**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 looked at 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 particularly area to investigate would have been within ‘Violence and Sexual Offences’. During my research I found many instances of domestic violence increasing. 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.

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 model has a high scope for further investigation.

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 [30/06/22]

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-134948) accessed [30/06/22]

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

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 [01/07/22]