

MSc Project - Reflective Essay

Project Title:	Investigating the effect of Coronavirus Lockdowns on Crime Rates within the UK
Student Name:	Elliot Linsey
Student Number:	210764484
Supervisor Name:	Prof. Norman Fenton
Programme of Study:	MSc Data Science and AI

For my MSc project, I used a combination of classical and Bayesian statistical methods to investigate the effect that the coronavirus lockdowns had on crime rates within the UK. During the coronavirus pandemic of 2020 – 2021, radical and sudden changes were introduced to UK society in terms of public mobility. This produced a unique environment in which the usual risk factors such as a perpetrator and victim's mobility (Farrell et al, 2020) and situational opportunity for a crime to occur (Clarke, 2012) were reduced. I felt this was an interesting topic to explore as the pandemic had recently ended and the aftereffects were still being experienced throughout the country. Papers I had researched on the topic had also only focused on short time periods within the pandemic, such as the immediate period after lockdown was initiated. During my own research more data was available and allowed a full look back on the entire pandemic period. Using classical statistical models allowed investigation into the data that was observed and the associations between variables. The Bayesian causal network was able to implement uncertainty about causes and was used to try and model the true and unobserved crime rates.

Analysis of Strengths and Weaknesses:

The first weakness and strength encountered within my project is related to the data, which was obtained from <https://data.police.uk/>. In general, the problems come down to the level of granularity that the data is viewed with. This data required a fair amount of data cleaning and manipulation and throughout this process some of the information had to be removed, combined, or reduced down for the sake of simplicity and resulted in a lower level of granularity (or a more pulled back view). An example of this being the combination of multiple types of theft into one 'Theft' variable. This meant that the nuances within this variable type may not be adequately represented or insights may be missed. It may have been found that burglary decreased but bicycle theft increased, however this is not possible to tell and the only conclusion we can draw is about theft as a whole. This problem is also present within the other variables, a particular area to investigate would have been within 'Violence and Sexual Offences'. During my research I found many instances of domestic violence increasing (Taub et al, 2020). However, due to the structure of the data I was not able to isolate that specific crime sub-type from within 'Violence and Sexual Offences' and so I was not able to confirm whether the papers from my independent research were accurate.

The same granularity problem occurs with the 'Region' variable, as 40 police constabularies are present in the data but were combined into 5 overall regions. This could also introduce confounding variables due to the difference between urban and rural areas present within each region. It could have been possible to separate the areas into these two variable types and identify differences between them to try and avoid this confounding variable. However, by doing this region could become the confounding variable. For example, is it right to compare a rural area from the North to a rural area in the South East? Would there be regional differences between them? These types of dilemmas were present throughout the project and had to be weighed up to try and make the best decision. I think that separating the country into 5 regions was the best decision for this particular project, as it provides a good level of view into the crime rates without being overbearing. Using the full 40 police constabularies could have provided more in-

depth information, but I wanted to get an overview of the effect of the pandemic on crime rates and did not require that high level of granularity. Also, the data visualisation of 40 area variables would have been an information overload and therefore redundant.

I think the ARIMA method of using previous non-covid data to predict expected crime rates and then comparing the observed rates was a strength of my project. It clearly demonstrated the difference between expected and observed rates, as well as using the 95% confidence intervals to show how close the predictions were. It highlighted the significant drop experienced by the initial lockdown over some of the 5 crime types as well as crime as a whole and then showed how crime reacted over the total pandemic period. As before, a large number of studies only took into account the first 6 months of the pandemic and did not show the effects over the full time period. The only issue with the ARIMA model used is that for the best predictions it is recommended to use data from 5 seasonal cycles. This would have required data from 2014 – 2019 which I did not use due to data size limitations and time constraints. This did not seem to have a major effect on my predictions, but may have affected 'Criminal damage and arson' due to the prediction not fitting the previously observed patterns.

Overall, I think that my use of classical statistics was sound. In using normalised data as well as percentage changes and probabilities, I was able to avoid the problems that would have been caused by each region having different population counts. This meant that each region and crime type were immediately comparable and issues such as extremely different ranges (Violent and sexual offences averaged around 1.5million each year whilst Drug offences were only 150,000) were not present. The chi-squared test was the weakest section of my investigation, this is due to the test being of no practical significance when used with high sample sizes (Lin et al, 2013) as well as the fact that using it on temporal data should most likely be avoided due to the difficulty in proving that each year is inherently independent from one another. In essence, knowing the crime rate for one year is likely to have an impact on the next and this should be avoided.

The Bayesian networks contained both strengths and weaknesses. The small network was good at investigating the relationships between the existing variables within the data, however this was not making full use of the causality or uncertainty that Bayesian networks are able to model. In essence, it was an extension of the 'Outcomes' section of my paper as you could generate the same probabilities seen there but also investigate multiple variables at the same time. A definite strength of the Bayesian networks was the ability to create interactive models that could be hosted online. These promote the accessibility of this type of statistical investigation and allow researchers to explore parts of the model that they are interested in, without the specific mathematical knowledge that other methods may require.

The larger Bayesian network tried to incorporate uncertainty into the model but I feel it could have been expanded more. I think that the justifications used to estimate the actual number of crimes were reliable overall, but there is an assumption that the actual number of crimes moved in the same manner as the reported number. For instance, it could be that some crime types and their true crime rates did not differ very much from non-lockdown to lockdown, yet people were less able to report it. An example could be domestic violence, that the true rate could have potentially stayed the same or increased due to the increased proximity of victim to perpetrator (Sri et al, 2021), but the reported rate could decrease due to fear of reprisal on the victim and less access to aid services (Women's Aid, 2020). I also think that the policy focus justifications were reliable to use, due to the research supporting them. However, it was difficult to know exactly what default level to set each one. If I know that the UK has a high focus on violent and sexual offences, should I set the probability of a high focus to 60%? Another researcher may set it at 70%, so at what point does it become an arbitrary decision?

I would have liked to have implemented region as another causal factor as one that both affects accuracy and also changes the reported and actual crime rates to reflect that specific regions results. Currently, the model is only able to generate the values over the entire UK and this means you cannot investigate regional differences. While the model has strengths and is quite close to the reported data, there are certainly more risk factors and causes that I would have liked to investigate.

Presentation of possibilities for future work

This project has a high scope for expansion. The higher granularisation of the data would allow more specific insights to be gathered. For instance, I could have investigated each police constabulary and grouped the results by region. This would provide a more in-depth look at each region and could identify which constabularies saw decreases or increases of certain crime types. From a practical point of view, those constabularies that saw greater decreases or increases could have their methods reviewed to see if there were any common factors that could be extrapolated and identified within other police forces around the country. For ultimate granularity, the crime types would also not be combined, allowing every crime sub-type to be investigated too.

If I wanted to go even further into the specifics within each constabulary, the data of each crime record contains both the latitude and longitude of where it occurred. Using these, I could have plotted locations of each crime and highlight areas where hotspots occur. It could be found that many thefts occur at a particular shop or street, for example. Unfortunately, the time of the incident is not recorded in the data which could be another useful feature.

This project could also be applied to other countries around the world and comparisons could be made between them to investigate the effect that their specific lockdown policies had on crime rates. A good control group to use could be Sweden, who issued no formal lockdown policy and therefore crime may not have been affected so strongly as for other countries. Using more countries, I could also have identified more causal links for a Bayesian network, such as average temperature or average number of sunlight hours. It could be expected that those countries that have longer nights could see a difference in crime rates compared to those that have shorter nights. I would have liked to investigate economic factors too, with those countries that supported their populace with schemes such as the UK furlough being compared with those that did not or provided reduced aid.

Critical analysis of the relationship between theory and practical work produced:

For the classical statistics section of my project, the theory and practical work joined together well. The only issue occurred within the chi-squared test as mentioned previously. Apart from that, the time series data was implemented effectively as well as the percentage changes and probabilistic outcomes components.

The main issue between the theory and practical work occurred during the Bayesian network. Whilst the network was able to model the uncertainty around certain variables, it is not able to answer counterfactual questions such as "Would we expect this amount of crime during 2020 if a lockdown was not in effect?" It is instead applying a reduction or increase factor depending on whether the 'Covid' node is set to True or False. In regards to seeing the difference between the observed rates of crime and using uncertainty to also model the true rates it works well; however, it is still a Bayesian network that at this stage can only investigate the association between variables.

Awareness of legal, social, ethical issues, and sustainability:

Due to the type of data used and the social implications of such analysis, there is a definite requirement for this type of investigation to be carried out ethically. To that end, the data used was all anonymised by the source due to the potential for personal and identifying data to be present.

In regards to the anonymisation of this data, there are a couple of issues that could be addressed. Firstly, although there are no names allocated to any of the crime incidents that took place, both the time (to the month and year) as well as location (within which police constabulary as well as to the longitude and latitude) are included within the data. If this was to be combined with other data such as local newspapers or local crime reports, it could be possible to identify the victim and/or perpetrator. The outcome would also be available and could be linked to the perpetrator if identified. This is a serious issue due to the UK GDPRs rules on the processing and dissemination of personal data (EU GDPR, 2016, Article 4.1), as well as the fact that this data contains very sensitive information due to it containing criminal or potential conviction information. If the data is truly anonymised, then it is ok to be processed, however when combined with other data there is the potential for identification which seriously impacts the security of the data.

The use of machine learning within criminal data also has ethical issues. This is due to the potential for bias to be included by the model creator. For example, if I had used the longitudes and latitudes to identify areas where more crimes occurred, then found that these areas had majority populations of certain ethnic groups, I could link these ethnic groups with higher crime rates and introduce bias into the model. These types of models already exist and are being used within policing systems (such as PredPol in the US). Due to them being the creation of private companies, they are not able to be examined and have their biases identified by external sources. Due to this, there are many instances of disproportionate policing and searches being made in primarily African-American areas within the US (Heaven, 2020).

Due to these issues, foundations such as IEEE (Institute of Electrical and Electronics Engineers) have created standards that should be followed when creating a machine learning model. These include the protection of universal human rights, as well as the transparency and accountability of researchers and companies using sensitive data and material. Awareness of misuse and the competency of the machine learning models should be taken into account when relying on it for any decision which will impact a natural person. Due to the severity of a mistake being made when using a machine learning model within the justice system, these models and analyses should be scrutinised to the highest standard.

I believe that my work does not contain any biases, with judgements such as in the Bayesian network being justified using independent research and the classical statistics being primarily observational.

Conclusion:

Overall, I have successfully investigated the effect of lockdowns on different crime types throughout the main regions of the UK. I discovered trends and patterns using different analytical techniques that I encountered throughout the modules I undertook this year at QMUL. There is certainly scope to investigate the data to a fine level of granularity, but in terms of creating easily communicable figures and results I am happy with what I have produced. The Bayesian network could be expanded upon with more causes as well as the implementation of true counterfactual reasoning. I am also aware of the ethical issues surrounding the use of machine learning and data analysis within the criminal justice system and how they can affect minority populations or the disenfranchised.

References:

Clarke, R.V. (2012). Opportunity makes the thief. Really? And so what? *Crime Sci* 1, 3 <https://doi.org/10.1186/2193-7680-1-3> accessed [03/08/22]

EU GDPR, (2016), Regulation (EU) 2016/679 of the European Parliament and of the Council (GDPR). Available at: <https://www.legislation.gov.uk/eur/2016/679/contents> [Accessed 03/08/2022]

Farrell, G., N. Tilley. (2020). Coronavirus: How crime changes during a lockdown. *The Conversation*, 02 April. [Coronavirus: how crime changes during a lockdown \(theconversation.com\)](https://theconversation.com/coronavirus-how-crime-changes-during-a-lockdown) accessed [03/08/22]

Heaven, Will. (2020). Predictive policing algorithms are racist. They need to be dismantled. <https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/> accessed [03/08/22]

Lin, Mingfeng. Lucas, Henry Jr. Shmueli, Galit. (2013) Research Commentary—Too Big to Fail: Large Samples and the p-Value Problem. *Information Systems Research* 24(4):906-917. <https://doi.org/10.1287/isre.2013.0480> accessed [02/08/22]

Sri AS, Das P, Gnanapragasam S, Persaud A. (2021). COVID-19 and the violence against women and girls: 'The shadow pandemic.' *International Journal of Social Psychiatry*.;67(8):971-973. doi:10.1177/0020764021995556

Taub, A., Bradley, J. (2020). As domestic abuse rises, UK failings leave victims in peril. <https://www.nytimes.com/interactive/2020/07/02/world/europe/uk-coronavirus-domestic-abuse.html> accessed [03/08/22]

Women's Aid. (2021). Shadow pandemic – shining a light on domestic abuse during covid. https://www.womensaid.org.uk/wp-content/uploads/2021/11/Shadow_Pandemic_Report_FINAL.pdf accessed [03/08/22]