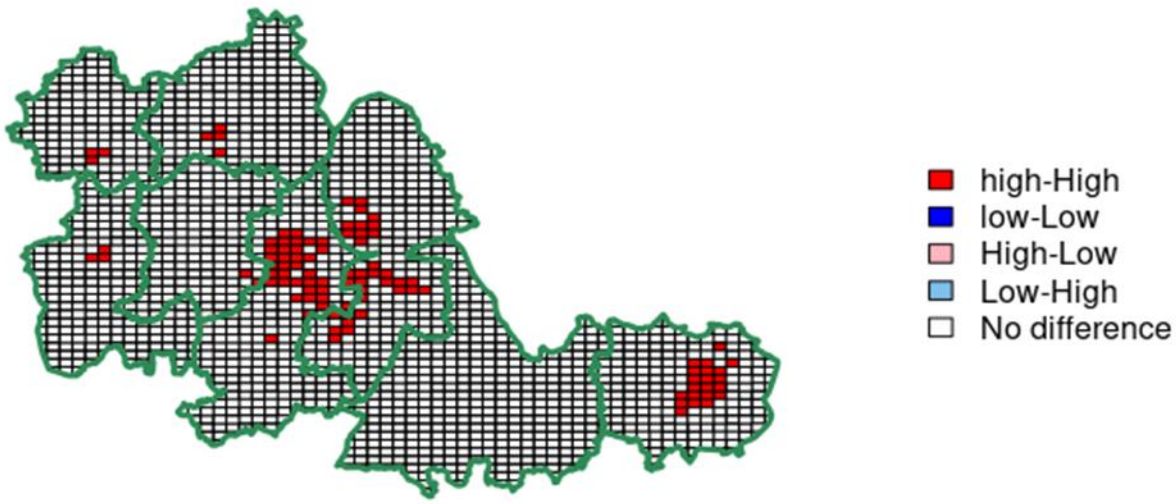

Figure 10: cluster map of local indicator of spatial association (2020 – 2021)

On the time dimension, it can be seen from the chart below that prior to 2009 there is a major structural change which would most likely be attributable to changes in recording practices:

[6] It should be noted that these are locations of likely clusters, not MSV crimes *per se*. Derived following calculation of a local indicator of spatial association statistic (see Anselin 1995). In this instance there are only clusters of relatively high numbers of MSV crimes.

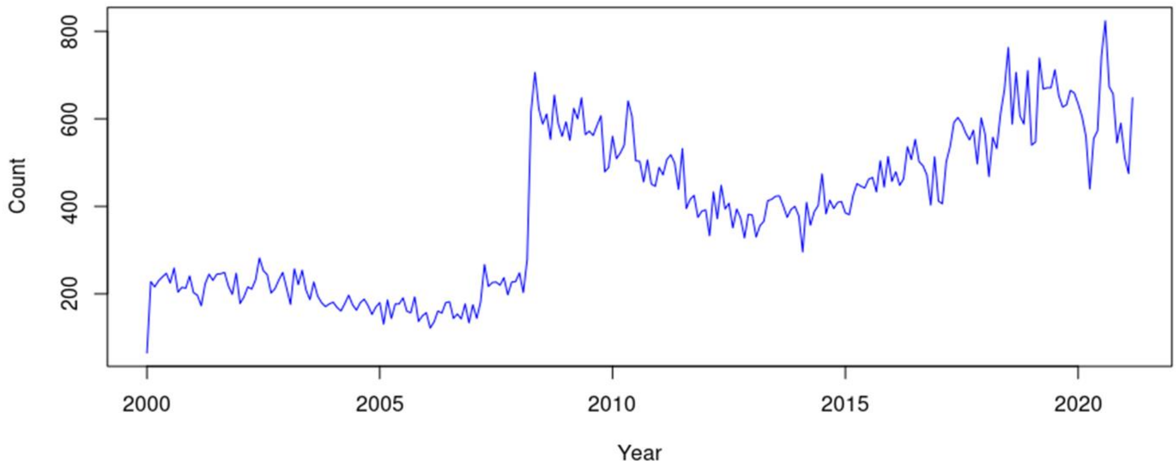

Figure 11: time series of MSV crimes for 20 years

Because of this data have been used from 2009 onwards only.

Looking at MSV crimes as a time series there would appear to be a degree of seasonality amongst MSV crimes, but of greater importance is an apparent long-term autocorrelation in the trend:

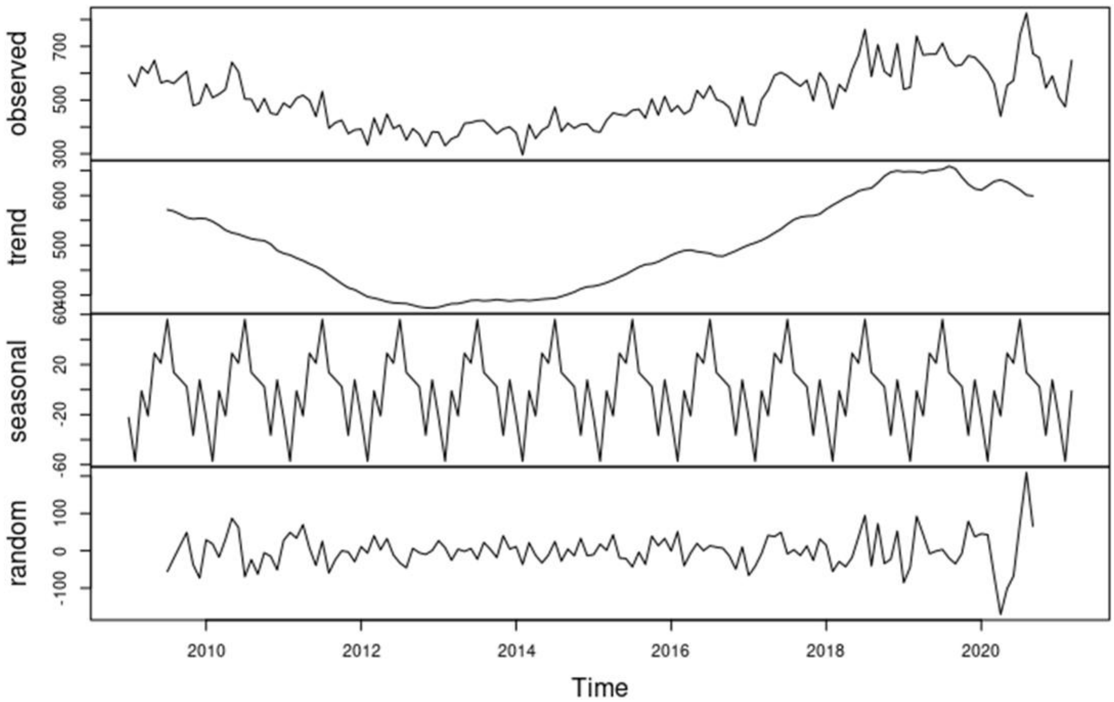

Figure 12: decomposition of MSV crime time series

This long-term autocorrelation is confirmed in the ACF and PACF charts below:

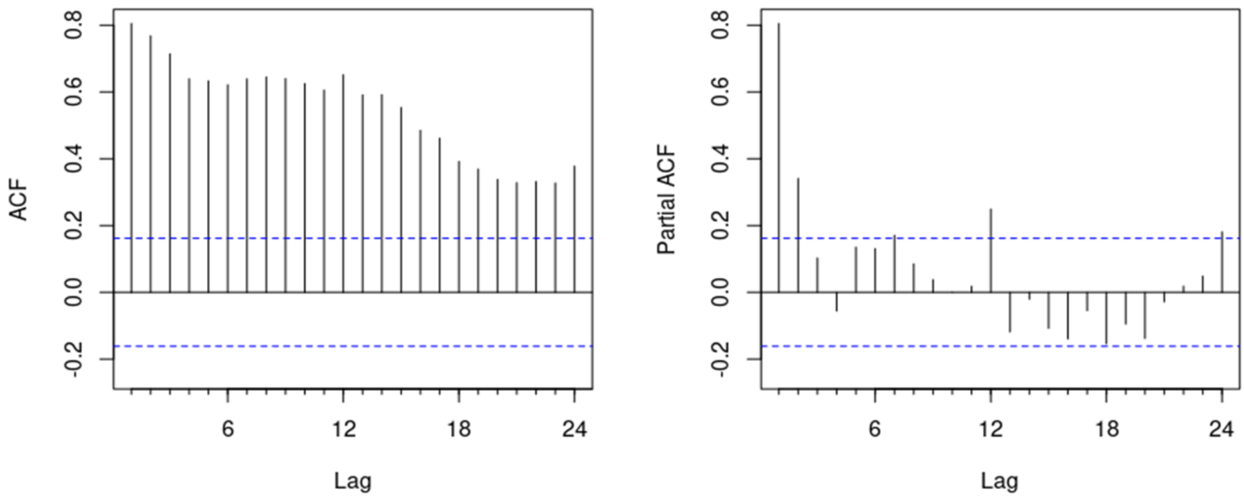

Figure 13: autocorrelation and partial autocorrelation functions of MSV crime (lags on the x axis, correlation on the y axis)

Figure 14 below shows the relative score of a space time scan statistic. This shows that there are locations that exhibit higher than expected counts of MSV incidents and that there are clusters of such areas.

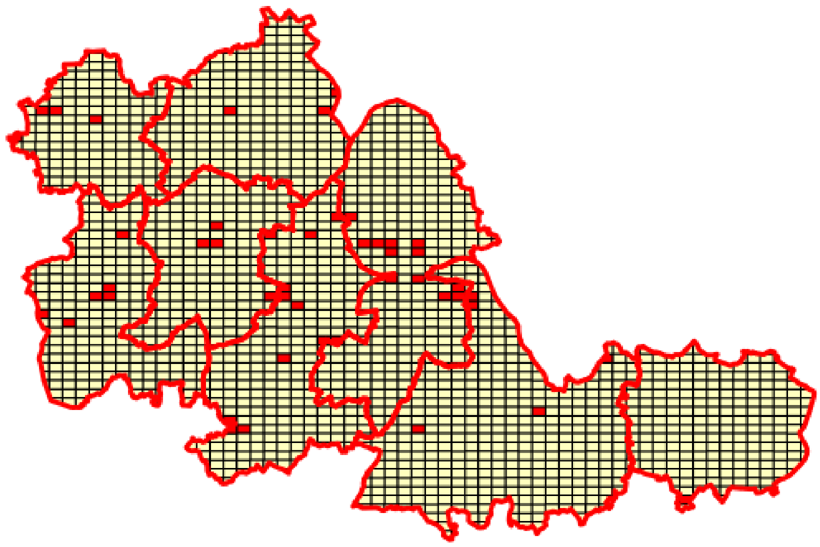

Figure 14: map of relative score of an empirical Bayes scan statistic

This points towards any relevant model potentially being space-time inseparable[7] (i.e. some form of interaction between space and time will likely be useful within any such model), however this is borderline and a space-time separable model could well prove preferable.

[7] Haining, R. and Guangquan, L., 2020.

From the above, there is information in both the temporal and spatial dimensions that can be useful for making predictions of MSV crime. Therefore, the approach taken here is essentially to use previous occurrences of MSV crime to predict (over a coming 4 week period) future MSV crime. This is akin to using a time series to predict future levels of that time series (an often used methodology). The data used are the count of 4 weekly MSV crimes over the previous 2 years.

For additional information the levels of (the previous time period) general crime are also used.

Identifying the best Model

The exploratory spatial data analysis (ESDA) undertaken shows that there is essentially correlation over both time and space of MSV crimes and this therefore naturally leads to the use of a spatio-temporal model. The results of the scan statistic noted in figure 13 are ambiguous regarding using space-time separable or inseparable models[8]. Both separable and inseparable model(s) have been tested and it was found that a separable model is preferable[9].

As a starting point, a number of different models were built to make predictions for one future observation (so one 4 week period over the 1,935 different grid squares).

The different models were:

Model	Explicitly spatial / spatio-temporal
Gradient Boosted Model 1	Yes (spatial)
Gradient Boosted Model 2	No
Neural network 1	No
Neural network 2	Yes (spatial)
Neural network 3	No
Neural network 4	Yes (spatial)
Random forest 1	Yes (temporal)
Random forest 2	Yes (spatial)
Random forest 3	Yes (spatio-temporal)
Conditional autoregressive 1	Yes (spatio-temporal with interaction)
GAM 1	Yes (spatio-temporal)
GAM 2	Yes (spatio-temporal with interaction)
Conditional autoregressive 2	Yes (spatio-temporal with interaction)

As can be seen from the chart below, using the root mean square error (RMSE) shows that the (Bayesian) conditional autoregressive model with a random walk of order 1

[8] A space-time separable model essentially models the data generation process as overall spatial + overall temporal whilst an inseparable model is applied to data generation processes that cannot be fully described by this structure; there is an interaction between space and time.

[9] The different types of models being tested by way of their predictive performance (see later in the appendix).

(noted as conditional autoregressive (CAR) 2 in the able above) produces the lowest (i.e. best) RMSE, but this is closely followed by the 2 GAM models:

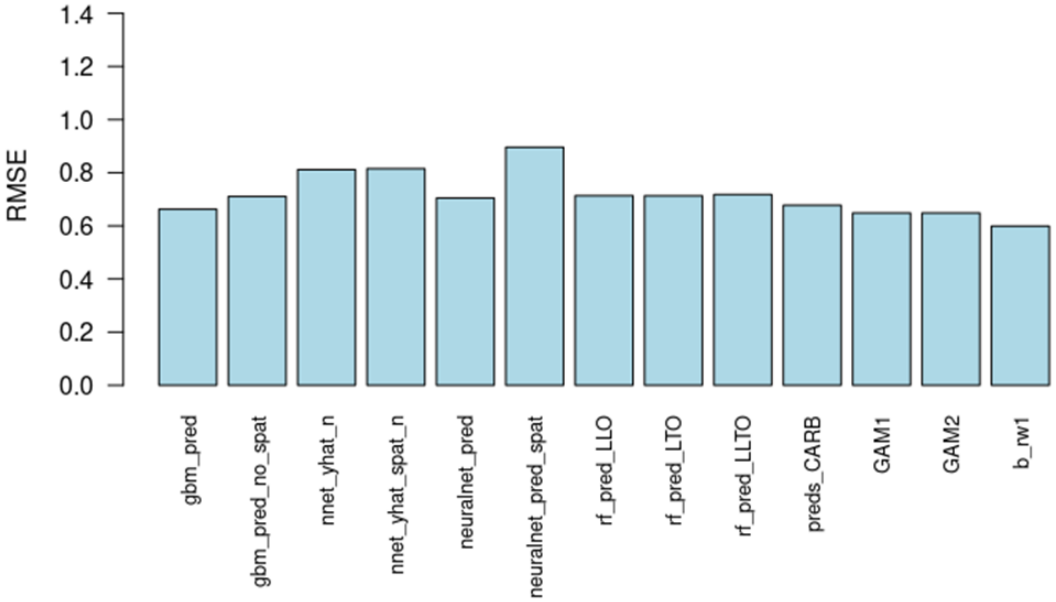

Figure 15: initial models (RMSE calculated over a single prediction period)[10]

The next means of choosing the final model was undertaken by way of using the 2 GAM models and the CAR model with a moving window approach to make predictions over 5 consecutive periods (so 5 four week periods, the predictions for each of which was based on the previous two years' of data which slides along from the beginning of the dataset onwards). Included in this was use of a naïve model (applying the mean) in order to compare the models against a more traditional hotspot approach.

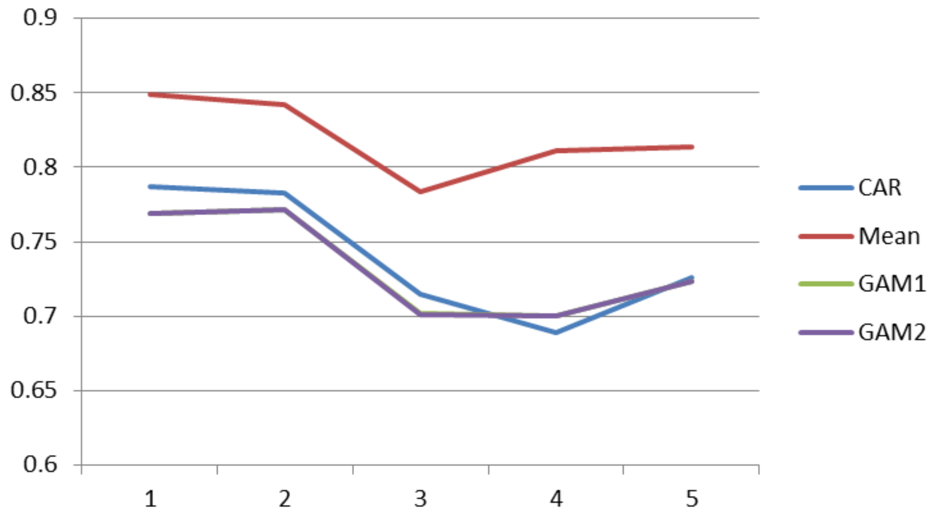

Figure 16: RMSE of final models

---

[10] The models were also assessed by way of a binary something occurred / didn't occur and a predicted is something going to occur / not occur. The result was qualitatively the same conclusion as reached via the RMSE.

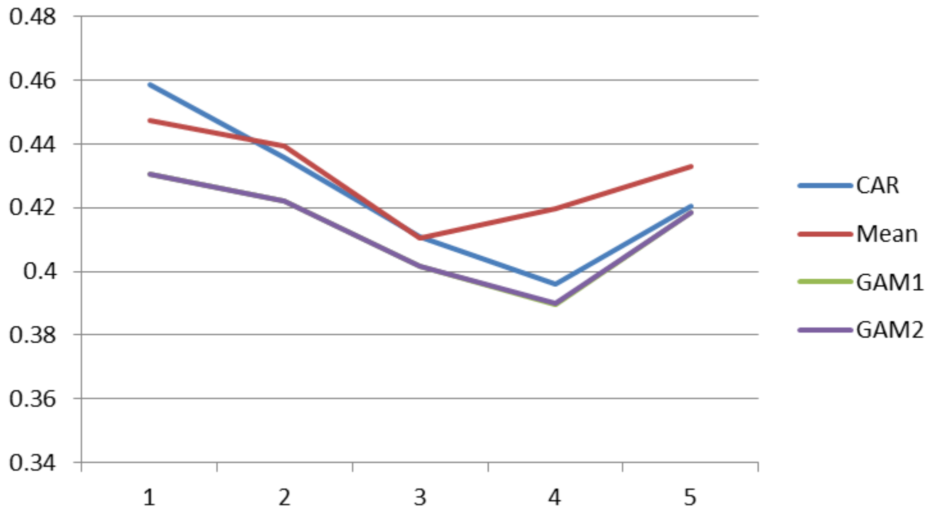

Figure 17: Mean absolute error (MAE) of models

Looking at both the RMSE and the MAE shows that all the models generally produce lower errors than the naïve approach with the two GAM models producing lower errors than the CAR model.

Another way to assess the efficacy of the models is to examine their performance as a binary something occurred / didn't occur process (so whether one or more MSV incidents are predicted in a grid matches up with whether a MSV crime actually occurred in the same grid for the period in question).

The results of this show similar results[11]:

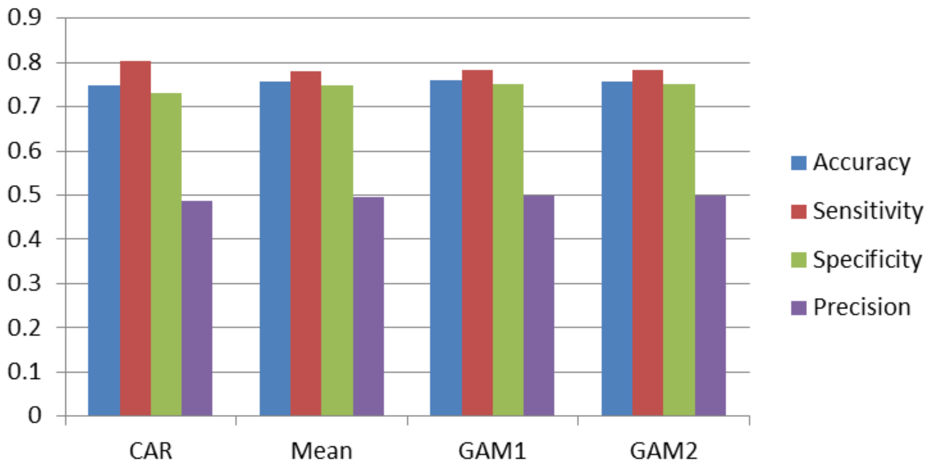

Figure 18: accuracy measures for moving window predictions

[11] For this the optimal cut-off points were ascertained by way of adjusting the true positive rate and the false positive rate to reach the optimal balance between the two. This was also applied to the naïve (mean based) model.

However, these results were also placed within a benefit-cost approach whereby

((Precision * APC) – (False Positive Rate * APC)) * (Sensitivity – 0.5)

where APC = the average PC salary. This measure takes into account the potential for the wasteful use of resources as a high false positive rate would mean that officers would be sent / resources used in areas that were unlikely to see MSV crimes. The last term penalizes the degree to which a model is inaccurate in terms of the true positive rate.

Using this measure the best model is the space-time seperable GAM model:

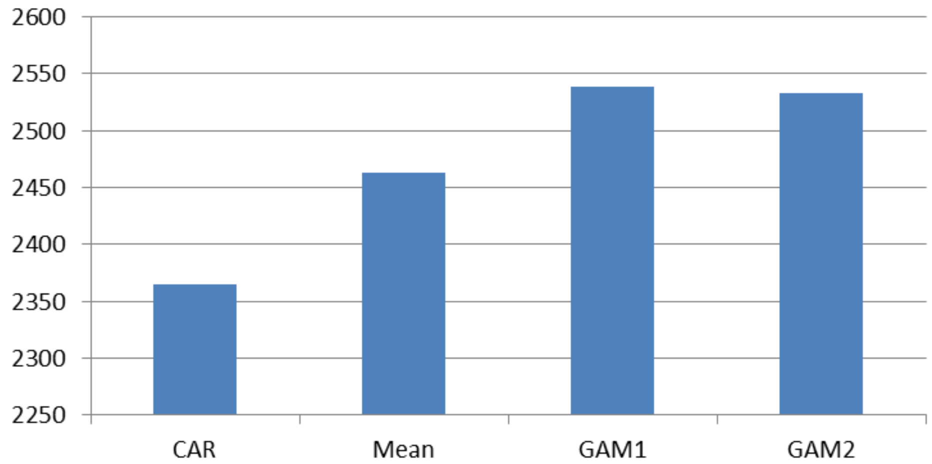

Figure 19: Benefit - Cost of the different models

On this basis, the GAM model is 5% – 12% better than the naïve approach.

The final model is a generalized additive model (GAM) making use of splines (so incorporating non-linearities) through the features (the previous time period's general crime, space and time; so spatial and temporal autocorrelation is taken into account by way of the splines). As the data are count data, the model utlises the negative-binomial family via a log link function:

The general form of a GAM is:

$$E(Y|x_{1,}x_{2,}, \dots, x_p) = \alpha + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p)$$

where Y is the target variable, x_i is a feature (predictor variable) and the f_j's are unspecified smooth functions (which are nonparametric). In general, the conditional mean $\mu(X)$ of the target variable Y is related to an additive function of the features via a link function g so that the model becomes:

$$g[\mu(X)] = \alpha + f_1(x_1) + \cdots + f_p(x_p)$$

In this case, the features are the lag of general crimes and the spatial and temporal patterns of MSV crimes.

Using this model, the output would be (for the first period of the moving window):

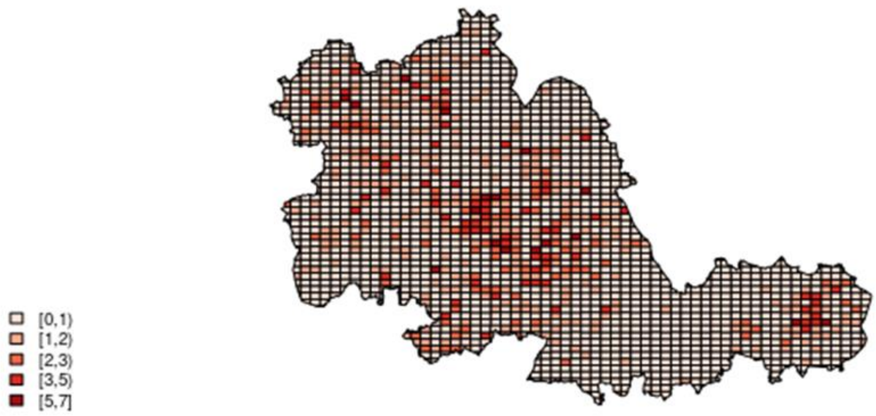

Figure 20: actual MSV crimes month 1

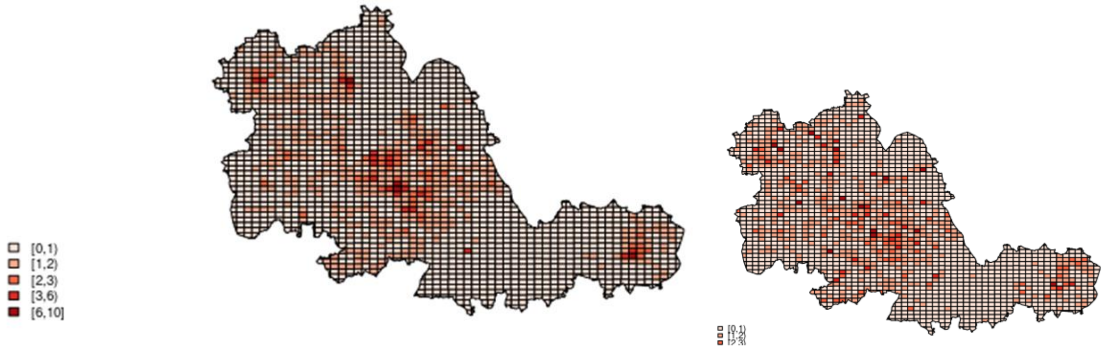

Figure 21: predictions from the GAM (left) and residuals (right)

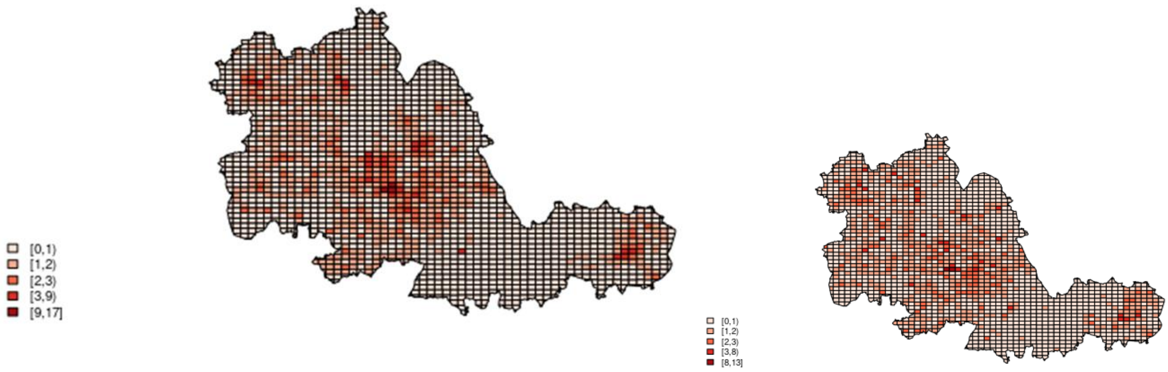

Figure 22: using the naive (mean-based) approach (left) and residuals (right)

References

Anselin, L., 1995, Local Indicators of Spatial Association – LISA, *Geographical Analysis*, Vol. 27(2), pp. 93 – 115.

Haining, R. and Guangquan, L., 2020, *Modelling Spatial and Spatial-Temporal Data A Bayesian Approach*, CRC Press, Boca Rotan, Fl.