Predicting Violent Crime

Data Analytics Lab

December 2020

This report presents the methodology used to produce a model with the aim of predicting the number of violent crimes within the WMP area. The report also includes a discussion of findings from exploratory analysis.

1 Table of Contents

2 Introduction

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 knife 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.

To this end, a statistical model has been trained to predict areas that are likely to contain violent crime (not including domestic abuse) in the next 4-week period (which is the tasking cycle). This is done by splitting the WMP region into a grid of 40x40 rectangles (each rectangle is approximately 1 km^2) and using data from the Crimes system to count the number of violent crimes in 4-week periods previously. The model is then trained on data from 02/01/2017 to 20/04/2020, with data from 18/05/2020 to 07/09/2020 kept for evaluation[3]. The model is then retrained on the entire available data and is then used to predict forward five 4-week periods for each area.

The outputs would be put into a Qlik dashboard for visualisation.

[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

[3] The Coronavirus pandemic has, and may continue to, have an impact on the number of violent crimes. The model may produce worse predictions in situations like these where there is a sudden change to the data generating process. However, over a period of time it will adjust to the new short-term trend.

3 Data

The data used includes:

- **Crimes**: The date, location and offence committed, from 2015 until the present day.

The Crimes are filtered by offence to only include 'violent crimes causing injury' as classified by a list of offence codes, previously agreed with a subject matter expert. See the identifying violent crimes section for a list of offences included, note that it does not include domestic abuse or sexual offences.

A grid is then constructed of the WMP area by dividing the width into 40 rectangles and the height into 40 rectangles, producing a total of 898 1.16 km^2 rectangles covering the entire area. The crimes are then mapped to each rectangle and then aggregated to 4-week periods. The grid size and aggregate time periods were discussed with subject matter experts to optimise model accuracy and operational usability. The final dataset consists of the following attributes.

Column	Description
idx	The unique ID of the grid square.
npu_code	The neighbourhood policing unit code.
four_week	The ID of the 4-week period.
four_week_start	The date of the first day of the 4-week period. The first day of a 4-week period is always a Monday.
num_crimes	The number of violent crimes in the grid rectangle in the specified 4-week period.

Data dictionary for model data

A sample of the described data is given below.

idx	npu_code	four_week	four_week_start	num_crimes
1	SH	1	2015-01-05	0
1	SH	2	2015-02-02	0
1	SH	3	2015-03-02	0
1	SH	4	2015-03-30	0
1	SH	5	2015-04-27	0
1	SH	6	2015-05-25	0

Sample of model data

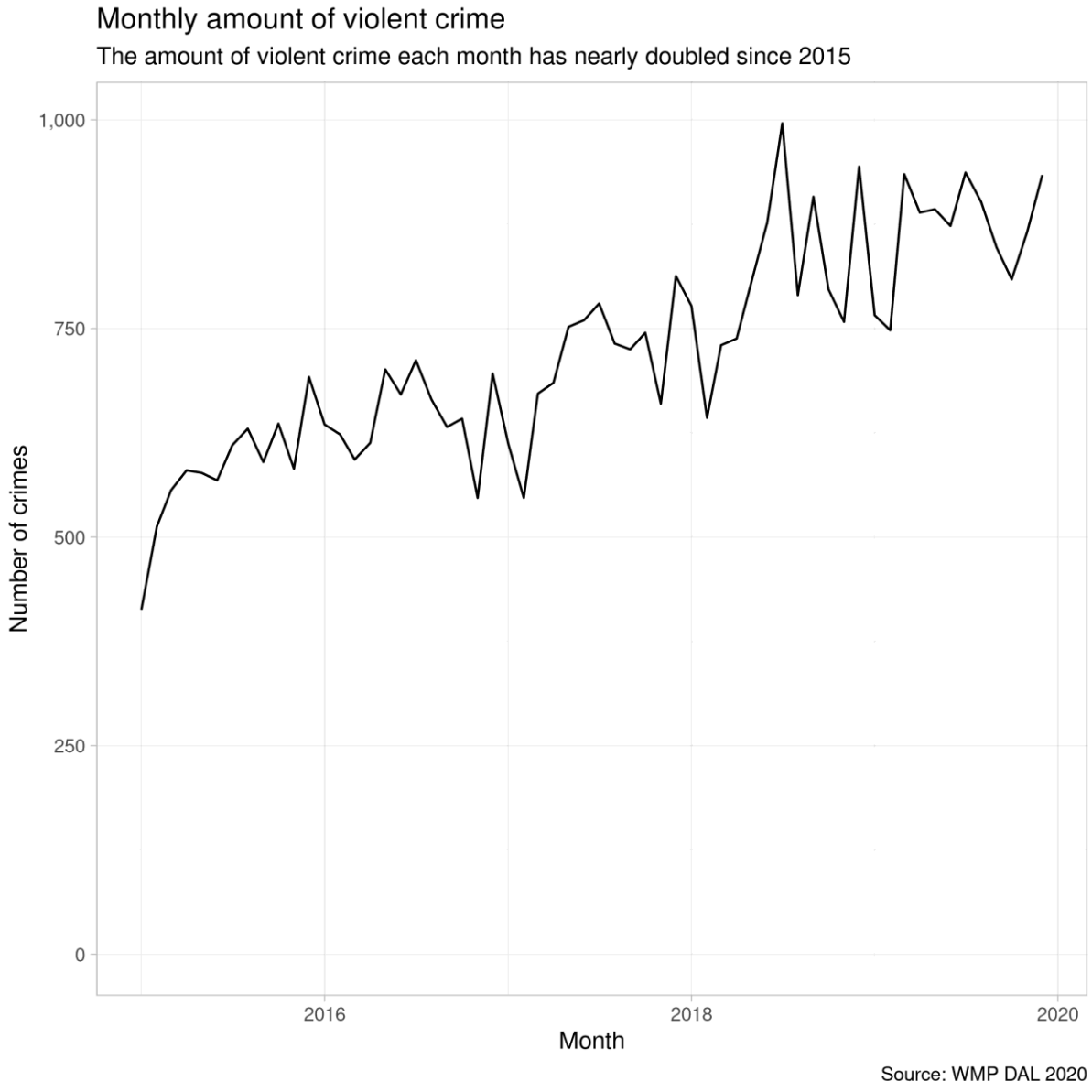
4 Data Exploration

Before a model is built, exploratory analysis is used to provide insights which will inform the model development.

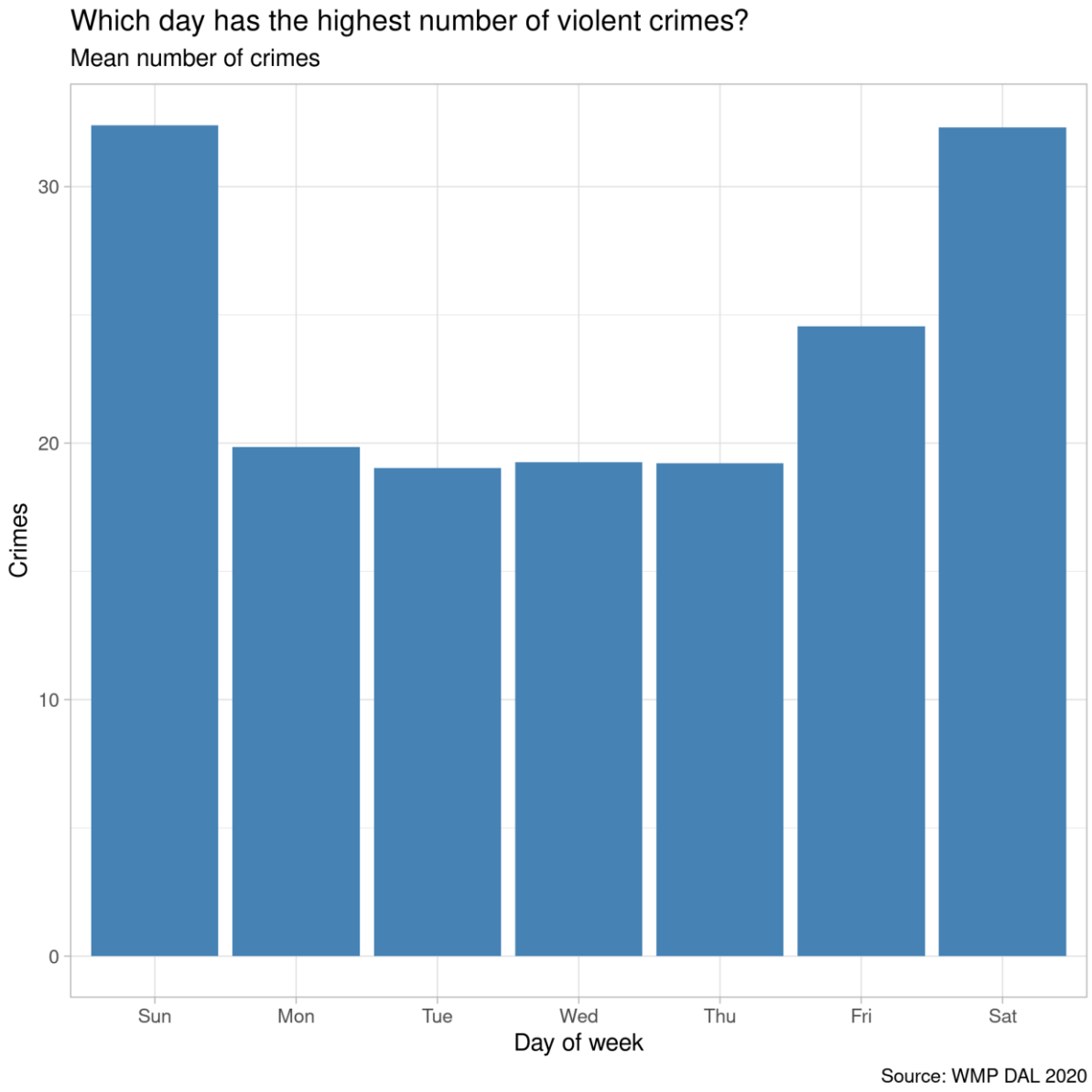
4.1 Temporal

First, temporal exploratory analysis is completed to see how the number of violent crimes vary over the years, months and days of the week.

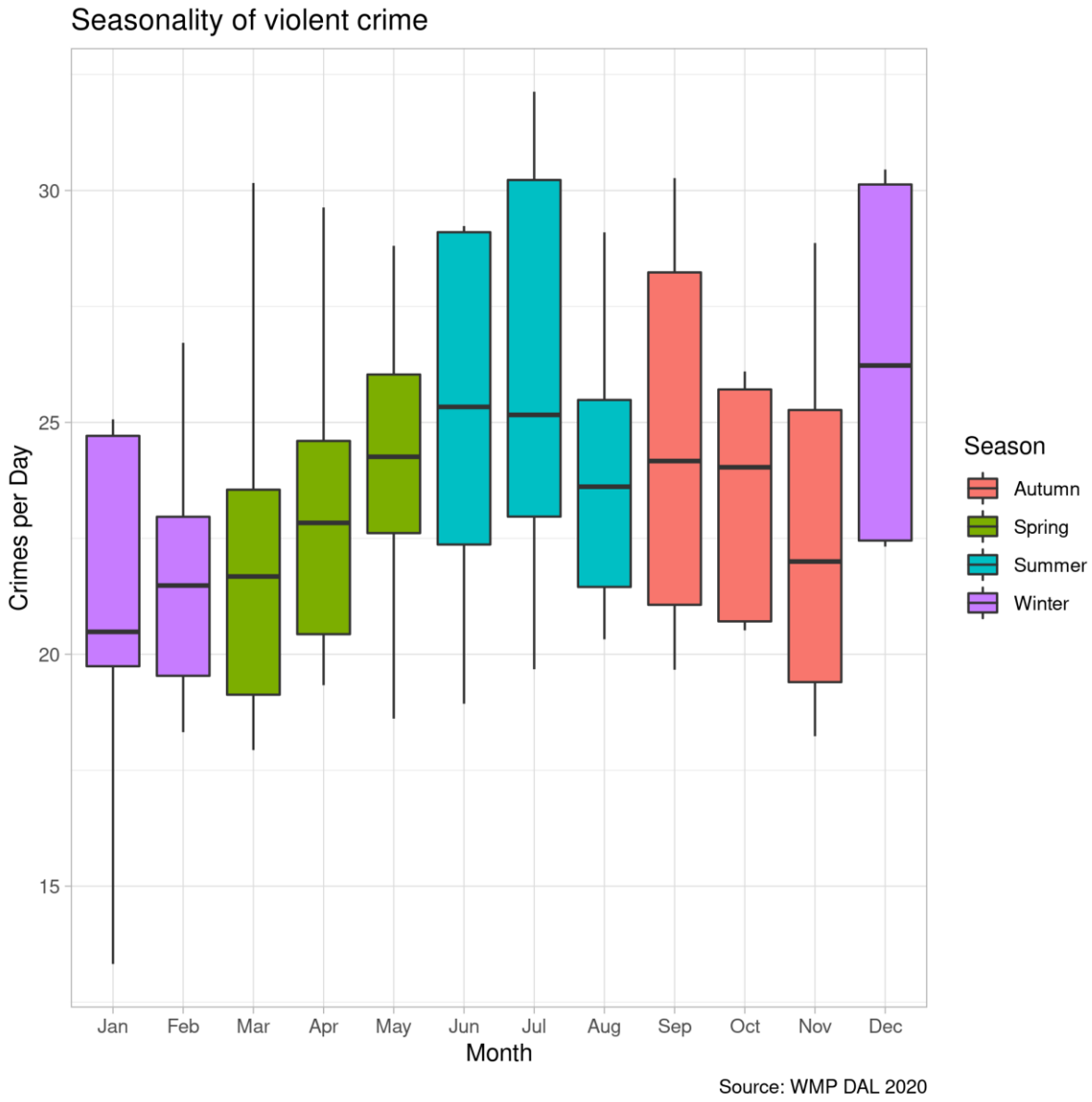
The filtered crimes have then been aggregated to give the total number of violent crimes in each month, in each year. This data is plotted below and shows that the number of violent crimes per month has steadily increased from 2015. This suggests that there is a temporal trend, with potentially some seasonality within the year.



Next, the mean number of violent crimes per day of the week is calculated and is plotted below. The plot shows that the number of violent crimes is higher on Fridays, Saturdays and Sundays. The model predicts over consistent 4-week periods, starting on a Monday, if this were to change to an inconsistent time period, say monthly, then the number of weekend days may need to be taken into account.



The mean number of violent crimes per day in each month is calculated and shown below in a boxplot, coloured by the season of each month. It shows that the number of violent crimes per day is higher during the summer months as well as during December.

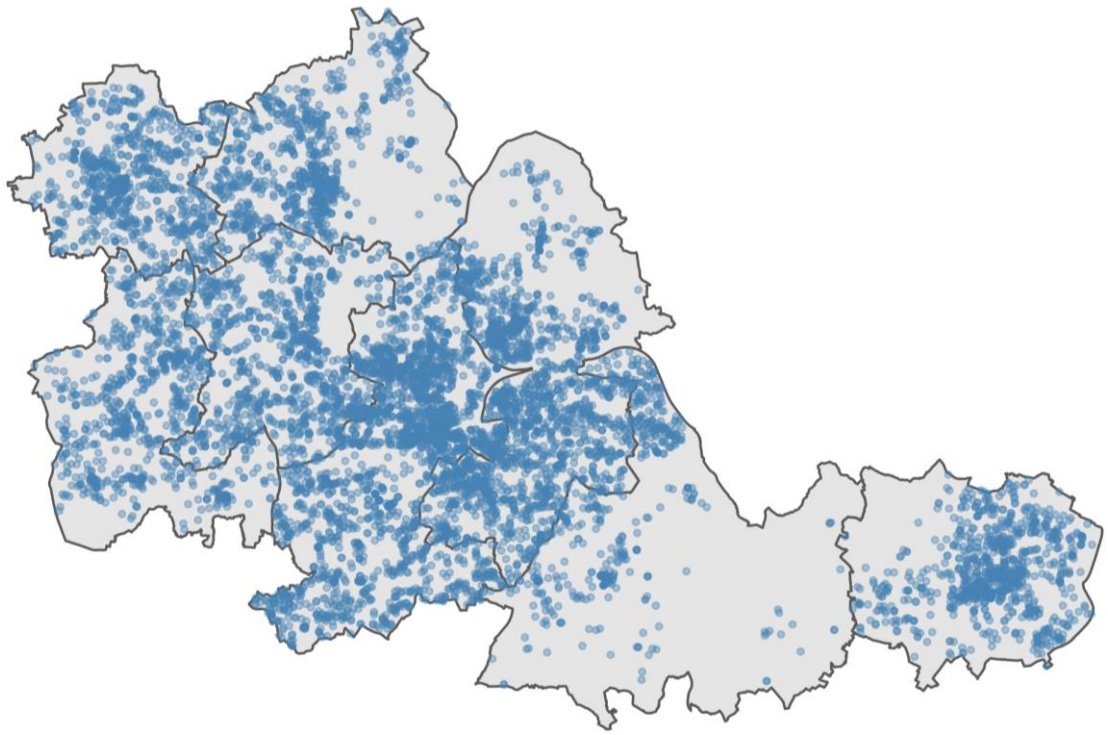


4.2 Spatial

Next, spatial exploratory analysis is completed to see how violent crimes are distributed across the WMP area.

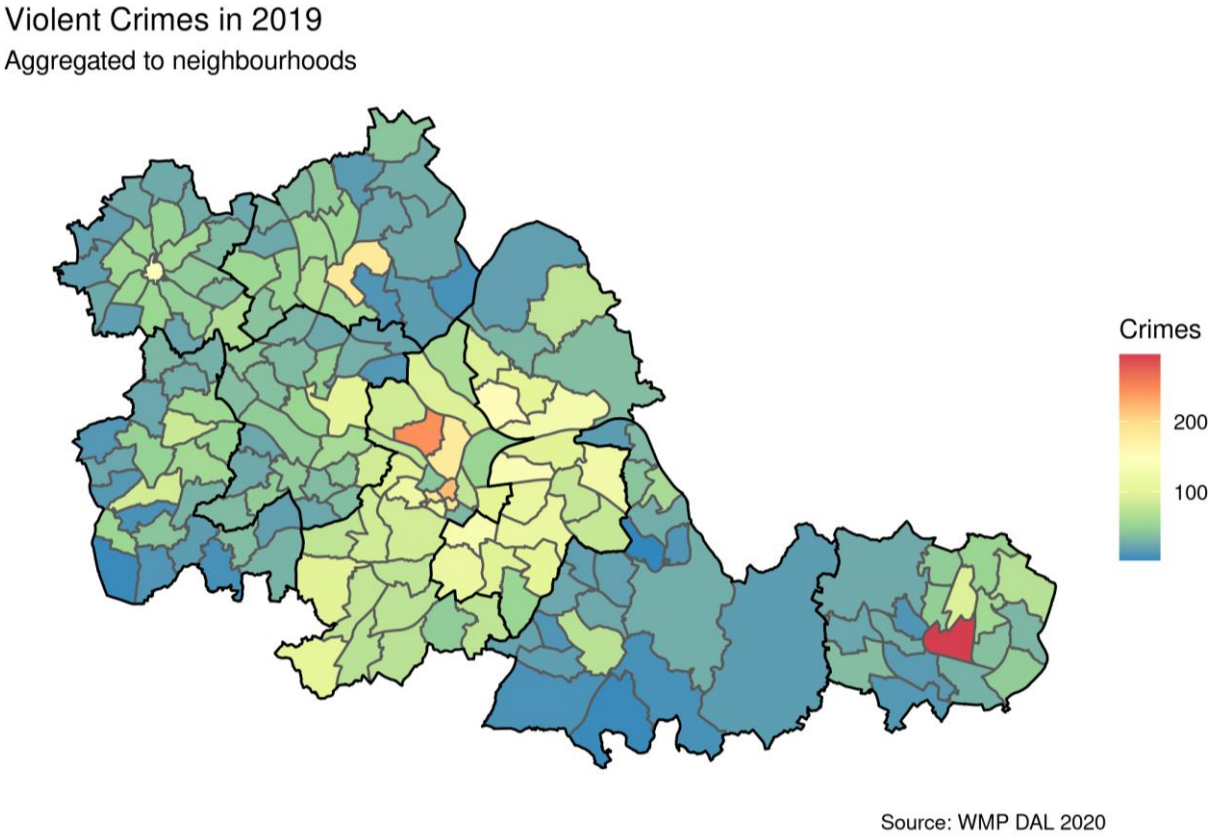
The locations of violent crimes in 2019 are shown below. The plot suggests that crimes are not randomly distributed across the WMP region, and that there is some other factor or process that is contributing to the location of the crimes. See the complete spatial randomness section for a more robust assessment of the potential for spatial randomness.

Locations of violent crimes in 2019

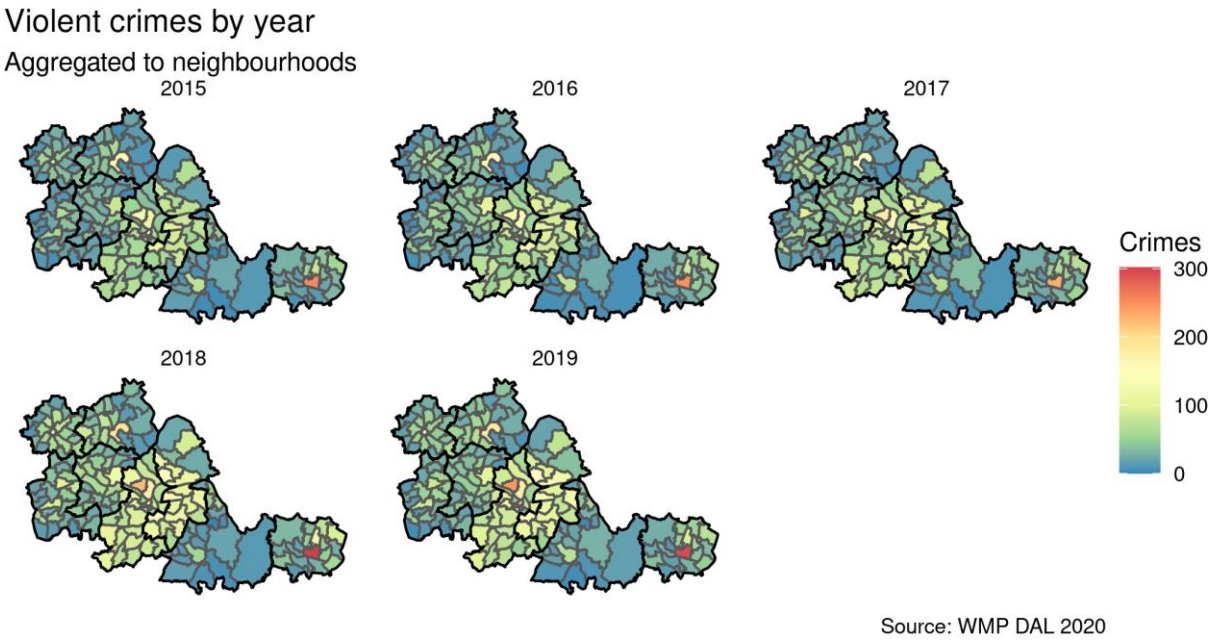


Source: WMP DAL 2020

When the violent crimes in 2019 are aggregated to policing neighbourhoods, the city centres are highlighted as areas with more violent crimes. In the chart below, warmer colours (red, orange and yellow) represent areas with more violent crimes, while colder colours (blue and green) represent areas with fewer violent crimes.



When the above plot is repeated for each year, a similar pattern emerges. Within each year, the city centres are highlighted as areas with more violent crime. The number of violent crimes in Birmingham city centre appears to have increased, particularly over the last two years. Also, the number of violent crimes in the other neighbourhoods in the Birmingham East and Birmingham West NPUs appear to have increased.



In summary, there appears to be temporal effects here which differ for each area.

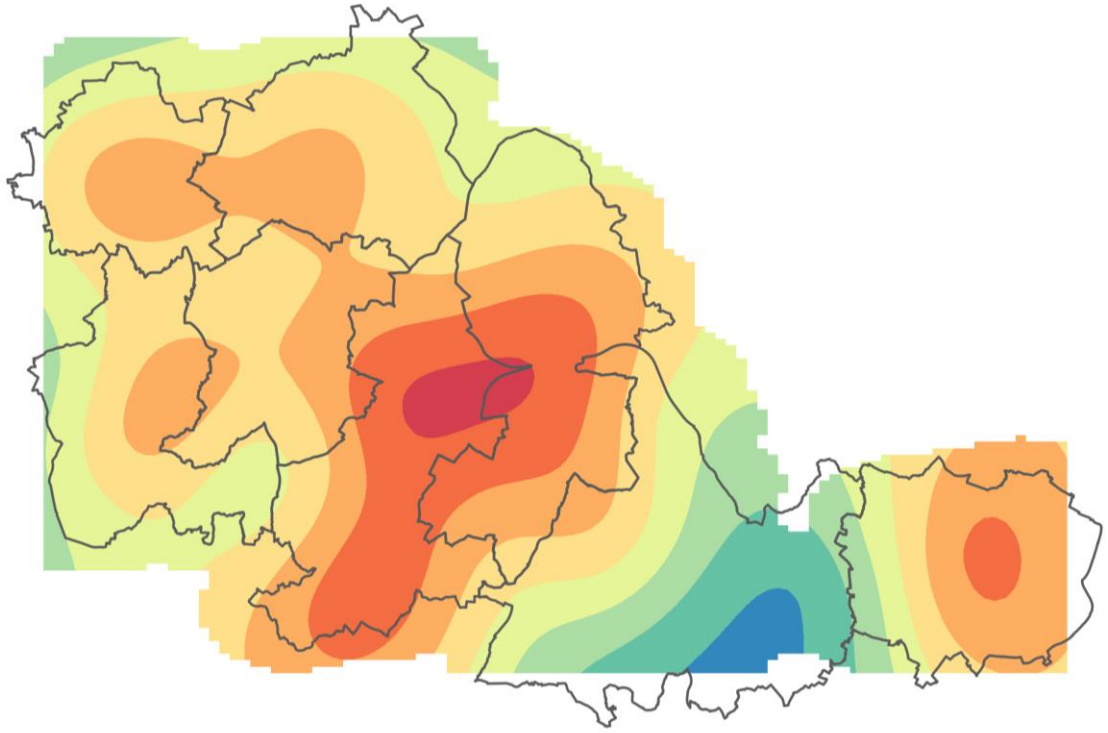
5 Model Building

5.1 Development

The results of one approach of spatial smoothing[4] are superimposed over the NPU boundaries below. The results tell a similar story to the previous plots, that city centre areas have a higher number of violent crimes, with a particularly low number of violent crimes in the rural, east side of Solihull.

Spatial smooth estimate

Red/orange areas contain more violent crime, green/blue areas contain less



Source: WMP DAL 2020

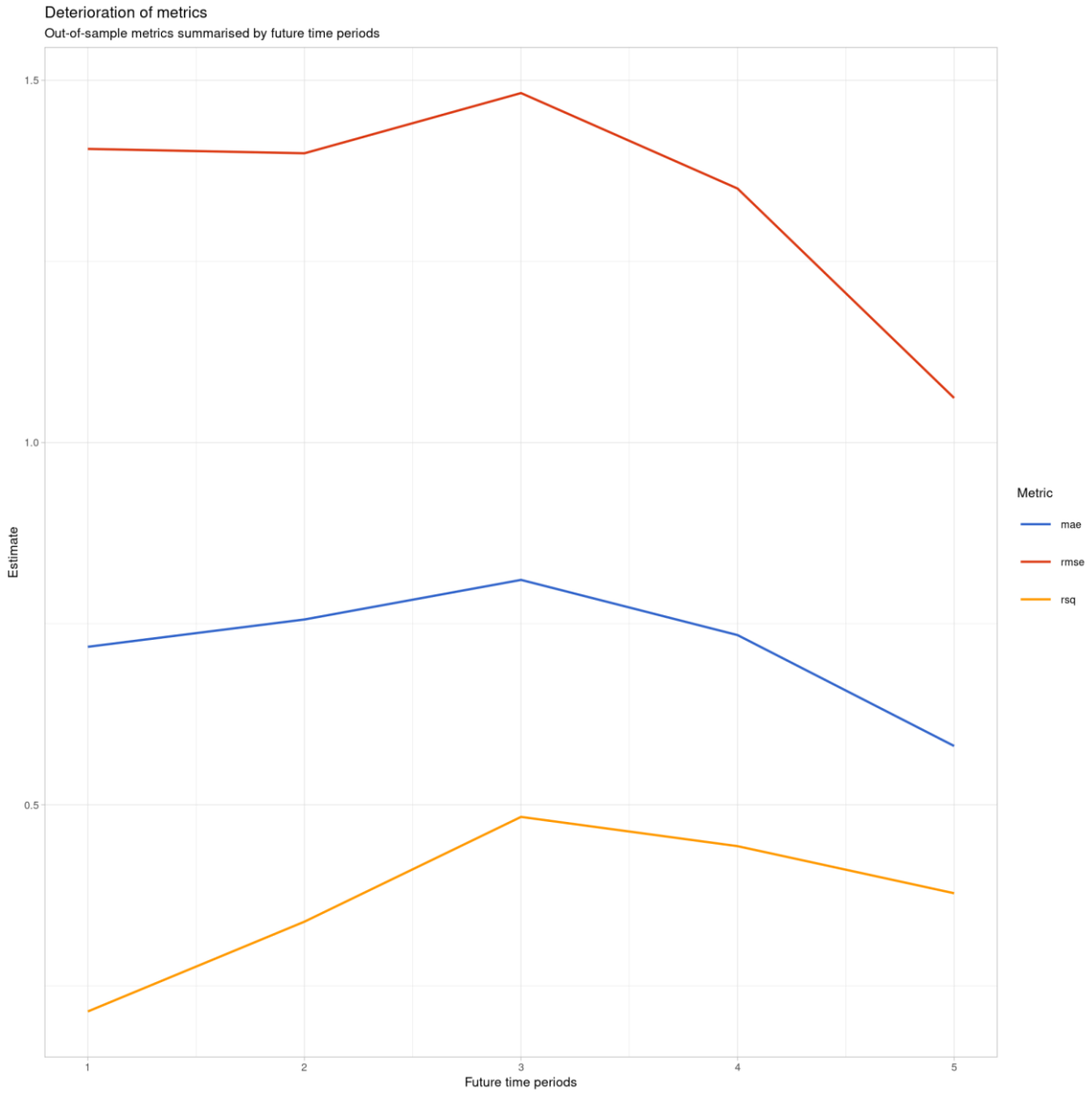
Space-time interactions were also analysed similar to the manner as described in (Blangiardo and Cameletti 2015). The conclusion of this analysis is that there is a temporal structure which is independent from the ones of neighbouring areas.

[4] An isotropic thin-plate spline is fit to the point pattern data to infer the empirical spatial distribution of violent crimes.

5.2 Evaluation

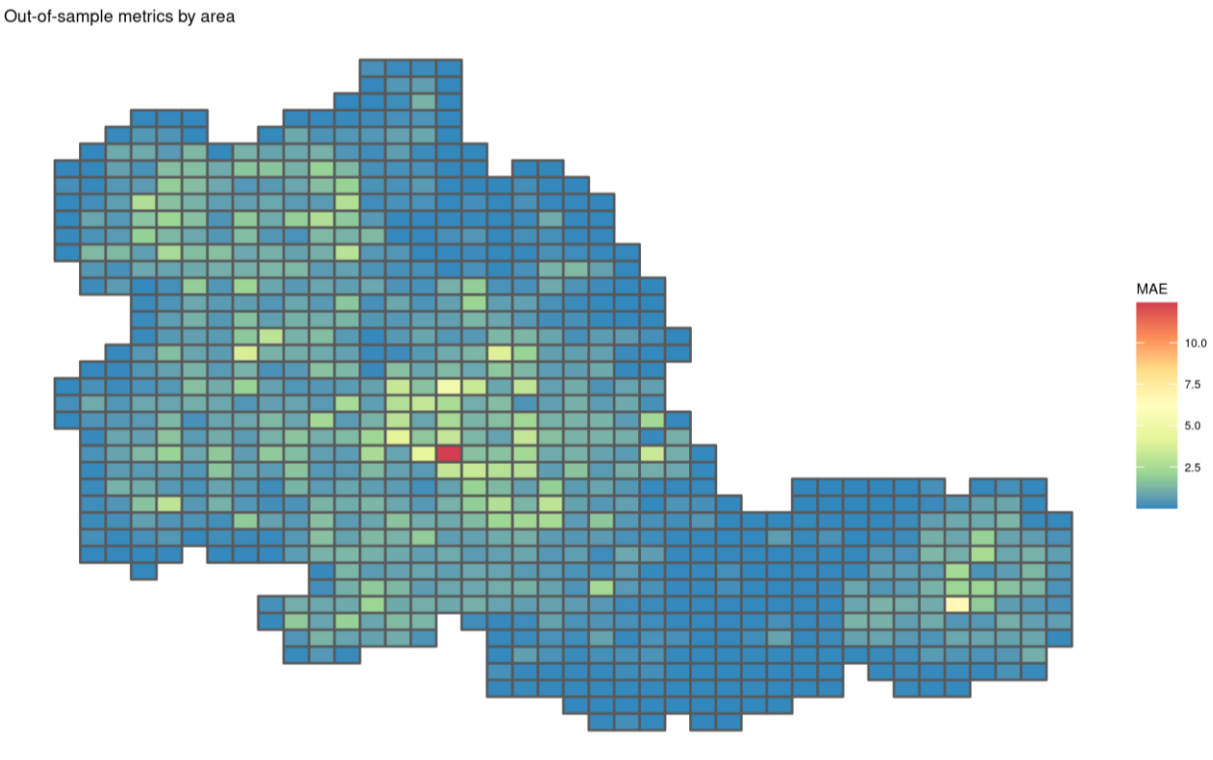
After the model has been trained on data from 02/01/2017 to 20/04/2020, the model is then used to make predictions for data from 18/05/2020 to 07/09/2020. These predictions are then compared to the observed number of violent crimes to evaluate the model. The model evaluation comprises of several model evaluation charts.

The first plot shows how the mean average error (mae), root mean squared error (rmse) and r-squared (rsq) vary over the future predictions. Each metric is calculated across the whole area for each separate 4-week future period. In particular, this graphic shows the slight improvement of model performance at five future periods. In this case, one future period is one 4-week period in the future, compared to what the model has been trained with.

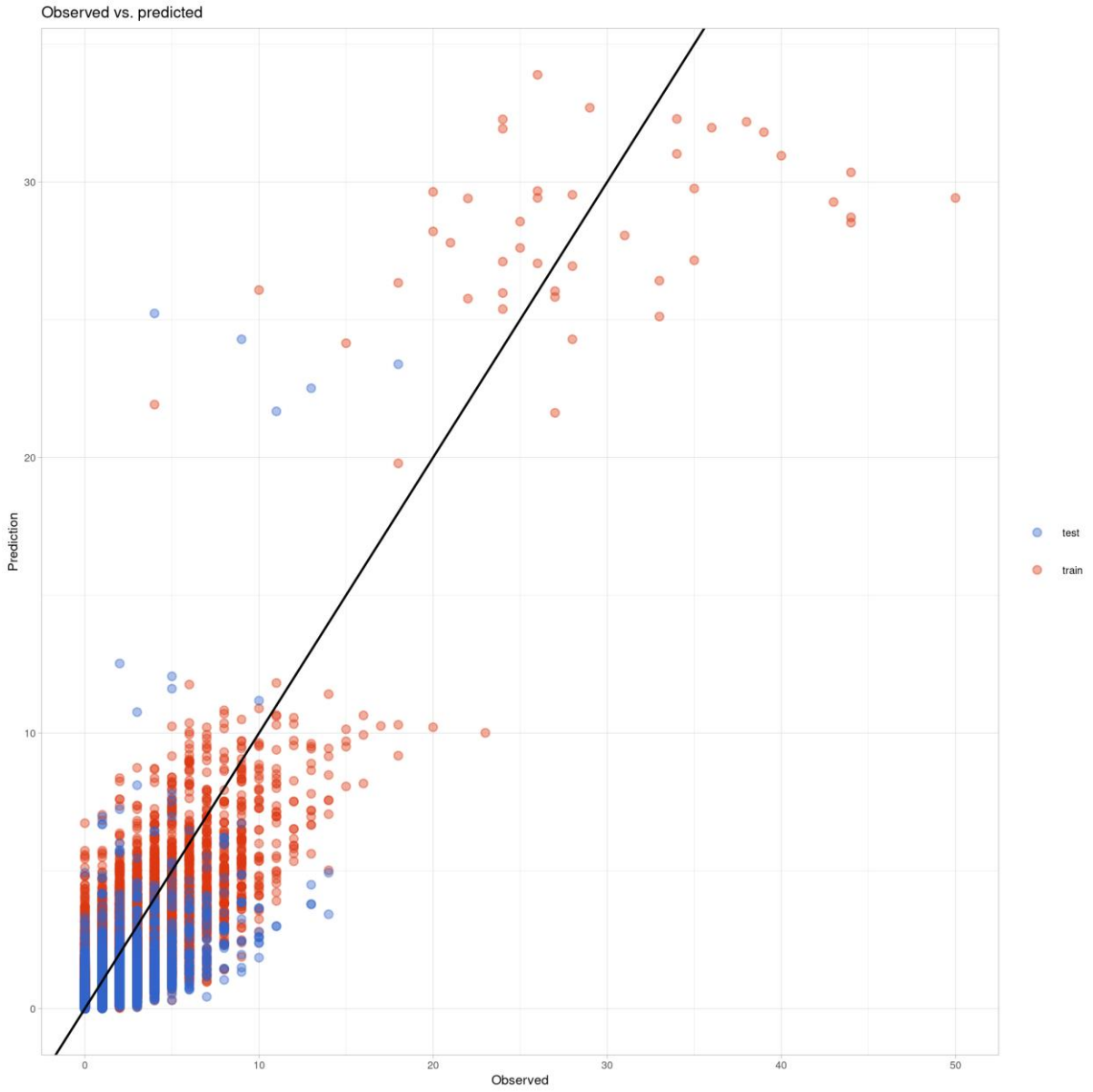


The second plot shows how the mean average error (mae) varies for each area over all five 4-week future periods. It highlights how the model performs well in most Solihull

areas, where there are generally fewer violent crimes, but not as well in areas such as Birmingham city centre where there is a higher variance.



The below plot shows the predictions against the observations for both the in-sample (train) and out-of-sample (test) observations with a reference line for perfect predictions (when the prediction equals the observation). The model appears to understand the main cluster of points well, although it perhaps under-predicts slightly. However, it is not able to predict the cluster of points with larger numbers of crimes with the same accuracy. These points are rarer because they typically have a larger number of violent crimes.



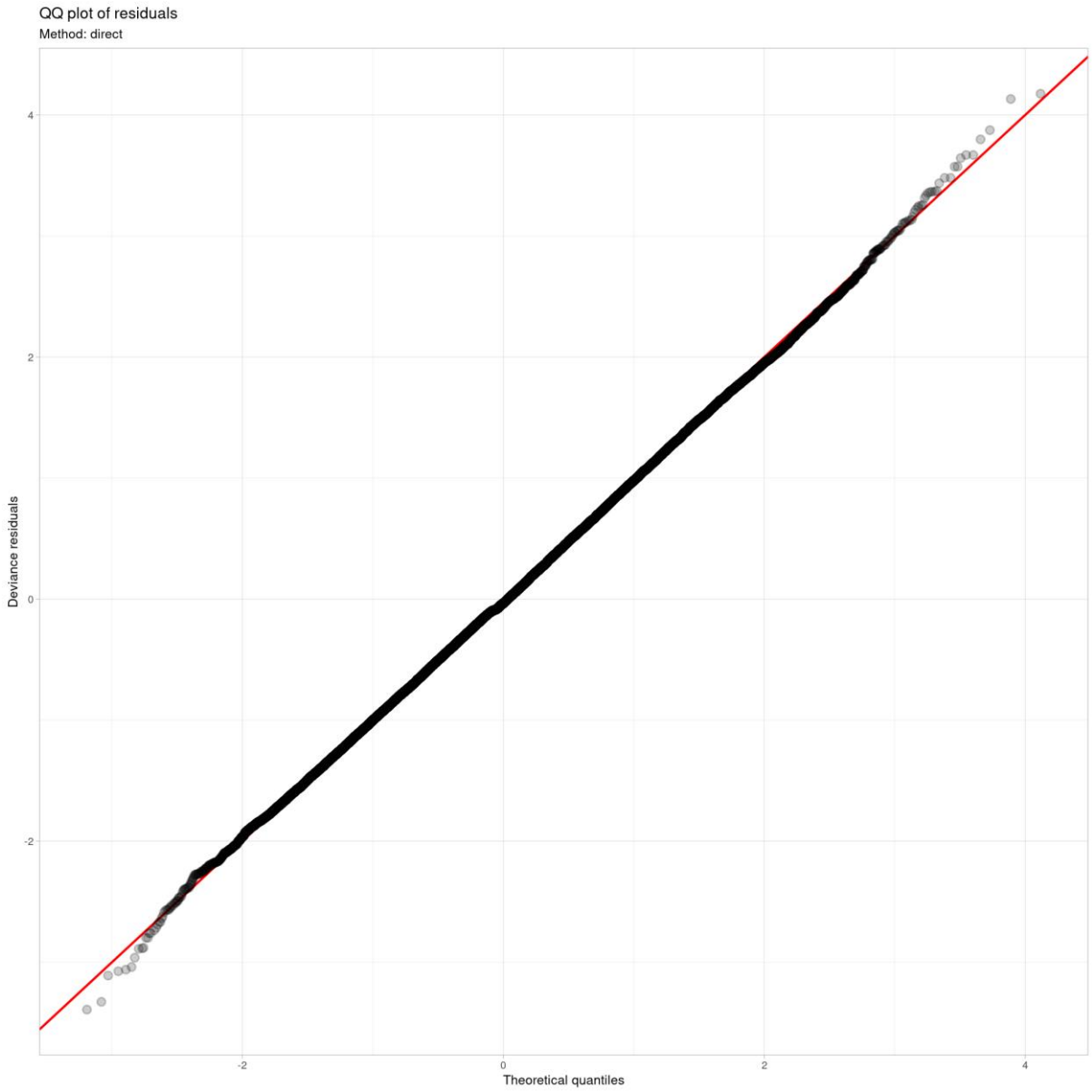
The table below summarises the mean absolute error (mae), root mean squared error (rmse), r-squared (rsq) and mean absolute percentage error (mape) for both the in-sample (train) and out-of-sample (test) observations. This table shows how the model is slightly worse at predicting the unseen data (test) compared to the seen data (train); this is expected.

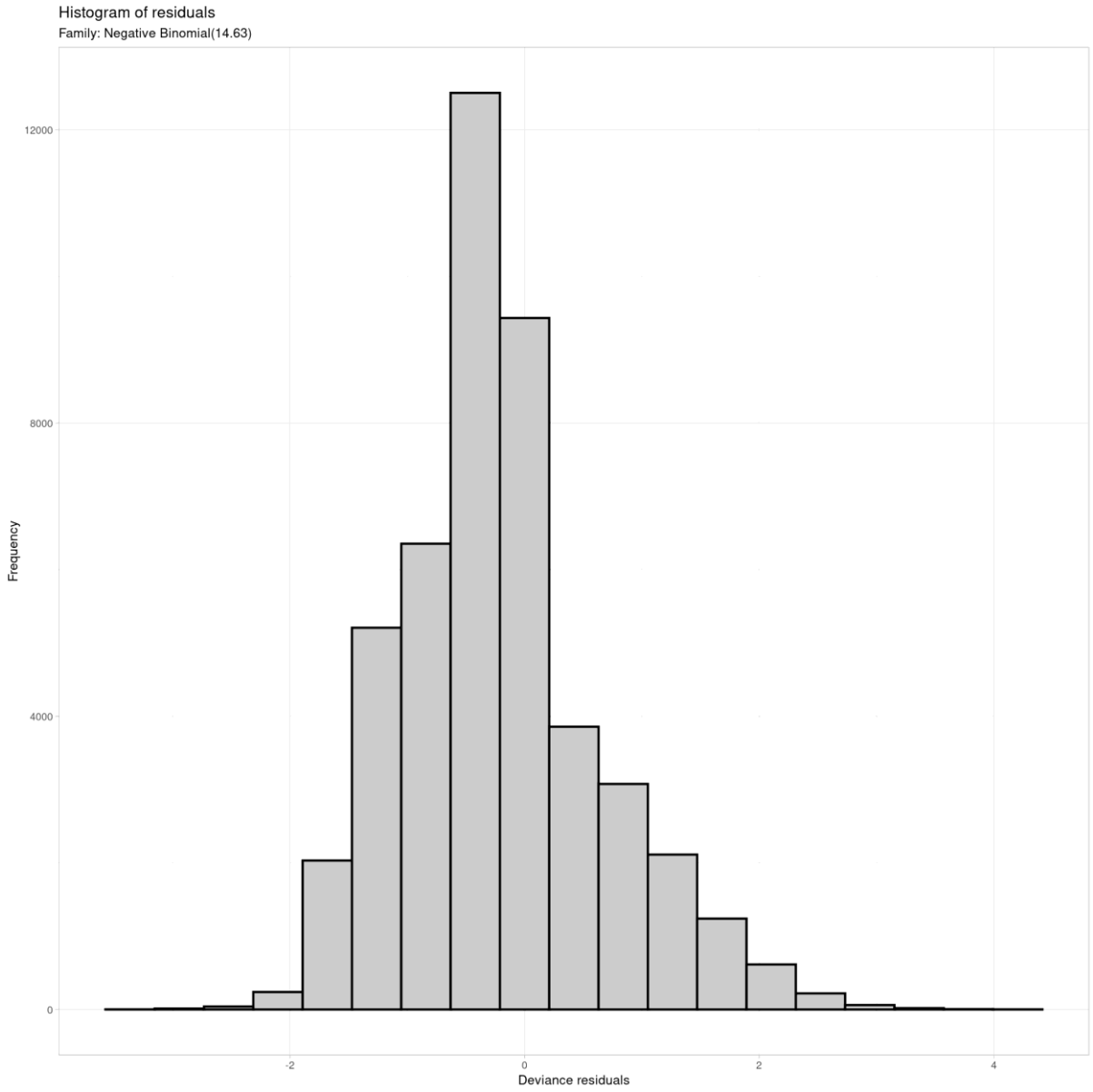
metric	train	test
rmse	0.9939230	1.3476544
mae	0.5971024	0.7199091
rsq	0.6618274	0.3601430
mape	47.4259174	56.3757136

Model metrics

The following plots are diagnostic plots, relating to the residuals (the difference between the predictions and observed values) in the in-sample data. These plots show

no areas for concern and that the model manages to capture the variance of the data well.





The model will be retrained every time a new set of predictions is made, every four weeks. This will allow the model to pick up changes in short term trends and respond to changes more quickly.

6 Conclusion

In conclusion, the WMP area has been split into a 40x40 grid and a model has been trained to predict the number of violent crimes in each square area over a 4-week period. This model takes into account temporal changes as well spatial differences, with each area effectively having its own temporal trend. The model appears to capture the variance of the data well, even at this level of granularity.

The output of the model will be put into a Qlik dashboard for visualisation. This will enable the Project Guardian team to see which localised areas are more likely to experience violent offending in the forthcoming four week period.

Whilst the main focus of Project Guardian and the VRU is long-term problem solving, this tool provides an indication of any emerging trends outside of the usual focus areas, which may require a policing response.

The predictive nature of this analysis means that rather than reacting to violent events after they have occurred, the team can take short term preventative measures by working with local neighbourhood policing teams, licensing teams and offender management teams. Tactics would include visible policing patrols and engagement with local educational establishments and partners.

7 Appendix

7.1 Identifying Violent Crimes

A list of offence codes from previous projects have been agreed with subject matter experts to include only 'violent crime' (not including domestic abuse). The same definition is used here and can be altered at a later date if deemed necessary. The current list is as follows.

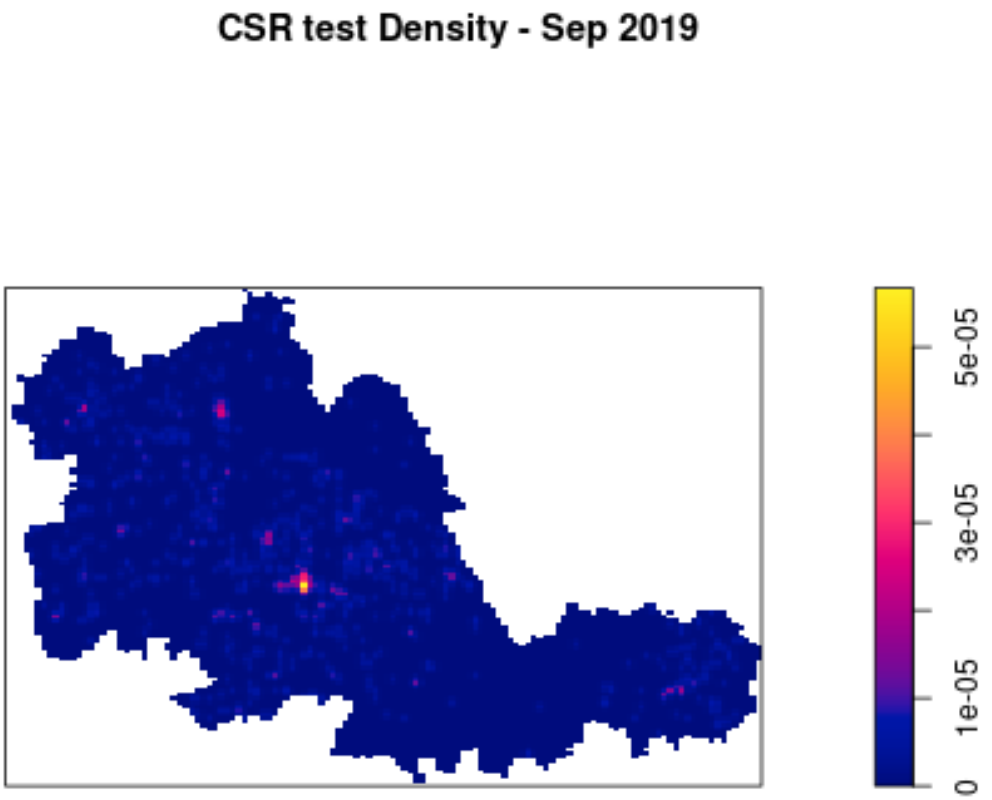
Offences	
RACIALLY AGGRAVATED MALICIOUS WOUNDING	ATTEMPT TO INFLICT GBH
WOUND W/I TO RESIST/PREVENT ARREST	CAUSE NOXIOUS THING TAKEN-ENDANGER LIFE
RACIALLY AGGRAVATED WOUNDING S.20	CAUSE POISON TO BE TAKEN-ENDANGER LIFE
POLICE - ATTEMPT TO CAUSE S.18 GBH WITH INTENT TO DO GBH	CAUSE GBH WITH INTENT
RELIGIOUSLY 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

Offences	
COUNSEL/PROCURE ACT OF FEMALE GENITAL MUTILATION OUTSIDE UK	APPLY CORROSIVE FLUID W/I
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

7.2 Complete Spatial Randomness

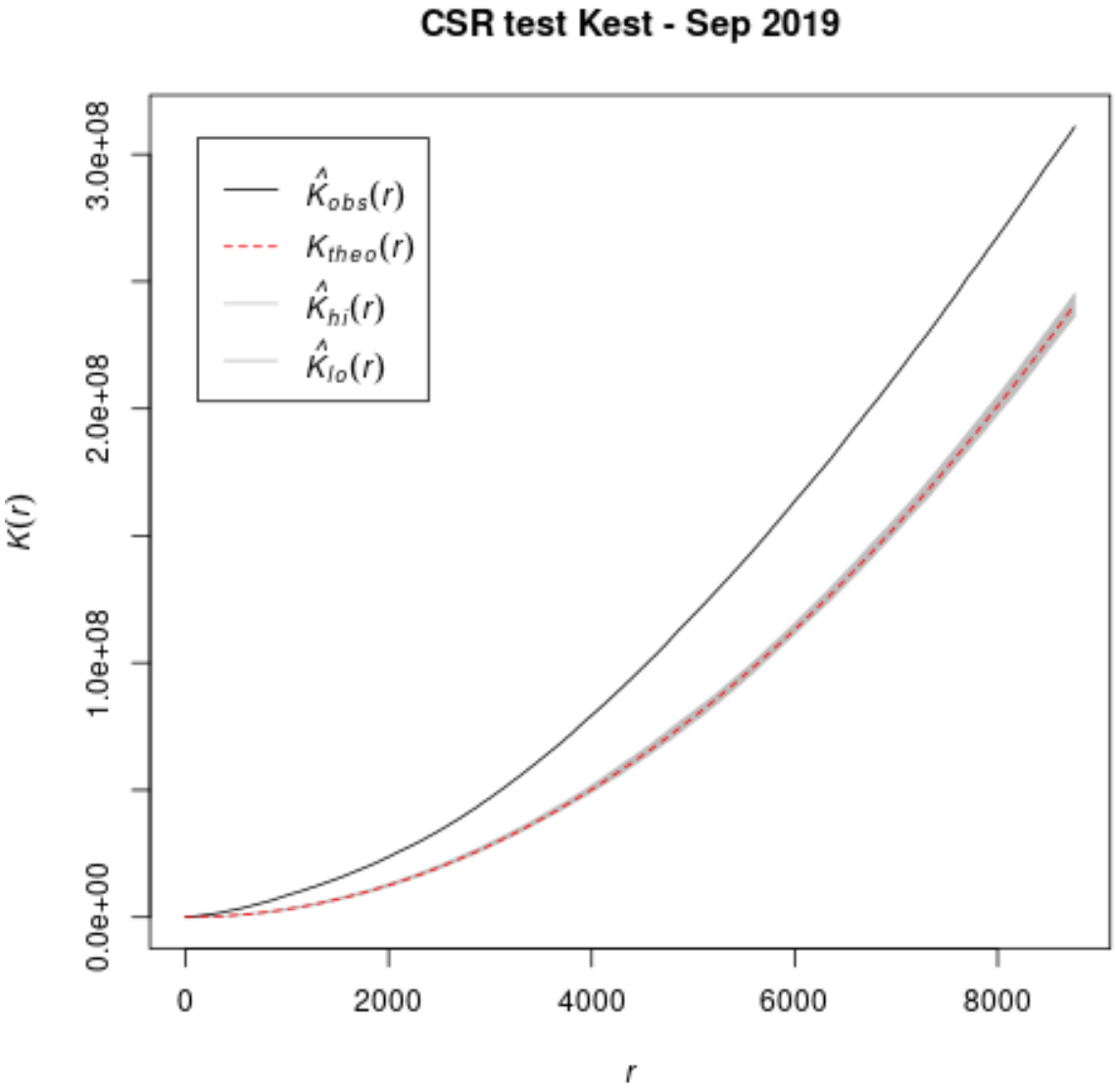
A point pattern, the location of violent crimes, is said to follow complete spatial randomness (CSR) if the points occur randomly within a study area.

The below density plot highlights the city centres as well as some surrounding inner-city areas as having a higher density of violent crimes. This is further evidence that violent crimes are not spatially random.

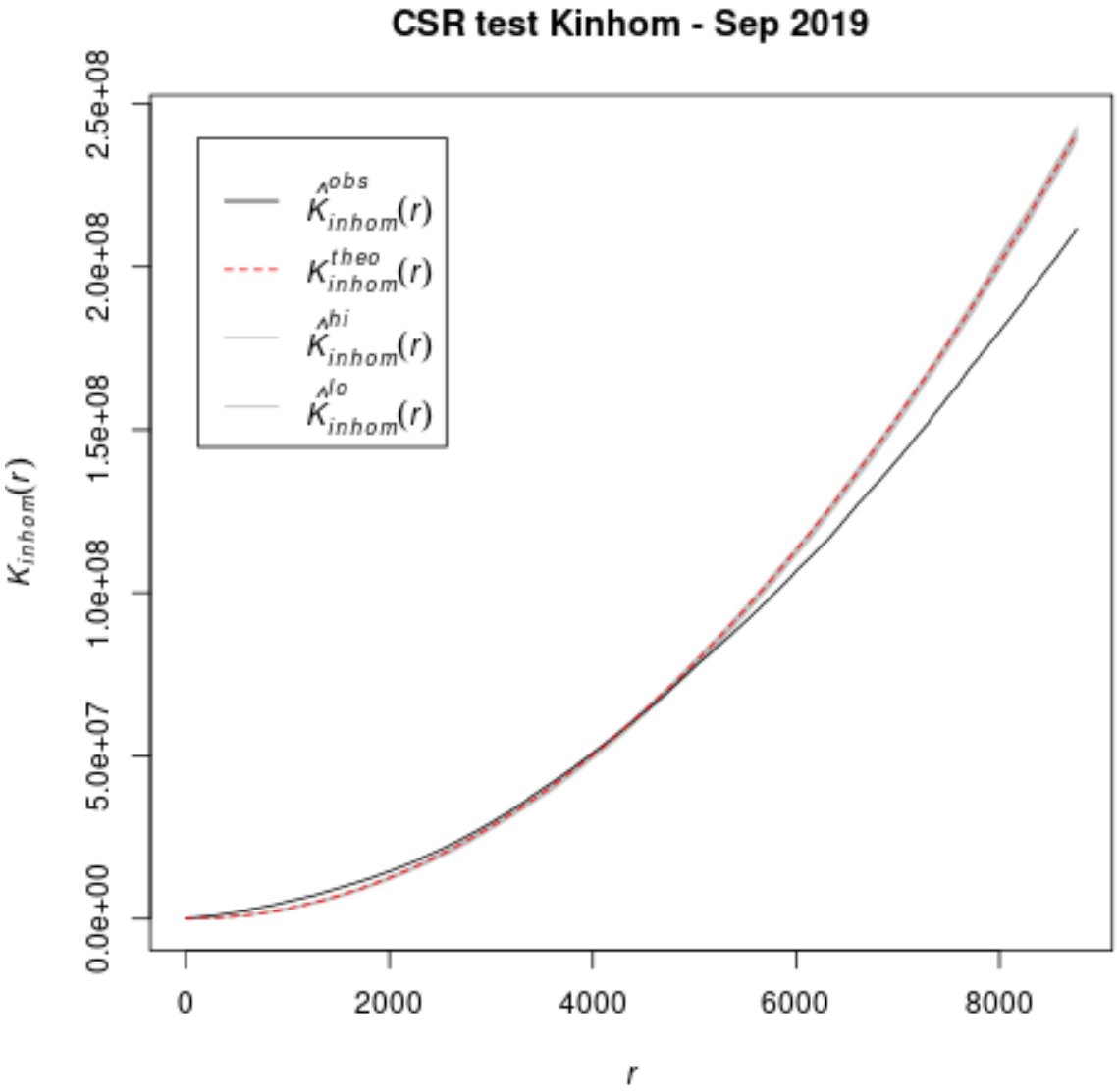


A further test for complete spatial randomness was applied and shown in the below plot. The observed statistic (black line) represents how many points are in a circle of radius r (x-axis), in metres, centred at a random point. The theoretical statistic (red dashed line) represents how many points in a circle of radius r are expected, if the points follow a poisson process. The interval (grey band) is how much the theoretical statistic varied over a repeated simulation.

The observed statistic is above the theoretical statistic and outside of its interval. This is further evidence that the violent crimes are not spatially random.



The result of a similar test which does not assume the point pattern is stationary is shown below. The observed statistic starts above the theoretical interval and then crosses it at around 4,000m and then remains below. This suggests that violent crimes are inhomogeneous at some scales and homogeneous at others.



Similar plots to the above have been generated for other months and yield similar results.

7.3 Model

The model is a negative binomial generalized additive model (GAM) with multiple smooth terms, summarised below.

Term	Description
s(idx, bs = 'mrf', xt = list(nb = grid_nb))	This Markov random field (MRF) term is similar to the spatial smooth plot earlier in the report. This type of term is typically used to model spatial data. It can also 'use' information from neighbouring areas (defined in 'grid_nb') to inform the prediction for that area.
s(four_week)	This thin plate spline term captures the non-linear temporal trend overall.

Term	Description
s(idx, by = four_week)	This term is a thin plate spline which captures the temporal trend for each area.

These terms are combined to generate a prediction as follows.

$$num\_crimes = s(idx, bs = 'mrf', xt = list(nb = grid\_nb)) + s(four\_week) + s(idx, by = four\_week)$$

8 Bibliography

Blangiardo, Marta, and Michela Cameletti. 2015. *Spatial and Spatio-Temporal Bayesian Models with R - Inla*. Wiley.