On Autoregressive Component of Crime Time Series

Kayvan Nejabati Zenouz[*]

22$^{\text{nd}}$ Dec, 2021

Abstract

Investigating patterns in crime numbers has been a topic of research for more than a century. Naturally, results of such research will have crucial impact, not least to effective policing, efficient allocation of resources, but ultimately can lead to saving human lives. Multiple studies have established seasonal patterns in crime numbers. However, a suitable minimal modelling framework for understanding and forecasting crime numbers has not been agreed upon. In this paper, we study the time series behaviour of monthly crime numbers in the UK. For the aggregate monthly crime numbers, we propose several suitable dynamic linear models, with at most 6 parameter estimates, which explain over 82% of variations in the data, provide robust seasonality indicators, and perform with low mean absolute percentage forecasting error. Next, we validate the models through residual analysis, investigate the relative importance of each parameter, discuss the effect of pandemic, and make comparison to auto ARIMA models. Finally, we explore the applicability of our modelling framework to other datasets of crime and to individual crime types and show that crime patterns acquire historic behaviour proportionate to the number of crimes.

Contents

[*] School of Computing and Mathematical Sciences, Queen Mary's Building, University of Greenwich, SE10 9LS, K.NejabatiZenouz@greenwich.ac.uk

1 Introduction

Enhancing an economy while considering the restrictions on resources and the need for constant maintenance of an orderly society mean that great efforts are required in optimizing the allocation of resources; in particular when under provision of a certain resource can also have devastating effects on human lives.

Efficient control and reducing the number of crimes in a society has been a topic of utmost importance throughout the history, however, policing resourcing are as ever either under strain, creating a need for optimisation, or a suitable allocation of these resources can lead to further efficiencies being made, and thus benefiting the society and enhancing the economy.

To this end, in order to provide sufficient policing, a sound understand of criminal behaviour is needed, which has been carried through by the criminologists over the centuries and with great care.

Recently, with the digital age and then the age of data, more and more criminology researchers have begun to focus on learning from the data of crimes and to create mathematical understanding of patterns they exhibit, with an aim to forecast future events, and thus be able to allocate sufficient resources.

Over the past two decades, data has changed the way of human life and with it we have developed sophisticated methods in creating meaning from large collection of observations useful information which would have been lost in the past or gathered without effective processing power.

Consequently, the aim of our manuscript is to add and strengthen the current literature in understanding variations of criminal activities over time through the use of statistical analysis.

We discuss our first result from an ongoing investigations, which aims to understand the patterns in data relating to the crimes recorded in the UK by different policing forces and for each crime category and ultimately provide reliable long term forecasts.

Our main results show that dynamic linear models (DLM) of a certain kind, also known autoregressive distributed lag (ADL) models, can suitably model the monthly crime numbers and provide reliable forecasts. In partciular, we show that models with at most 5 or 6 parameter estimates can be both statistically valid and provide significant forecasting capability. We compare these to automatically selected autoregressive integrated moving average (ARIMA) models and also apply our selected ADLs to other datasets to further test their modelling and forecasting features as well as validity.

Several questions remain as to why these particular models or choice of parameters provide such powerful predictive ability. We go further to investigate if the proposed model structures can form a valid framework for different categories of crimes and for numbers recorded by different forces and show that aggregate crime time series can acquire systematic behaviours whereas the constituent time series may not, however, there is evidence that the aggregate crime time series with larger numbers overall are composed of more systematic constituent time series.

The rest of the manuscript is organised as follows. In Section 2 we conduct a literature review on existing and relevant previous studies. In Section 3 we provide a summary of our main results and discuss the significance of our modelling framework. In Section 4 we discuss the data used and provide visual information, and look at the basic theory of ADL and ARIMA models. In Section 5 we discuss our modelling framework, choose a best one, validate it, and create forecasts for crime numbers using several models for 20 months in the future. We shall also use apply our framework to the crime dataset from Chicago and crime types data and discuss their applicability. Finally, Section 6 we conclude our studies and discuss future related works.

2 Literature Review

Current literature in the usage and applications of ARIMA models as well as time series decomposition in studying the temporal behaviour of crimes is relatively developed and is known to fall into the category of predictive policing [KRAL20]. However, systematic modelling procedures or governing best minimal models are often not agreed upon, or not fully investigated for either the total or each category of crime.

One of the most relevant early studies is by [MLP12], where authors use a variant of the classical time series decomposition in understanding seasonal cycles in crime numbers on a dataset from 88 US cities covering years from 1977 and 2000. They conclude all major crime rates have similar seasonal cycles. They also discuss the advantage and disadvantage of using ARIMA models in studying the seasonality in crime.

The authors use models with at least 12 parameter estimates and use the same model for each category and each area. Their main investigation is not creating a best model for a certain category, but to determine the seasonality characteristics of their data, e.g., in which months the crimes are higher or lower.

More recently and with crime data in the UK the authors of [LDF21] investigated whether the first lockdown has had an effect on the number of crimes for each category. They use the R package forecast [HK08] with automatically selected ARIMA models based on minimizing AIC for each category of crime. They compare the forecasts from the models to the observed data and show that in several categories forecast is higher than the observed values, aiding them to suggest that the lockdown may have induced a sharp, short-term decline in crime.

The authors do not study the details of each time series for each category of crime, the mathematical significance of the models, or the statistical validity of their models, they only use these models to arrive at their conclusions. Particularly, if such declines in crime numbers is observed, how can we construct models which account for the decline and produce reliable forecasts going forward.

In [CYS08] authors study a dataset relating to property crime in one city of China. Their results show that property crime x_t over t in weeks may be suitably modelled with models of type

$$x_t = \beta_0 + \beta_1 x_{t-1} + \epsilon_t \text{ with } \epsilon_t \sim N(0, \sigma^2).$$

In this case the crime numbers is not seasonal, e.g., with 12 months periodic cycles, which indicates that different categories of crime or data from other locations may require different ARMIA modelling and have different seasonality behaviours.

The authors do not discuss whether this model applies to other categories of crimes. In fact it would be useful to know if there is a natural clustering of crime types according to the ARIMA models they are best described by.

Related ideas are also discussed in [BZB+18] where the authors investigate crime data for San Francisco, US, and Natal, Brazil. They divide the regions into subregions and apply time series decomposition to crime time series for each subregion to produce time series feature and ultimately arrive at a spatio-temporal forecasting for crime numbers.

In [CCT16] the authors also apply auto ARIMA models to the number of crimes in Chicago and create models with above 80% accuracy of forecasting for two years. They consider the number of crimes per week and decide that the time series is best modelled by $\text{ARIMA}(1,1,1)(0,1,2)_{52}$.

They do not consider modelling their data through distributed lag models or make any comparisons. We note that although their ARIMA model fits the data fairly closely, on forecasting it always overestimates. This is the issue with ARIMA models as they do not naturally allow for a linear trend in their description, and their data in this case does have a negative linear trend.

In fact, we apply our models and partially reproduce their results see Subsection 5.1 and also can account for the overestimation through ADL modelling. Furthermore, auto ARIMA models do not often present an enlightening simple structure as they may change with the modelling window.

Likewise, there are numerous other studies on the crime forecasting either through other sophisticated methods, or only applying time series methods and use the results and applications for example see [WYB+19]. In [MWW12] the authors use ADL models to related drug related crimes to the usage of drugs.

3 Summary of the Results

We shall briefly discuss our main results in this section. Let us denote by C_t the aggregate number of crimes in all categories recorded over a particular month t which is in the range Jan 2014–Sep 2021. The time parameter is suitably translated into a numerical value of

$$t = 2014 + \frac{i}{12},\ i = 0, ..., 92.$$

We assess several dynamical linear models (also known as autoregressive distributed lag models [cf. LS14]) for C_t that fit the dataset particular well (see Section 5 for reasons why). The main candidates we consider in this paper are

$$C_t = \beta_0 + \beta_1 t + \beta_2 C_{t-1} + \beta_3 C_{t-12} + \epsilon_t \tag{1}$$

$$C_t = \beta_0 + \beta_1 t + \beta_2 C_{t-1} + \beta_3 C_{t-12} + \beta_4 C_{t-13} + \epsilon_t \tag{2}$$

$$C_t = \beta_0 + \beta_1 z + \beta_2 C_{t-1} + \beta_3 C_{t-12} + \beta_4 C_{t-24} + \beta_5 C_{t-25} + \epsilon_t \tag{3}$$

$$C_t = \beta_0 + \beta_1 t + \beta_2 C_{t-1} + \beta_3 C_{t-12} + \beta_4 C_{t-24} + \beta_5 C_{t-25} + \epsilon_t \tag{4}$$

where $\epsilon_t \sim N(0, \sigma^2)$ for all t. In equation number (3) we set $z = 0$ if t lies in Jan 2014–Feb 2021 and $z = 1$ if t lies in Mar 2021–Sep 2021, the ”lockdown” (after-)effect. Then β_i coefficients for $i = 0, ..., 5$ are estimated through maximum likelihood method via dynlm [Zei09] package, Dynamic Linear Regression, of R.

Our results show that although 4 models above can be valid when fitting on the dataset, equations number (3) and (4) seem to provide the best models with highest

R^2, above 82%, lowest mean absolute percentage trend error (MAPTE), at most 2.6%, model (4) can be tested to show relatively low mean absolute percentage forecast error (MAPFE) on a two years validation set, and they have low AIC in comparison to other models.

We use a publicly available dataset on crimes in UK, excluding Manchester Police Force (data for this force is missing since Jun 2019), to estimate the coefficients $\beta_0, ..., \beta_5$ for model number (4). Fitting the same models to the subset of data Jan 2014–Jun 2019, which includes Manchester, will have similar discussion. The coefficient estimates, their significance, and their relative importance is given in the table below.

Table 1: Coefficient Information for Model of Type (4)

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)	Rel R^2
β_0	11215834.289	4314213.761	2.600	0.012	-
β_1	-5551.842	2147.330	-2.585	0.012	7.2
β_2	0.597	0.091	6.578	0.000	37.2
β_3	0.309	0.119	2.604	0.012	28.1
β_4	0.605	0.120	5.034	0.000	3.4
β_5	-0.518	0.108	-4.813	0.000	6.6

Rel R^2 in the above table corresponds to what percentage of the variation in data is explained by each parameter; the numbers are calculated from ANOVA table for the model, a ratio of sums of squares corresponding to the parameter divided by the total sums of square.

The significance of β_1 indicates a declining trend in overall crime numbers, which may be a result of the "lockdown" or maybe a genuine decline in crimes numbers as a result of other factors or a combination of both.

The seasonality indicators are picked up by $\beta_2, \beta_3, \beta_5, \beta_4$ according to their importance in the ANOVA. Rounding the estimates to one decimal place from Table 1 above, the equation creating the model trend is given by

$$C_t = 11215834.3 - 5551.8t + 0.6C_{t-1} + 0.3C_{t-12} + 0.6C_{t-24} - 0.5C_{t-25} + \epsilon_t.$$

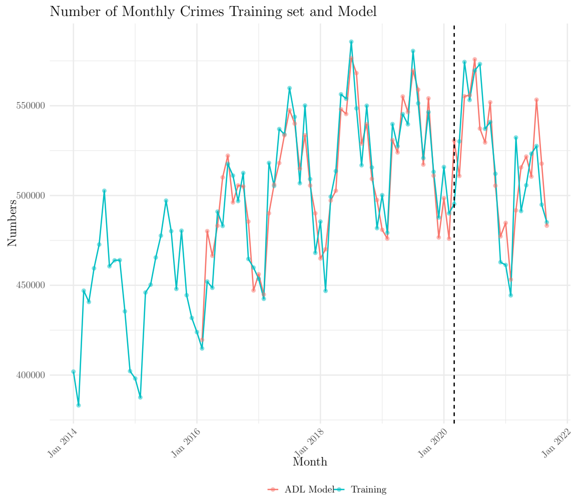
In this case we had expected that crime to be yearly seasonal, so lag 12 is justified, but it is not clear why the other lags are significant, particularly lags $1, 25$, or 13, or in fact 24. We see that the model accounts for large cycles in data, for example high in July to July by the coefficient of C_{t-12}, and also month to month variations through the coefficient of C_{t-1}. The question remains if the other coefficients have significant interpretations, may be excluded or not, or why lags of 24 and 25 are significant suggesting two-year cycles in data.

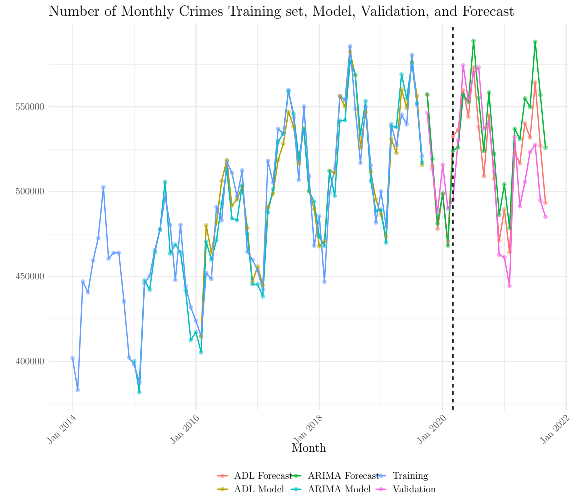
We can fit the same model (4) to a subset of data for validation purposes. We can use the modelling window Jan 2014–Sep 2019, keeping 24 months for validation data (using this subset the β_1 coefficient will no longer be significant as a parameter estimate, however it will still carry significant weight in the ANOVA). Then forecasting from the model we find the MAPFE is 3.4%.

Forecast stays close to the validation data. Here, β_1 is providing a decrease in numbers at each month and including it for long therm may be unreasonable, although it seems to perform well on the 2 years forecast. In fact fitting the model without linear trend to the data before pandemic (before Mar 2020) is both valid and has 89% value for R^2 with 2% value for MATPE.

Hence, β_1 seems to be indicating either a "lockdown" decrease or a genuine decline, or a combination of both, in crimes numbers, which seems to occurred just preceding or around Mar 2020 (which may not also be easily modelled with an auto ARIMA).

The ADL models can be compared to auto ARIMA models, the comparison leads to a better MAPTE for auto ARIMA, but lower MAPFE for ADL. Also the auto ARIMA may change rapidly as the modelling window changes, whereas ADL can be kept the same allowing one to gain a better theoretical understanding of the patterns involved.





The ADL models on graphs above are produced by equations of type (4). Dashed lines on the graphs indicate when the first Lockdown began Mar 2020. The auto ARIMA in the right hand side is produced by an $\mathrm{ARIMA}(1,0,1)(1,1,0)_{12}$ which is automatically selected to minimise AIC. See also [CCT16], ARIMA models tend to over(under)estimate if there seems to be a trend in data. Trends or sudden changes in data however can be accounted for by ADL models.

We further show that similar type ADL models can be compared to ARIMA models for a crime dataset in Chicago and also for time series of different types of crime in UK. In Section 5 we shall introduce the concept of historicity and show that crime time series tend to acquire historic behaviour as described by model (4) more often as the total number of crimes increase.

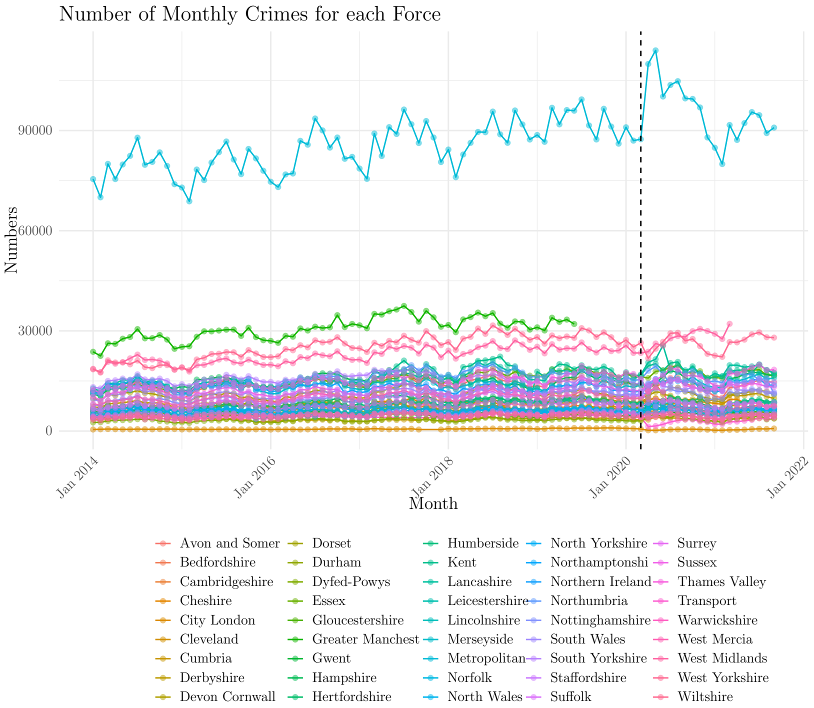
4 Data and Method

During the past decade data relating to criminal activities investigated by the police in the UK has been gathered and made publicly available in https://data.police.uk/. The immediate available data corresponds to monthly records for crimes in 45 different police forces where each police force records crimes for 14 different categories, the data is monthly updated.

There is an archive for the past data, including it the range of available data is Jan 2014–Oct 2021. The master dataset constructed by collating all the individual .csv files contains 47,978,692 observations on 12 variables: *Crime ID*, *Month*, *Reported by*, *Falls within*, *Longitude*, *Latitude*, *Location*, *LSOA code*, *LSOA name*, *Crime type*, *Last outcome category*, and *Context*.

Now one set of analyses may concern with understanding the number of crimes per month for *Falls within* and *Crime type*, thus this results in 630 ”independent” time series, 45 times series for all crimes numbers recorded by *Falls within*, 14 times series if counted by *Crime type*, and a total time series of all *Crime type* and all *Falls within*. The results of this paper are mainly on the data related to the latter time series. Subsequent investigation then focus on any of the other time series.

Several concerns deserve to be mentioned in analysing the data above. First to note is that, out of 45 police forces, Manchester Police Force, which records the second highest number of crimes, has stopped providing data since July 2019. Similarly, West Midlands Police, the third largest recorder of crimes has stopped providing data since Mar 2021.

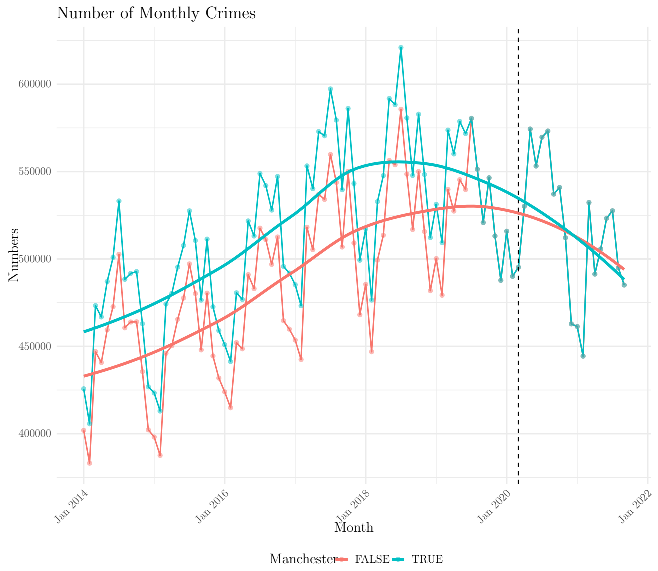


Hence, we shall exclude Manchester Police Force in our analysis in Section 5, and assume that lack of data from West Midlands Police has negligible overall effect, and will not exclude West Midlands Police, however our analysis and models remains robust with or without the inclusion of these forces. Other similar issues with data supply exist but are negligible; for example, British Transport Police has data missing for Jan 2016–Jan 2018. Authors of [TJA$^+$15] have discussed further details on the accuracy of the data available in particular in spacial dimension.

The second issue is the possible drop, or not, in crime numbers due to lockdown and pandemic, and if so how would the model provide useful forecast when the validation set includes data from lockdown period. There seems to have been a change in the numbers for different categories of crimes as also indicated by [LDF21], some raising, anti-social behaviour or possession of weapons, and many falling, though on the total crime number these seem to manifest as a smooth drop.

There are essentially three distinct ways to fit a total model. This is because we may assume that the effect of lack of data from Manchester Police Force or lockdown effect cannot be neglected. Manchester Police Force has not provided data since Jun 2019 and we may assume the lockdown effect started in Mar 2020.

Therefore, one setting is to only use the data from Jan 2014–Jun 2019 so Manchester can be included. The second setting is to exclude Manchester Police Force data altogether, assume the lack of data from West Midlands Police since Mar 2021 is negligible, also the effect of lockdown is natural and use data Jan 2019–Sep 2021. The third setting is to use data excluding Manchester from Jan 2014–Feb 2020.



Time Series Decomposition, ARIMA, and ADL Models

There are various ways of understanding the behaviour of a time series, we shall briefly review these without going into details. Let us consider a times series given by $\{x_t : t = 1, ..., n\}$. The classical time series decomposition splits the time series depending on the whether it is additive or multiplicative through

$$x_t = T_t + S_t + \epsilon_t \text{ or } x_t = T_t S_t \epsilon_t,$$

into trend, seasonal, and error components.

Another way to understand time series is through autoregressive integrated moving average (ARIMA) models. In this exposition an ARIMA$(p, d, q)(P, D, Q)_m$ is given by

$$\phi(B)\Phi(B^m)\left(1-B\right)^d\left(1-B^m\right)^D x_t = \delta + \theta(B)\Theta(B^m)w_t \text{ with } w_t \sim \text{wn}(0, \sigma^2)$$

where $Bx_t = x_{t-1}$ is the backshift operator, p, P are the orders of the polynomials $\phi(B), \Phi(B^m)$ respectively, q, Q are the orders of the polynomials $\theta(B), \Theta(B^m)$, the positive integers d, D are coefficients of differencing respectively, m is the seasonality, and δ is a drift parameter [cf. SS17, p 140]. The values w_t are usually taken as mean zero white noise from a distribution with constant variance σ^2. Some examples, which are used in this paper, are given below.

$$\begin{aligned}
&\text{ARIMA}(1,1,1)(0,1,2)_{52}\\
&(1-\phi_1 B)(1-B)\left(1-B^{52}\right)x_t = (1-\theta_1 B)\left(1-\Theta_1 B^{52}+\Theta_2 B^{104}\right)w_t\\
&\text{ARIMA}(1,0,1)(1,1,0)_{12}\\
&(1-\phi_1 B)\left(1-\Phi_1 B^{12}\right)\left(1-B^{12}\right)x_t = (1-\theta_1 B)w_t\\
&\text{ARIMA}(2,1,0)(1,1,1)_{12}\\
&\left(1-\phi_1 B-\phi_2 B^2\right)\left(1-\Phi_1 B^{12}\right)(1-B)\left(1-B^{12}\right)x_t = \left(1-\Theta_1 B^{12}\right)w_t\\
&\text{ARIMA}(0,1,1)(2,1,2)_{12}\\
&\left(1-\Phi_1 B^{12}-\Phi_2 B^{24}\right)(1-B)\left(1-B^{12}\right)x_t = (1-\theta_1 B)\left(1-\Theta_1 B^{12}-\Theta_2 B^{24}\right)w_t
\end{aligned}$$

The left hand side of these equations is called the autoregressive (of order (p, P)) process and the right hand side is called the moving average (of order (q, Q)).

Alternatively, one may only study the autoregressive part of the time series or investigate the effect of one time series on another. This can be achieved through Dynamical Linear Models (DLM) or also known as Distributed Lag Models (DLM). A basic dynamical linear model for two time series x_t, z_t with linear trend may be represented by

$$x_t = \beta_0 + \beta_1 t + \beta_2 x_{t-p_1} + \beta_2 x_{t-p_2} + \cdots + \beta_r x_{t-p_r} + \gamma_1 z_{t-q_1} + \gamma_2 z_{t-q_2} + \cdots + \gamma_s z_{t-q_s} + w_t$$

$$x_t = \beta_0 + \beta_1 t + \sum_{i=2}^{r} \beta_i B^{p_i} x_t + \sum_{j=1}^{s} \gamma_j B^{q_j} z_t + w_t$$

for some (non-zero) coefficients β_i, γ_j where $i = 0, ..., r, j = 1, ..., s$. We may assume $0 < p_1 < p_2 < \cdots < p_r < n$ and $0 \leq q_1 < q_2 < \cdots < q_s < n$ and $w_t \sim \text{wn}(0, \sigma^2)$ we usually require $w_t \sim \text{N}(0, \sigma^2)$ be i.i.d, independent and identically distributed.

For such series, or any times series with an autoregressive part, we may define the (maximum) **memory** of x_t by

$$M(x_t) = \max(p_r, q_s)$$

and the **history intensity** of x_t by

$$H(x_t) = \sum_{i=1}^{r} p_i + \sum_{j=1}^{s} q_j.$$

We see that $M(x_t)$ is the minimum length of data we need in order to construct one piece of x_t and $H(x_t)$ is the composition of the minimum historical information needed in order to construct x_t. Finally, we recall that the absolute energy of a time series is given by

$$AE(x_t) = \sum_{t=1}^{n} |x_t|.$$

As an example, for

$$C_t = \beta_0 + \beta_1 t + \beta_2 C_{t-1} + \beta_3 C_{t-12} + \beta_4 C_{t-24} + \beta_5 C_{t-25} + \epsilon_t$$

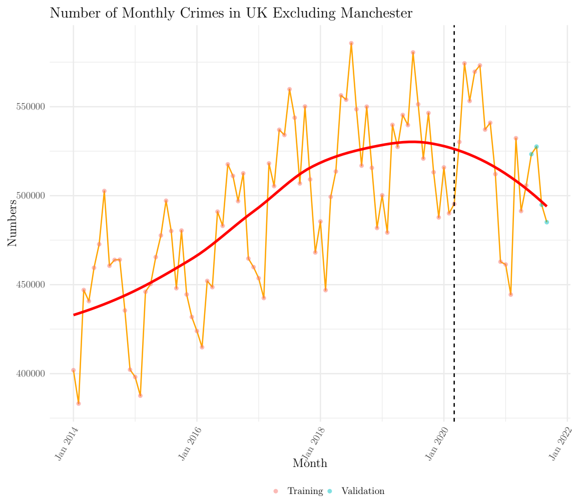
we have $M(C_t) = 25$ and $H(C_t) = 62$ and if C_t is given by $\text{ARIMA}(1,0,1)(1,1,0)_{12}$, we have $M(C_t) = 25$ and $H(C_t) = 75$. If C_t is positive, then $AE(C_t)$ is the sum of numbers C_t for all t.

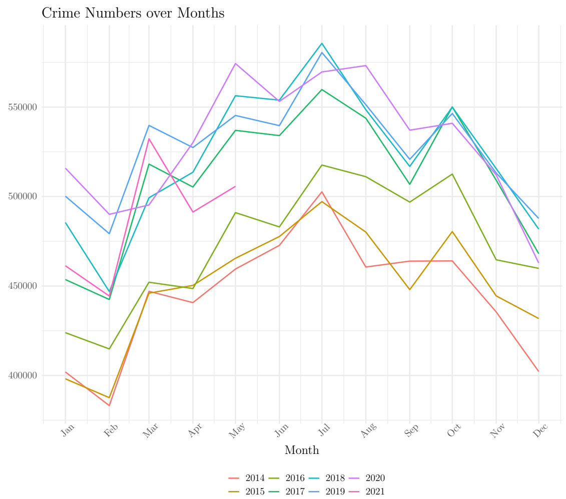
5 Results and Discussions

We shall work for the remainder of this section mainly with data excluding Manchester. We shall keep four months validation points so the training data is from Jan 2014–May 2021. The dashed black lines on the graphs show when the first lockdown started in Mar 2020.

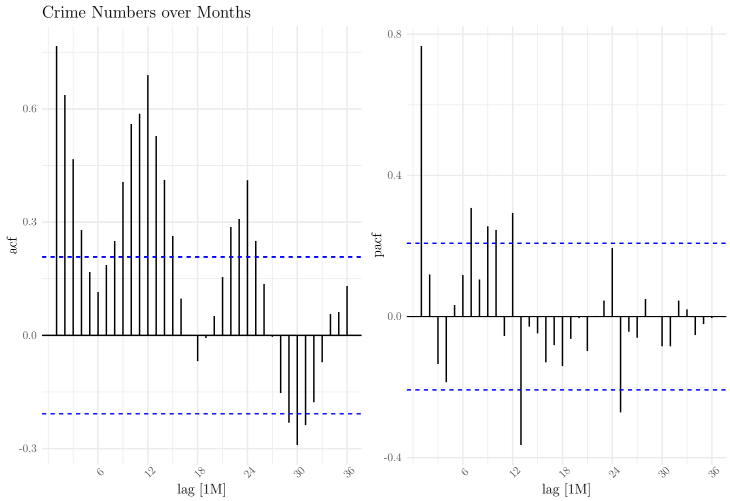
We note the number of crimes have been somewhat stabilising in trend, then during the pandemic and afterwards they have decreased in smooth trend. It is not clear whether there was a one off drop in the number of crimes during and after the lockdown or the drop was linearly graded and continuing. A Mann-Kendall test for trend has p-value of less than 0.05 indicating presence of a trend, however an augmented Dickey-Fuller p-value of less than 0.05 shows that the time series is stationary.

In general the number of crimes seem to reach it's peak in July and trough in February as also studied by [MLP12] using a different dataset. However, there seems to be local maximums in March and October, and local minimums in April and September.





The autocorrelation function indicates most significant positive lags 1 and 12 and perhaps 24, and partial autocorrelation function indicate negative significant lags at 13 and 25.



We note that a classical time series decomposition would require around 12 parameter estimates, and would be a method of modelling for this data, however we investigate the models with smaller number of parameters which may even provide a better and more elegant understanding of the data.

As such and partially informed by the ACF and PACF graphs, we look for autoregression distributed lag models of the kinds given by

$$C_t = \beta_0 + \beta_1 t + \beta_2 C_{t-1} + \beta_3 C_{t-12} + \epsilon_t \tag{1}$$
$$C_t = \beta_0 + \beta_1 t + \beta_2 C_{t-1} + \beta_3 C_{t-12} + \beta_4 C_{t-13} + \epsilon_t \tag{2}$$
$$C_t = \beta_0 + \beta_1 z + \beta_2 C_{t-1} + \beta_3 C_{t-12} + \beta_4 C_{t-24} + \beta_5 C_{t-25} + \epsilon_t \tag{3}$$
$$C_t = \beta_0 + \beta_1 t + \beta_2 C_{t-1} + \beta_3 C_{t-12} + \beta_4 C_{t-24} + \beta_5 C_{t-25} + \epsilon_t \tag{4}$$

where $\epsilon_t \sim N(0, \sigma^2)$ for all t. In equation number (3) we set $z = 0$ if t lies in Jan 2014–Feb 2021 and $z = 1$ if t lies in Mar 2021–Sep 2021, the "lockdown" effect.

The coefficients β_i for $i = 0, ..., 5$ are estimated through maximum likelihood method via dynlm package of R. We fit the models (1), (2), (4) with or without the linear trend. We believe the effect of lockdown has created a significant decrease in numbers making β_1 a negative estimate in all models and significant in all for the period ending Sep 2021. If the modelling period is changed to pre Mar 2021, then β_1 estimate will no longer be of significance, although it will carry significant weight in the ANOVA.

A table summarising regression models information is provided below. The validation is done on 4 months and model (3) has been evaluated with $z = 1$ for forecasting.

Table 2: Comparison of Several Models

Model	Trend	Lag1	Lag2	Lag3	Lag4	R^2	AIC	MATPE	MAFPE
(4)	t	1	12	24	25	83.8	1431	2.52	4.88
(3)	z	1	12	24	25	83.6	1432	2.48	5.17
(4)	–	1	12	24	25	81.9	1436	2.59	5.42
(2)	t	1	12	13	–	80.9	1724	2.80	6.18
(2)	–	1	12	13	–	80.2	1725	2.90	6.71
(1)	t	1	12	–	–	80.5	1750	3.00	7.97
(1)	–	1	12	–	–	78.8	1755	3.10	9.63

We see that the final two models (3) and (4) are the best out of the 7 and the performance of these two is almost the same, except model (4) has lower forecasting error on 4 months validation set. Both of these models have similar properties on the training set and same validity. We have shown coefficient estimate for model (4) with trend t in Section 2 using the full range of data.

In short term forecasting for future numbers model (3) will give a one off drop for the number of crimes for months after the pandemic rather than a continual decline in crimes in time as given by (4), in longer term forecasting (4) may not provide reasonable estimates as numbers will continue to decline to zero, however on two years forecasting the decline is reasonably slow.

The coefficient estimate table for both models (3) and (4) with trend t is given below.

Table 3: Model of Type (3)

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)	Rel R^2
β_0	40524.189	29129.053	1.391	0.169	–
β_1	-16328.955	6579.747	-2.482	0.016	1.0
β_2	0.568	0.100	5.695	0.000	43.8
β_3	0.233	0.120	1.943	0.057	28.2
β_4	0.677	0.128	5.266	0.000	3.8
β_5	-0.538	0.110	-4.893	0.000	6.8

Table 4: Model of Type (4) with t

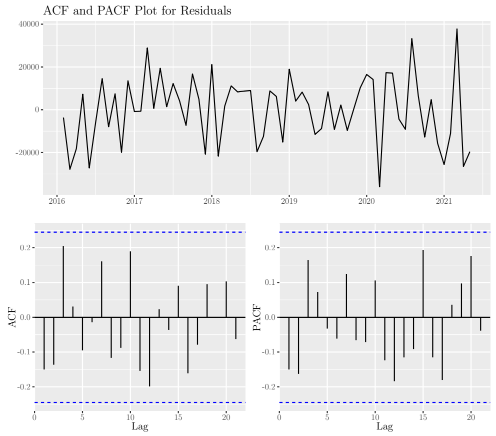
Coefficient	Estimate	Std. Error	t value	Pr(>|t|)	Rel R^2
β_0	11170521.538	4282192.129	2.609	0.012	–
β_1	-5532.836	2131.363	-2.596	0.012	9.6
β_2	0.569	0.099	5.777	0.000	35.7
β_3	0.319	0.121	2.641	0.011	30.3
β_4	0.615	0.125	4.929	0.000	3.3
β_5	-0.492	0.117	-4.212	0.000	5.0

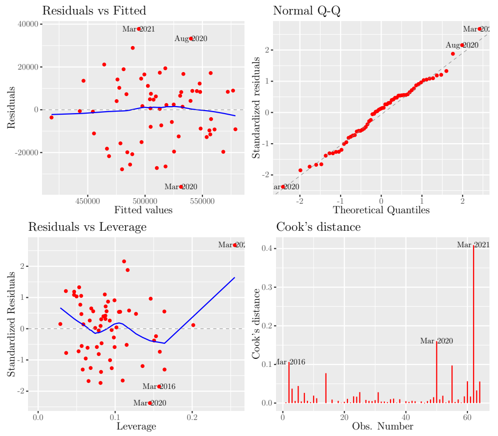
Model of type (4) can be applied to data pre pandemic whereas models of type (3) cannot, so if we didn't know the pandemic was coming, then in fact it'd be better to use a model of the form (4). In the summery of our main result Sections 3 we have shown the forecasting capability of models of type (4) when pulled back to earlier date range.

We believe that models of type (4) are going to provide a better short term future forecast. We shall validate model (4) and forecast for the next two years using several different models.

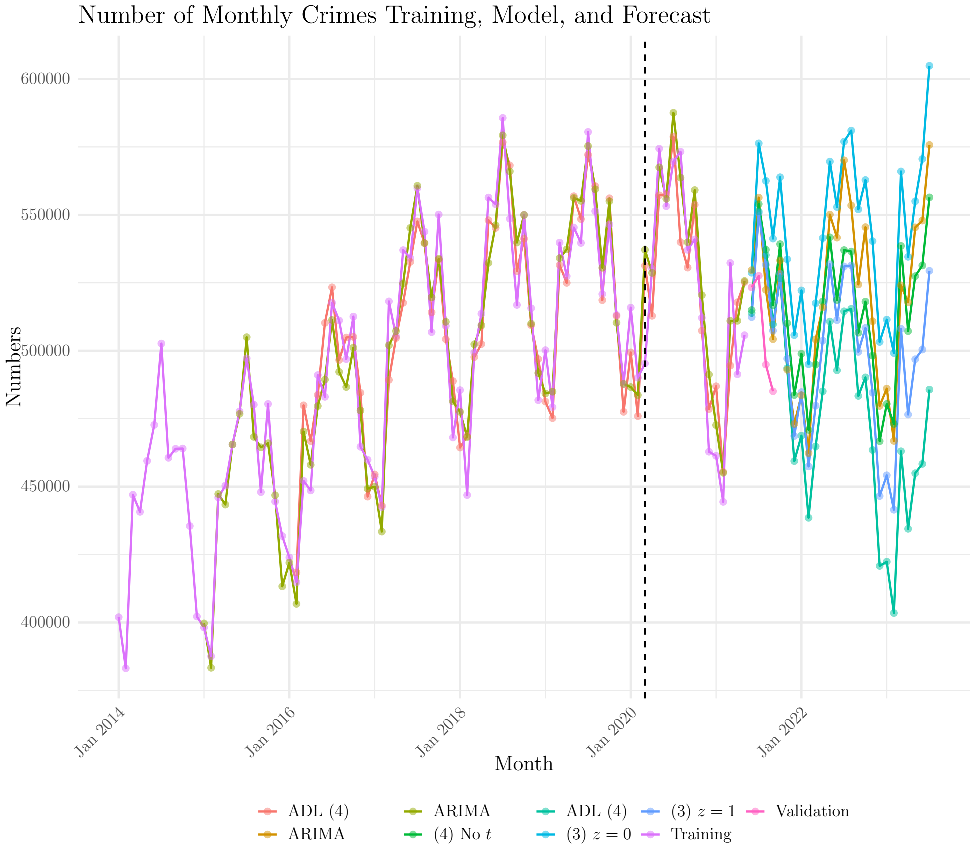
The diagnostic checks show no major autocorrelations and obvious pattens on the residuals. The Shapiro-Wilk normality test has p-value 0.72, no evidence for rejecting the normality of residuals. Breusch-Godfrey test for serial correlation of order up to 24 has p-value 0.07, no evidence that heteroscedasticity is present in the residual. Finally, Box–Pierce has p-value 0.2, so no evidence for rejecting the independence of residuals.

Further checks show no alarming major model violation of normality of residuals. Providing some indication that the residuals are independent white noise from a normal distribution with constant variance.





Now depending on what we assume about the drop in crime numbers whether it is lasting, i.e., $z = 1$ for the foreseeable future, or not $z = 0$, the forecast for overall crimes numbers may take different looks as shown below.



We have also plotted an auto ARIMA model, generated by $\text{ARIMA}(2, 1, 0)(1, 1, 1)_{12}$. Perhaps the assumption that the one off drop in the number of crimes being lasting is not necessary realistic, we may argue that $z = 0$ is a better approximation. However, by trying different validation sets, it appears that models with a linear trend or with z and forecasting $z = 1$ perform better on forecasting. We have already provided a forecasting graph for models of type (4) in Section 3 for a different validation set.

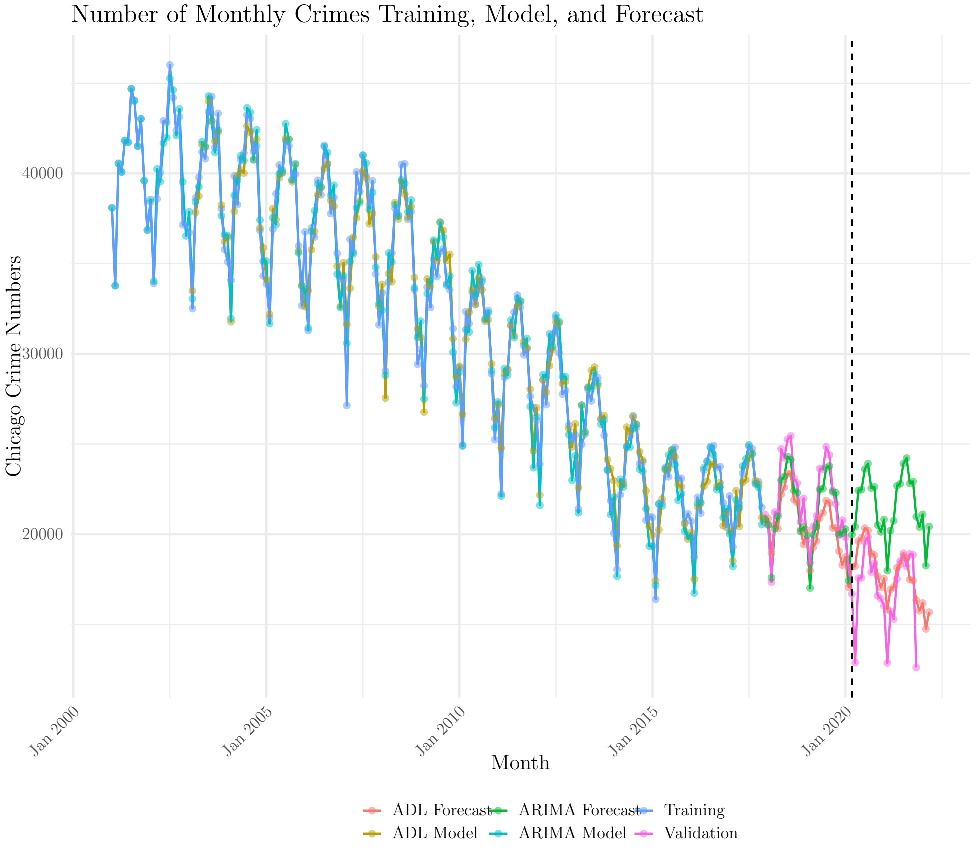
It is not immediately clear why the particular lags $1, 12, 13$ or $1, 12, 24, 25$ provide such a predictive capability. Crime is known to be seasonal, the main contribution to the crime in UK by category is Anti-social behaviour which is seasonal, so we may expect lags at 12 or 24 account for large variations, but it appears that lags $1, 13, 25$ provide control over smaller variations.

In all models we have lags $1, 12$ which are the first and second largest contributors in explaining the overall variability in data, however, lags $13, 24, 25$ seem to provide more accuracy in forecasting. We note that when lag 24 and 25 are present parameter estimate for lag 13 becomes insignificant, hence it has been excluded in some models.

Note, the linear trend is providing a reduction in the number of crimes per month, so it appears that the inclusion of trend helps with a downforce in forecasting, especially recently with pandemic decrease, the parameter helps to explain significant variations in the data, but the parameter estimate for trend may not always be significant as the range of date for modelling changes.

5.1 Comparison with Chicago Crime Data

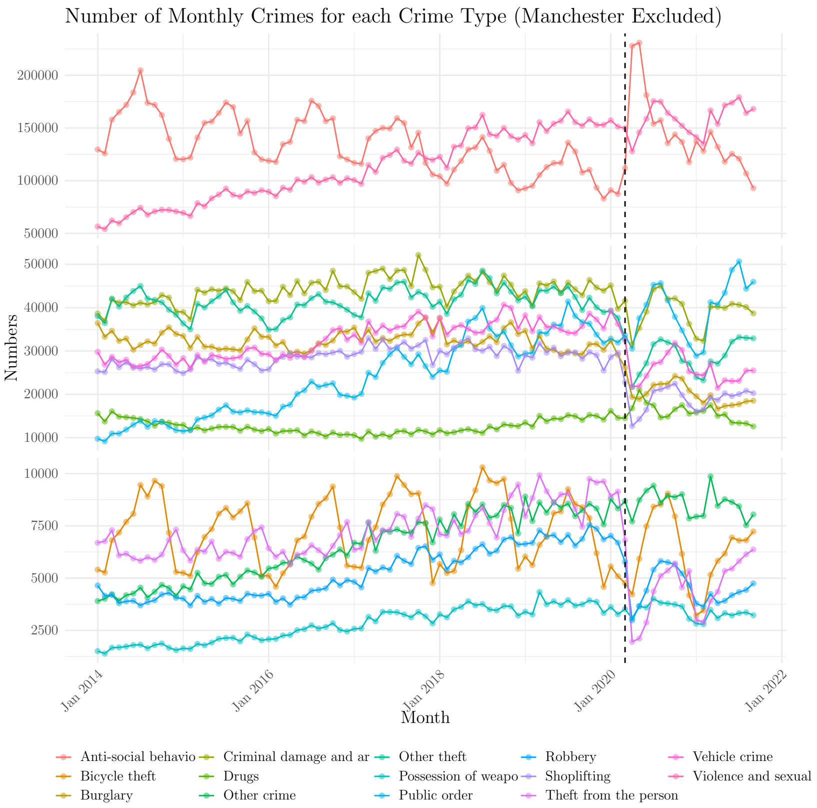
We can apply the same model (4) to crime data from Chicago (data available in https://data.cityofchicago.org/), results shown below [cf. CCT16]. Although the validity of model (4) here may be further investigated (after applying the model a large lag is left in the residuals at lag 24) the model matching the data and forecast from the model remain reasonable even for long periods and stay closer to the validation in comparison to an auto ARIMA.



The ARIMA model is generated by $\text{ARIMA}(0,1,1)(2,1,2)_{12}$. We note that, the crime composition of data from Chicago is different to the UK; for example, in Chicago the major contributor to crime is theft, whereas it is anti-social behaviour in the UK, however, the graph above shows that the same ADL may have the capability to produce reasonable forecasts for two completely different crime datasets. We note that here the ARIMA model seem to produces the amplitudes better than the ADL in forecasting.

5.2 Models for Different Crime Types

We can apply our models to see how they fits for time series in UK split by crime type. Note, it does appear that after lockdown the time series behaviour of several crime types have changed; for example, see anti-social behaviour, criminal damage and arson, or theft from a person.



We apply our models to aggregate crime type time series data, excluding Manchester, and data prior to lockdown, period Jan 2014–Feb 2020. An interesting question for investigation would be how to model the effects of lockdown and produce reliable forecast for after the lockdown period since there has been a noticeable change in the behaviour of time series for certain crime types, as also discussed by authors in [LDF21].

In general we fit three types of models, without trend, with lags $1, 12, 13$ type one, $1, 12, 24, 25$, type two, $1, 12, 13, 24, 25$, type three, and select the most suitable model to each crime type. A summary of selected models is given below. In all these models the coefficients for lags $1, 12, 24$ are positive and $13, 25$ are negative. The lags selected perform well in explaining the variations in data, almost comparable to auto ARIMA in terms of mean absolute percentage error.

Crime Type	L1	L2	L3	L4	L5	R^2	MAPE	$AE(x_t^i)$	AutoARIMA	MAPEA
Anti-social behaviour	1	12	13	24	25	94.2	3.5	9872507	(0,0,4)(0,1,1)[12] wd	3.2
Violence and sexual offences	1	12	13	24	25	97.3	2.4	8174551	(1,0,0)(0,1,1)[12] wd	2.0
Criminal damage and arson	1	12	-	24	25	58.4	2.9	3261576	(0,1,4)(0,1,1)[12]	1.8
Other theft	1	12	13	-	-	82.1	2.3	3064100	(3,1,0)(0,1,1)[12]	1.7
Vehicle crime	1	12	13	24	25	84.5	2.6	2419141	(0,1,1)(0,1,1)[12]	2.0
Burglary	1	12	13	24	25	76.2	2.6	2402495	(0,1,0)(0,1,1)[12]	1.8
Shoplifting	1	12	13	-	-	59.8	3.2	2122849	(2,1,0)(0,1,1)[12]	2.1
Public order	1	12	13	-	-	97.7	3.4	1743177	(0,1,0)(1,1,0)[12]	3.0
Drugs	1	12	-	24	25	86.3	4.2	927043	(0,1,2)(2,0,0)[12]	3.4
Theft from the person	1	12	13	24	25	87.5	4.3	540040	(2,1,0)(1,1,0)[12]	3.6
Bicycle theft	1	12	13	24	25	92.4	5.4	538920	(1,0,0)(0,1,1)[12]	4.1
Other crime	1	12	-	24	25	88.9	3.9	478145	(2,1,0)(0,1,1)[12]	2.6
Robbery	1	12	-	24	25	94.7	3.4	388539	(0,1,1)(2,0,0)[12]	3.3
Possession of weapons	1	12	-	24	25	92.2	3.5	204709	(0,1,1)(1,1,0)[12]	3.3

These give further evidence on the tendency of crime time series to be explainable by lags in the set $\{1, 12, 13, 24, 25\}$, though it is not clear why some time series have one year cycle and for some others a two year cycle fits better. We note that all 14 categories of time series above can be modelled with significant lags only using $1, 12, 13$.

5.3 History Intensity and Average Significant Lags

We may study the properties of the time series for each crime type further by studying the constituent time series which make up the total time series for each crime type.

Note that time series for each crime type is the sum of time series over *Falls Within* which has 45 different levels (including Manchester). In general, if a time series x_t is a sum of smaller times series say

$$x_t = \sum_{i,j=1}^{m,n} x_t^{i,j}$$

so for each i, j we have a time series $x_t^{i,j}$, and we can sum over i or j to have marginal aggregates so

$$x_t^i = \sum_j^n x_t^{i,j} \text{ and } x_t^j = \sum_i^m x_t^{i,j}.$$

For example, in our this manuscript our aggregate time series is the sum of *Crime Type*, say i, and *Falls Within* say j. To each $x_t^{i,j}$ we can associate an auto ARIMA or a dynamical linear model or just a best autoregressive component $ARC(x_t^{i,j})$ and thus a memory $M_{i,j} = M(x_t^{i,j})$ and a history intensity $H_{i,j} = H(x_t^{i,j})$.

Now suppose by a way of investigation we know that $H_{i,j} \in \{\alpha_1, ..., \alpha_k\}$, this is an ordered set of positive integers. For each i we can count over j the number of time series with $H_{i,j} = \alpha_\ell$ and let that number be $n_{i,\ell}$. Then create the weighted average

$$AH_i = \frac{\sum_{\ell=1}^k n_{i,\ell}\alpha_\ell}{\sum_{\ell=1}^k n_{i,\ell}},$$

which shows the tendency of a time series to have longer memory intensity or have constituent time series which had a higher tendency to have history intensity made of $\{\alpha_1, ..., \alpha_k\}$.

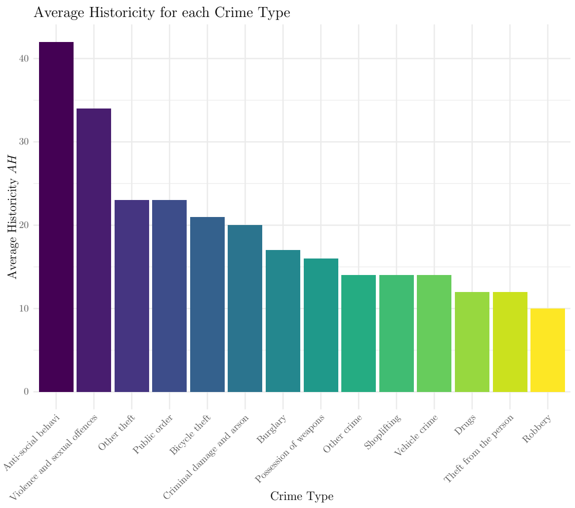
Next we fit models of type $1, 12, 13$ and $1, 12, 13, 24, 25$ to each of the individual $14 \times 45 = 630$ time series, and select the model in which average historicity is higher, i.e., some of the significant lags is larger.

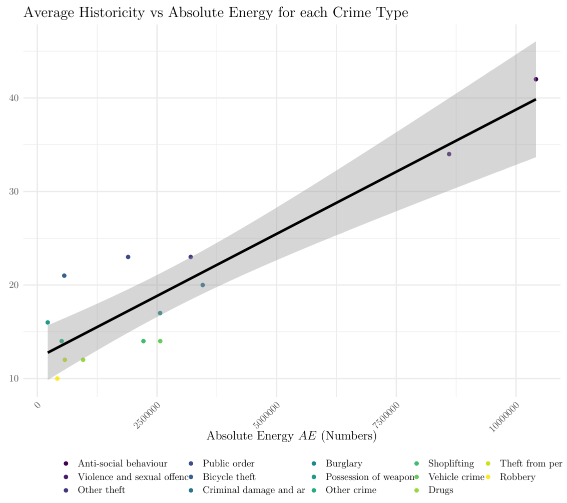
The result of this analysis in the current dataset shows that in general anti-social behaviour and violence of sexual offences have constituent times series which are more regular over *Falls Within*, whereas robbery and theft from the person have on average less regular constituent time series.

Furthermore, it appears that AE_i, the absolute energy $AE_i = AE(x_t^i)$ of a time series, has a positive correlation with AH_i, with significant Spearman's rank correlation $\rho = 0.646$. In a regression we have

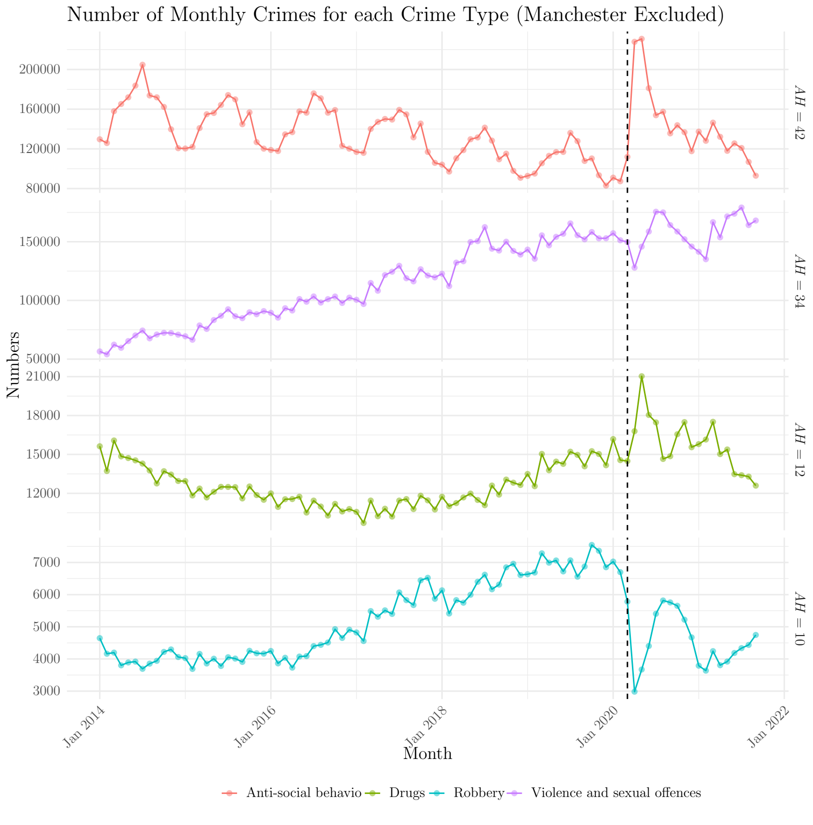
$$AH_i = 12.18 + 0.00000266 AE_i + \epsilon \text{ with } \epsilon \sim N(0, \sigma_\epsilon^2),$$

with $R^2 = 83\%$. Therefore, the average historicity over sub-time series increases as the absolute energy of the total time series increases, and also crime types may be organised according to their average historicity values as seen on the left graph below.



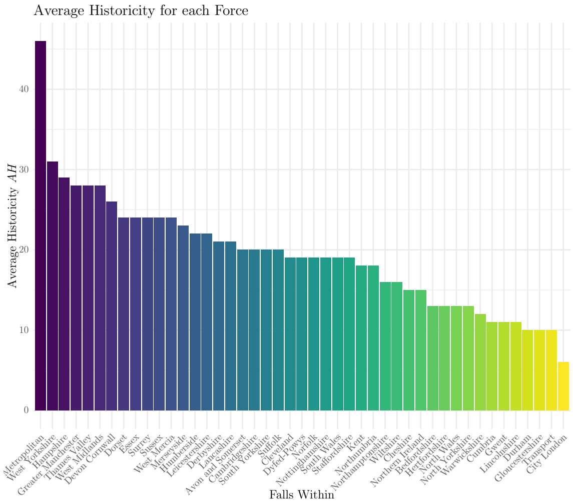


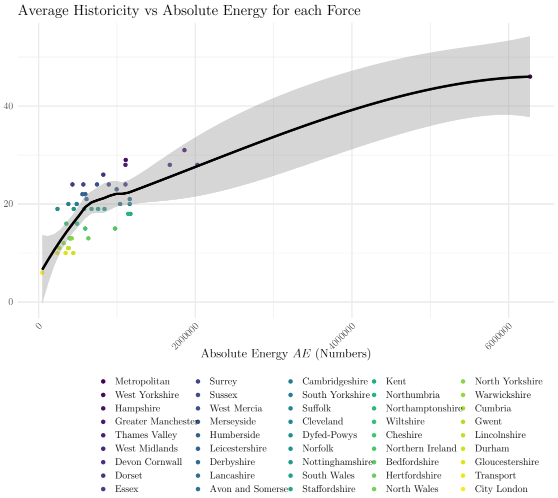
In simple terms it means that as the total number of crimes increases, the historical information afforded by the sub-time series increase. However, it should be noted that the two largest type of crime, antisocial behaviour and violence of sexual offences, seem to have a significant influence in creating this correlation as seen on the right graph above.



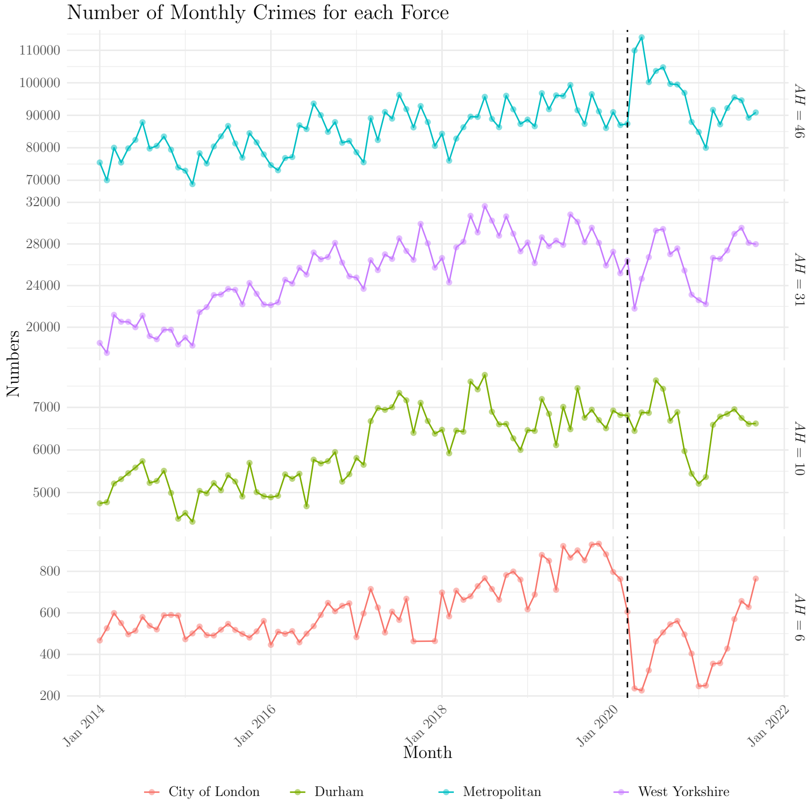
The graph above should the four time series for crime types with low AH and two with high AH. The time series for crime types with higher AH should have more regular seasonal constituent time series.

Repeating this procedure for *Falls Within* produces similar graph, but a weaker correlation. Note in investigating individual time series by crime type and force, data from Manchester can be included.





Again this provides some statistical evidence that crime patterns acquire historic behaviour as the number of crimes increase. The crime time series for four forces, two from each extreme ends, is provided below.



The city of London has the lowest crime numbers, and the pattern in its time series is barely visible, on the other hand the rest of the crime time series have move visible cycles.

6 Conclusions and Future Work

We have tested the validity of an autoregressive distributed lag model of the type

$$C_t = \beta_0 + \beta_1 t + \beta_2 C_{t-1} + \beta_3 C_{t-12} + \beta_4 C_{t-24} + \beta_5 C_{t-25} + \epsilon_t$$

in explaining the variations is monthly crime numbers in the UK. The model is simple, statistically valid, has a high R^2 value, and can provide reasonably low forecasting errors. Most of the variation is absorbed by lags $1, 12$, and there seems to be a negative trend after taking into account the contribution by lags when using the largest possible modelling window.

We have compared our results from the models above to auto ARIMA models and showed that ADL models can provide further degrees of flexibility in forecasting; at the same time they can be robust, easier to interpret, and be kept fixed over datasets. It is clear that for long term forecasting the auto ARIMA models tend to stay on the same path without a trend, however ADL model can take into account sudden drops or trends in data. We have tested the ADL model above on a crime dataset from Chicago to show that it also preforms reasonably well and can account for trends where an auto ARIMA may not.

This was the first step, as we have only modelled the aggregate numbers. The results here goes to show that auto ARIMA models may not necessarily be the best option when it comes to crime forecasting, one for their inflexibly in accounting for trends suitably, but also they can rapidly change as the modelling window changes, offering no overall understanding over the patterns.

Next we tested models of type

$$C_t = \beta_0 + \beta_2 C_{t-1} + \beta_3 C_{t-12} + \beta_4 C_{t-13} + \beta_5 C_{t-24} + \beta_6 C_{t-25} + \epsilon_t$$

in understanding the behaviour of sub-time series by crime type and by force and showed that they can model the categories close to auto ARIMA models. It is not clear that the same ADL model is valid for different categories of crimes or data from other countries adhere to the same model. For example, it is certainty not the case that same type of models perform as well for anti-social behaviour violence or drugs as for the aggregate numbers. We have selected the best models using only the lags in the model above for the all 14 categories of crimes type and further investigated into how crime time series seem to acquire such seasonal behaviour.

The framework of this paper shows that a simple governing modelling structure provides a significant understanding over crime times series and explains universal behaviours in these types of data. Although many questions remain, one possible next step is to study what best models each category of crime, give also the effect of pandemic, and produce reliable forecasts.

References

[BZB+18] Julio Borges, Daniel Ziehr, Michael Beigl, N. Cacho, A. Martins, A. Araujo, L. Bezerra, and Simon Geisler. Time-series features for predictive policing. In *2018 IEEE International Smart Cities Conference (ISC2)*, pages 1–8, 2018.

[CCT16] Eugenio Cesario, Charlie Catlett, and Domenico Talia. Forecasting crimes using autoregressive models. In *2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, pages 795–802, 2016.

[CYS08] Peng Chen, Hongyong Yuan, and Xueming Shu. Forecasting crime using the arima model. In *2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, volume 5, pages 627–630, 2008.

[HK08] Rob J. Hyndman and Yeasmin Khandakar. Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software*, 27(3):1–22, 2008.

[KRAL20] Ourania Kounadi, Alina Ristea, Adelson Araujo, and Michael Leitner. A systematic review on spatial crime forecasting. *Crime Science*, 9(1):7, 2020.

[LDF21] S. Langton, A. Dixon, and G. Farrell. Six months in: pandemic crime trends in england and wales. *Crime Science*, (10), 2021.

[LS14] Spyrides M. Lucio P. Lopo, A. and J. Sigró. Uv index modeling by autoregressive distributed lag (adl model). *Atmospheric and Climate Sciences*, (4):323–333, 2014.

[MLP12] David McDowall, Colin Loftin, and Matthew Pate. Seasonal cycles in crime, and their variability. *Journal of Quantitative Criminology*, 28(3):389–410, 2012.

[MWW12] Steve Moffatt, Wai-Yin Wan, and Don Weatherburn. Are drug arrests a valid measure of drug use? a time series analysis. *Policing: An International Journal of Police Strategies & Management*, 35(3):458–467, 2021/12/02 2012.

[SS17] R.H. Shumway and D.S. Stoffer. *Time Series Analysis and Its Applications: With R Examples*. Springer Texts in Statistics. Springer International Publishing, 2017.

[TJA$^+$15] Lisa Tompson, Shane Johnson, Matthew Ashby, Chloe Perkins, and Phillip Edwards. Uk open source crime data: accuracy and possibilities for research. *Cartography and Geographic Information Science*, 42(2):97–111, 2015.

[WYB$^+$19] Bao Wang, Penghang Yin, Andrea Louise Bertozzi, P. Jeffrey Brantingham, Stanley Joel Osher, and Jack Xin. Deep learning for real-time crime forecasting and its ternarization. *Chinese Annals of Mathematics, Series B*, 40(6):949–966, 2019.

[Zei09] A Zeileis. Dynlm: Dynamic linear regression. *CRAN.R-project.org*, 2009.