University of Glasgow

College of Science and Engineering
Graduate School

MSc Statistics Dissertation

Supervisors: Dr Nema Dean, Prof. Michele Burman, and Prof. Ade Kearns

Comparing Crime Hotspots at Different Areal Resolutions in Strathclyde

Rebecca Miriam McKay

Date of Submission: 31/10/2018

Abstract

Crime hotspots are used by police and government agencies to target interventions and resources in key high crime areas. It is therefore of interest to look at how hotspots are identified. Hotspots can be identified by clustering and then finding the clusters with a high crime level. The modifiable areal unit problem (MAUP) can have an impact on the clusters identified. MAUP means that if the data are aggregated to different areal units, the results can differ. This impact was investigated using crime data provided by Strathclyde police (now Police Scotland) which covered all crimes (bar crimes of a sexual nature) over the financial year 2011 by clustering this data at two different levels of aggregation (output areas and data zones where output areas are nested within data zones). Clustering was carried out using 4 different cluster methods (k-means, finite mixture modelling, Local Moran's I and Getis Ord Gi*). Maps were produced to visualise this and the adjusted Rand index (a measure of similarity between clusterings) was calculated for each cluster method at the output area and data zone level. The results showed that there was not much similarity in the clusterings produced at the two different areal levels. At the output area level, the methods, k-means, finite mixture modelling and Getis Ord Gi*, clustered over 90% of the output areas in the lowest crime cluster and therefore the lowest crime areas. However, Local Moran's I had less than 7% in the low crime cluster and this shows there can be a great dissimilarity between cluster methods. When comparing these results at the data zone areal level, there was a distinction between using methods which assumed spatial contiguity and those which made no assumptions. Both k-means and finite mixture modelling produced clusters which had most data zones lying in the low crime cluster while Local Moran's I and Getis Ord Gi* had most data zones in the medium crime cluster (or non-significant cluster). This shows that at the output area level, most output areas are in the low crime cluster but at the data zone areal level, most data zones are in the medium crime cluster highlighting the difference in clusters identified at each areal unit. This highlighted the MAUP and the importance of choosing the correct areal level for the analysis.

Maps were again used to visualise the clustering output for both output areas and data zones at the output areal level and the adjusted Rand index was calculated and the results showed that there were similarities in the k-means and finite mixture modelling clusterings and also between the clusterings identified by Local Moran's I and Getis Ord Gi*. Therefore, this shows the importance of choosing areal units and methods wisely, based on the analysis to be undertaken.

Acknowledgements

I would like to acknowledge my supervisors, Dr Nema Dean, Professor Michele Burman and Professor Ade Kearns without whose constant guidance and support this thesis would not have been possible. I would also like to thank Professor Gwilym Pryce whose guidance at the start was greatly appreciated. Also, Dr Ellie Bates whose help was extremely important during my studies in guiding me when using ArcGIS software. Dr Ellie Bates was also kind enough to provide the steps which she had taken in order to create the shape file for Strathclyde at the 2001 data zone areal level.

My parents, Moira and Seaton, and my fiancé Frank Chalmers whose patience and constant support was greatly appreciated as without this I would not have been able to succeed with the dissertation.

Declaration

I have prepared this thesis myself; no section of it has been submitted previously as part of any application for a degree. I carried out the work reported in it; except where otherwise stated.

Contents

Tables

Figures

Chapter 1 - Introduction

Crime hotspot analysis is important as it is mainly used by police and government agencies who wish to identify high crime areas in order for them to target interventions and resources in this area. The Modifiable Areal Unit Problem (MAUP) can influence the hotspots (clusters with high crime levels) identified, therefore, it is of interest for this thesis. The MAUP is an issue whereby if data are aggregated to different areal units, the results produced at each areal unit can be different. For this thesis, two areal units will be used, and these are output areas and data zones. These are spatial units created using census information with output areas nested within data zones. The hotspots can also be influenced by the clustering methods used. The four different cluster methods utilised in this thesis are: k-means, finite mixture models (both non-spatially contiguous), Local Moran's I and Getis Ord Gi* (both spatially contiguous). I will look at the different crime clusters / hotspots that are produced by different cluster methods at two different areal unit levels within Strathclyde. I will identify crime hotspots based on the recorded crimes in the Strathclyde dataset provided by Police Scotland (formerly Strathclyde Police). I will examine how the clusters identified are affected by the type of areal units used (output areas and data zones). If these clusterings are seen to be very different this will highlight the MAUP. I will also examine the variation due to the different clustering methods used.

Crimes of Place

The following section explains how crime relates to place and the beginnings of crime mapping where maps were produced to look for areas with high crime. In this thesis, I will use maps as a way to visually represent the crime clusterings at the different areal levels and for the different methods.

Crime Mapping

The first links between crime and place were through the Cartographic School which was influential from around 1830 to 1880. The Cartographic School had an emphasis on mapping crime and looking at the "relationship between society and the physical environment" (Courtright & Mutchnick 2002, 176) thus linking crime and place. The beginnings of statistics being used in criminological thinking began when a national report was produced in France. In 1827, the 'Compte', was published which had each crime included with whether an offender had been caught, charged or acquitted along with a range of other information relating to the offender (Courtright & Mutchnick, 2002). Guerry (1831) and Quetelet (1842) used this document to conduct research into relationships between crime and social factors (Vold & Bernard, 1986; Wolfgang & Ferracuti, 1967).

Guerry used the term 'moral statistics' to refer to the links he was making between crime rates and social factors and he was able to identify areas which were less 'moral' than other areas based on high crime rates thus linking moral thinking with crimes (Schafer, 1969). Guerry used the French statistics to create maps which looked at crimes in relation to social factors leading him to publish the first research on "scientific criminology" (Vold & Bernard 1986, 131). This was the first documented crime mapping. Guerry looked in particular at the link between economic conditions and property crime establishing that high crime rates were often found in the more affluent areas which led him to deduce that property crime occurred here due to there being better goods which could be stolen (Courtright & Mutchnick, 2002). He also researched violent and personal crimes which he discovered were more likely to occur in rural areas (Brantingham & Brantingham, 1981). While Quetelet focused on locational and environmental attributes for causing an individual to commit a crime (Courtright & Mutchnick, 2002). He agreed with Guerry that these statistics could show the moral standards of an area as, if an area had high crime rates, then this would suggest that the moral standards were failing in this area (Courtright & Mutchnick, 2002) thus linking crimes to place. These discoveries would appear to be the beginnings of the social theory that Cohen and Felson would later identify as Routine Activities Theory. Guerry looked into the crime rates in England as well and was able to compare these to crime rates in France providing the first comparative research using criminal statistics (Courtright & Mutchnick, 2002).

Environmental Criminology

With crime being thought of as tied to place, theories which link crimes to place make up a field known as Environmental Criminology which has its roots in the 'environmental backcloth', a theory developed by Brantingham and Brantingham. Brantingham et al. (2009) explained that the backcloth was formed by the environment in which we live:

> "What surrounds us in an urban environment includes centers of activity, roads and pathways, well known landmarks, and parks as well as neighbourhoods with different socio-economic and demographic character. We move around the urban environment from one activity node to another sometimes with fixed location goals (such as a specific restaurant) and sometimes with general area goals (the entertainment district)." (Brantingham et al. 2009, 90).

The Environmental Backcloth links crime to being influenced by environment which suggests that crimes can be thought of as being linked to certain places. Crimes of place then highlight crime hotspots as these crimes will occur in similar places and this can lead to hotspots policing. At its core, the environmental backcloth is all of the environmental elements which combine to influence an individual's behaviour and may cause them to commit a crime (Andresen, Brantingham, & Kinney, 2010). A well-kept park which has a groundskeeper may reduce the number of crimes in the area as this park is seen to have a suitable guardian and thus the chance of being caught is higher. However, a building which is disused and derelict may attract crime to an area as no-one is believed or seen to care about the building and thus no-one is likely to report crime (e.g. vandalism) in the area. Therefore, hotspots analysis can be used to identify areas which could be targeted for interventions to reduce crime.

Social Theories

For certain types of crime it could be assumed that it is a crime of place as opposed to a crime dependent on people (Sherman, Gartin, & Buerger, 1989). There are several social theories that link crime and place such as Routine Activities Theory (RAT), Defensible Space Theory and Broken Windows Theory. These can provide some background to why crime hotspots are important. Routine Activities Theory (RAT) was formulated by Cohen and Felson (1979) in their belief that crime and place are linked. They believed "crime is tied to the characteristics of the environment and to events in time and space" (Courtright & Mutchnick 2002, 179). RAT centres around the idea that in order for a crime to occur there has to be a convergence of:

(1) motivated offenders,
(2) suitable targets, and
(3) a lack of suitable guardians
(Cohen & Felson, 1979).

Felson (1987) believed that the focus could be on the routine activities of place. If the targets were less suitable, or there was an increased guardianship, or the motivated offender numbers were reduced in an area, this could lead to a reduction in crimes (Sherman et al., 1989) suggesting that merely changing an area could reduce crime rates. This would imply that if crime hotspots were identified, further work could research the reason why these areas became crime hotspots and thus interventions could be targeted.

Defensible Space Theory (DST) was developed in the 1970's by Oscar Newman who believed that crime could be reduced by designing the environment in such a way that crime becomes much more difficult to accomplish (Shjarback, 2014). DST has at its core that the perceptions of an area are important as this will influence whether people want to live there and look after an area. Newman believed that residential areas could be set up to link three key components to a neighbourhood being a safe and 'defensible space' are:

(1) territoriality which means having defined barriers either actual or perceived;
(2) surveillance which means having the ability to 'watch over' your area;
(3) image which means the perceptions of the area
(Shjarback, 2014).

A criticism of Defensible Space Theory appears to be Newman's neglect of defining the unit of analysis and continually using the same unit of analysis in his work as he appears to use the term 'space' to mean a number of different areal units such as an apartment complex or a neighbourhood consisting of a number of streets (Reynald & Elffers, 2009). Newman appears to leave the term 'space' as open as possible in order to enable the theory to be applied at multiple levels, from street to apartment to neighbourhood, but this leads to the terms involving territoriality being left ambiguous as there is not a definition for what this means at each different level (Reynald & Elffers, 2009). This shows the importance of areal units highlighting the consequence of the Modifiable Areal Unit Problem (see Chapter 2 for further discussion).

Also, in the 1970's researchers noticed that there was an increase in public perceptions of crime in areas which appeared to be 'uncared for' which could be seen through the physical and social 'signs of incivility' (Hunter, 1978; Taylor & Harrell, 1996). Through examining 'Defensible Space Theory', Wilson and Kelling (1982) established Broken Windows Theory (BWT). This has at its core that if people perceive an area to be a high crime area then they can find it acceptable to commit crimes in this area. They argue that something as 'small' as a broken window or one piece of graffiti in an area can begin a chain reaction which escalates into more violent crime. This can lead to crime hotspots appearing in areas which are seen to be easy targets or areas which do not appear to have suitable guardians which link Broken Windows Theory with Routine Activities Theory.

It is through the Cartographic School that the criminological thinking shifted focus from biological and individual factors to environmental and societal factors which link crime to place. Most of the environmental criminology social theories appear to have three key components linking them. I believe the two main theories are Routine Activities Theory and Defensible Space Theory as all the other social theories link specifically to these ones. In particular, Routine Activities Theory links closely with Defensible Space Theory, through the concepts of increased guardianship and reducing targets in areas to ensure the area remains as crime-free as possible. Defensible Space Theory also links closely with many of the other social theories such as 'Broken Windows Theory' which connect based on decreasing targets and increasing guardianship through the uses of surveillance and planning an area in such a way as to reduce opportunities for crime.

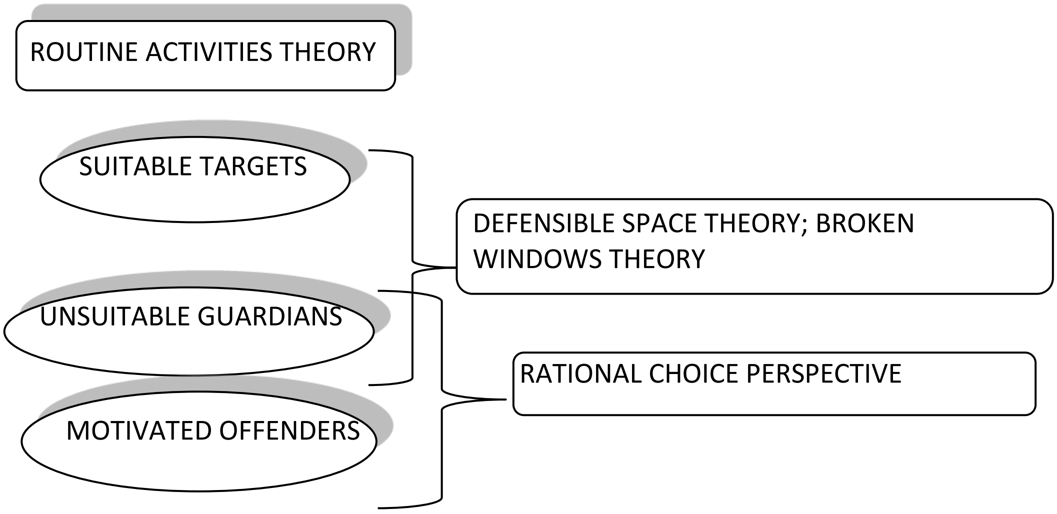

ROUTINE ACTIVITIES THEORY
SUITABLE TARGETS
UNSUITABLE GUARDIANS
MOTIVATED OFFENDERS
DEFENSIBLE SPACE THEORY; BROKEN WINDOWS THEORY
RATIONAL CHOICE PERSPECTIVE

Figure 1.1: Overlap of the social theories related to crime and place

Overview of the other chapters

This chapter has introduced the concept of crime being linked to place. The next chapter (Chapter 2) will look at the Modifiable Areal Unit Problem (MAUP) in detail and how this relates to crime hotspots. I will also look in detail at the reasons why crime hotspots analysis is interesting.   Chapter 3 will look at types of data used in crime studies and will provide an overview of the Strathclyde crime dataset used for this study. Chapter 4 will introduce the clustering methods utilised in this thesis for detecting both non-spatially contiguous clusters using k-means and finite mixture models, and spatially contiguous clusters using Local Moran's I and Getis Ord Gi*. Spatially contiguous means areas are clustered with other areas only if they share a border or are within a pre-specified distance from each other. Spatially non-contiguous means the areas that are clustered together do not need to be neighbours and there are no constraints placed on the areas being near each other. In chapter 5, I will look at the results of applying these four different methods to the 2011 Strathclyde crime dataset at different areal levels. The adjusted Rand index can be used to identify whether the clusterings produced by the different methods at each areal unit are similar. It will then be used to compare the clusterings at the output area level for the clusterings identified at output area and data zone levels and if these are found to be dissimilar, this will highlight the MAUP. Chapter 6 provides a summary and looks at the limitations and future work that could be done.

Chapter 2 - Modifiable Areal Unit Problem and Crime Hotspots

This chapter will look at how clusters and crime hotspots are defined. It will also provide a background to the Modifiable Areal Unit Problem. There will also be a brief overview of why hotspot analysis is important, in particular, looking at how and why police use hotspots analysis.

Spatial Data

Spatial data are any form of statistical data which have geographical locations attached and generally come in three forms:

1) point-referenced data – a set of observations which are taken at certain spatial locations
2) areal data – partitions the overall spatial region into a set of non-overlapping subregions, known as areal units, and aggregates the other covariates at this level i.e. a county split into output areas.
3) point pattern data – spatial data where the location itself is of interest i.e. the aim is to describe the pattern of the locations.

(Anderson, Lee, & Dean, 2014)

Cluster Analysis and Hotspots

A cluster is defined as a grouping of objects which are very similar to other objects within the cluster but different to objects from other clusters. Burns and Burns (2008) define a cluster as "A group of relatively homogeneous cases or observations" (Burns & Burns 2008, 553) where the aim is to minimise the within-cluster differences while maximising the between-cluster differences (Burns, 2008; Gordon, 1996; Kaufman & Rousseeuw, 1990). Figure 2.1 shows three distinct clusters, one at (-2,-2), one at (2,2) and one at (6,6).

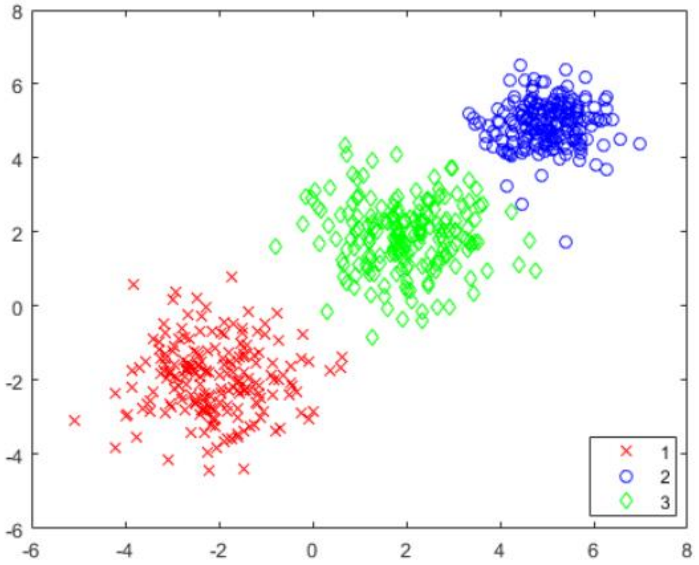


Figure 2.1: Example of three distinct clusters

Cluster analysis has been used in a variety of fields such as anthropology (Driver & Kroeber, 1932), psychology (Tyron, 1939; Zubin, 1938) and banking (Burns, 2008) as far back as the 1930's. It is used in the banking sector to target marketing initiatives by identifying what different groups/clusters of clients are looking for (Burns, 2008).

A hotspot is a cluster which has a higher mean level of the variable being studied compared to other clusters. Hotspot clustering looks for areas on a map where there is an "excess level" of the event being studied (Lawson 2010, 233). Lawson (2010) defined clusters as "where an intensity threshold or level threshold is used and *any* area of a map above the threshold counts as a cluster" (Lawson 2010, 232). Clusters need to be defined in terms of location, size, shape, and 'threshold' intensity values (Lawson, 2010). This appears to be a definition of a hotspot as opposed to a cluster as he discusses intensity of the object being studied.

Hotspot analysis has been used in other disciplines such as disease mapping (environmental causes of cancer (Mason, McKay, Hoover, Blot, & Farumeni, 1985)), transportation (vehicle fatalities (Baker, Whitfield, & O' Neill, 1987)) and ecological science (Kumar & Chandrasekar, 2010). Indeed, for transportation, accident hotspots are used by many insurance groups to identify areas where there is an increased likelihood of accidents happening and many maps are produced which enable the general public to see where clusters of accidents occur (MCE Insurance, n.d.; Which, 2013). This allows people to plan routes avoiding these areas which could lead to fewer accidents happening and fewer insurance claims which explains why accident hotspots are mapped by insurance companies. Based on the context of this thesis, crime hotspots can be identified as clusters of areas which have a high mean crime count or rate.

Several studies have looked into the heightened risks associated with nearby locations to recent crime events (Ratcliffe, 2010). Links between the risk of burglary to not only the house which has been burgled but to nearby houses was looked at in the UK and in Australia in studies by Bowers and Johnson (2004); Johnson and Bowers (2004a; 2004b); Townsley et al. (2003) (Ratcliffe, 2010). Near repeat patterns were found in shootings in Philadelphia in a study by Ratcliffe and Rengert (2008) (Ratcliffe, 2010). Townsley et al. (2008) looked at the location of IED's in Baghdad to discover if locations near other IED's were likely to be at a heightened risk of having IED's (Ratcliffe, 2010).

Hotspot analysis or crime mapping can trace its origins in crime analysis to moral statisticians Guerry (1831) and Quetelet (1842). Guerry (1833) and Quetelet (1842) provided a comprehensive analysis into the crime rates within French provinces (Wortley & Mazerolle, 2008) thus distinguishing between areas which had high crime rates and low crime rates. Shaw and McKay (1942) used crime mapping to look at juvenile delinquency in Chicago. Since then, new methods have been developed which overcome some of the previous issues of crime mapping such as technological and data limitations (Maltz, Gordon, & Friedman, 1991; Weisburd & McEwan, 1997), organisational issues (Openshaw, Cross, Charlton, Brunsdon, & Lillie, 1990), the inability to convert digital addresses to maps (Bichler & Balchak, 2007; Harries, 1999; Ratcliffe, 2001, 2004a) and functional obstacles including police databases not set up to record the crime location in a usable format (Ratcliffe & Mccullagh, 1998; Ratcliffe, 2010).

Ratcliffe (2004b, 2004a) argued there were 3 spatial event categories:

(i) dispersed (no pattern),
(ii) clustered (happens at one part of a hot street), and
(iii) hot street (crime consistently happens over and over).

These ideas provide the foundation for the idea of crime hotspots. However, the term 'crime hotspots' is usually first associated with the Sherman et al. (1989) article which looked into predatory crime hotspots.

Often the easiest method to identify a cluster or hotspot is to look at the data visually, however, due to datasets becoming larger, this is not always possible (Burns, 2008). Therefore, statistical techniques are needed to identify clusters if the dataset is too large. This leads to cluster analysis being used to identify the hotspots as this enables police to allocate resources and use pro-active policing as opposed to reactive policing (Grubesic, 2006). First the data are clustered and then the clusters with high mean levels are identified as hotspots.
Clustering can be carried out using both spatially contiguous methods and spatially non-contiguous methods. Spatially contiguous means areas are clustered with other areas only if they share a border or are within a pre-specified distance from each other. Spatially non-contiguous means the areas that are clustered together do not need to be neighbours and there are no constraints placed on the areas being near each other. Note that after clustering, you can separate non-contiguous clusters into multiple spatially contiguous clusters and vice versa.

Modifiable Areal Unit Problem

Whenever a study looks at spatial/areal data, the modifiable areal unit problem or MAUP must be considered. The MAUP means that depending on how the spatial data are aggregated, the results produced can be different. In terms of crime hotspots analysis, this means that if the areal data are aggregated to data zone level for example, a relatively coarse partition of the region, this can hide hotspots which lie at the output area level, a finer partition. A data zone might overall have an average low-medium crime level but if the same data were studied at output areal level, this could be made up of several low crime areas and only one specific high crime output area which would be a crime hotspot. Thus, the level of aggregation used for the spatial data can cause hotspots to be hidden. This means it is very important to define the resolution of the areal units which will be used for the study and why these areal units are chosen.

The modifiable areal unit problem (MAUP) was first identified in 1934 by Gelke and Biehl but it was Openshaw in the 1980's who made the issue much more prominent (Lembo Jr, Lew, Laba, & Baveye, 2005; Manley, Flowerdew, & Steel, 2005). MAUP can cause issues for any research which uses physical geographical locations as it arises where an analysis carried out on the same data could produce different results depending on how the data is split into 'neighbourhoods' or 'areal units' (Manley et al., 2005). Openshaw (1984) identified the term 'areal units' to mean a geographic area which is bounded clearly and which could have data recorded in it (Manley et al., 2005). This, therefore, links directly with the issue of hotspots and scale as different hotspots can be identified depending on the scale which is used by the researcher.

Indeed, Harries (1999) believed that the main issue with the MAUP (Openshaw, 1984) was that a localized hotspot might not register at a regional level (Grubesic, 2006). Indeed, Grubesic and Murray (2001) argue that it is the scale of the data used that is key to identifying hotspots.

Examples of MAUP Studies

Since Quetelet (1842) and Guerry (1833) looked at country level data, there has been an increasing movement to look at crime at smaller area levels such as Glyde (1856); Burgess (1916); Shaw and McKay (1931; 1942); Sherman et al. (1989); and Weisburd et al. (2004 and Weisburd et al. (2009) who all looked at crime at street-level. Sherman et al. (1989) believed that there was evidence that in some bad neighbourhoods there were locations which were never involved in crime and likewise, in some good neighbourhoods there were locations which were crime hotspots (Andresen & Malleson, 2013). The ideal is homogeneity between all the smaller units of space within a larger unit of space i.e. all underlying smaller units would have the same value as the larger unit. In order to identify homo- or hetero-geneity in these smaller units, more than one unit of analysis needs to be looked at in order to identify if the underlying units are the same as the larger spatial unit (Andresen & Malleson, 2013). However, this can be difficult if the data is only available at one spatial level such as county level, or not required if the aim of the study is to replicate previous study or the research question is focused only on one specific scale of interest (Andresen & Malleson, 2013).

There are three potential outcomes when spatial heterogeneity/MAUP is investigated and these are no impact (no statistically significant differences); a quantitative impact (overall results remain the same); and a qualitative impact (this leads to incorrect statements and potentially false results) (Fotheringham & Wong, 1991). MAUP can lead to qualitative issues when inference about a population is based on analysis at one spatial scale and then applied to another spatial scale (Andresen & Malleson, 2013). When analysis is carried out at a larger scale and then applied to a smaller scale this can lead to the ecological fallacy (i.e. that which is said to be true of the whole is not necessarily true of the parts) and similarly when analysis is carried out at a smaller scale then applied to a larger scale this can lead to the atomistic fallacy (Andresen & Malleson, 2013; Dark & Bram, 2007). This means that explanatory variables can have different effects on crime depending on the level of aggregation (Hipp, 2007; Ouimet, 2000). The ecological fallacy led to the decline of interest in geographic criminology in the 1950's through to the 1980's (Bernasco & Elffers, 2010; Robinson, 1950).

Mitigating MAUP

There are many ways in which the MAUP can be reduced. Openshaw (1984) first identified four distinct solutions of which there are realistically only three as the original solutions ii) and iii) are very closely linked and have been combined to form solution ii). These issues are

i) ignore the issue;

ii) correctly identify the scale at which to analyse the data and investigate each individual variable separately; and

iii) structure the hypothesis in such a way as to take account of MAUP
(Dark & Bram, 2007).

With solution i) the issue would still persist which I believe makes this an unsatisfactory solution. Solution ii) requires each variable to be looked at individually and could be time-consuming as it would mean identifying an areal unit for one variable and then, perhaps, a different areal unit for the next variable. This means that areal units of analysis would need to be chosen carefully which linked with the questions being investigated. An investigation would be required to decide which areal unit is "best". However, if only one variable is being looked at then this would prove a good option. Solution iii) would be reasonable provided this is carried out in the correct way, i.e. the hypothesis is stated before any analysis is carried out (a priori). However, as cluster analysis is an exploratory method, this would not usually involve a hypothesis being specified before analysis is carried out as the aim would be to generate a hypothesis based on the clustering.

Other solutions have also been put forward by Fotheringham (1989) and Tobler (1979) but there remains no single solution with which to minimise the issue of MAUP. The MAUP is only a true issue in datasets where the data is analysed at multiple scales and it is this that can produce conflicting results. Provided the spatial units are sensible and meaningful to the data being studied, then MAUP should not be a major concern. This thesis will look into how the MAUP affects the clusters identified at two spatial scales.

Why study Crime Hotspots?

Hotspots Policing

Hotspots policing is the focus of police resources on crime being linked to place instead of the traditional policing route which focused on people (Weisburd & Telep, 2014). It is also referred to as place-based policing because of its focus on crimes being tied to place (Weisburd & Telep, 2014). It leads to resources such as interventions or officers being deployed in high crime areas (crime hotspots). The interventions/resources are targeted based on the individual area's needs (Weisburd, 2008). One of the first studies which looked at hotspots policing was Sherman & Weisburd (1995) in their study of the Hot Spots Patrol Experiment in Minneapolis.

Several articles show the importance of hotspots policing and the focus on small areal units to identify crime hotspot areas (Weisburd & Telep, 2014). The aim is to look at small areal units and identify areas with higher levels of crime and then target police interventions in these areas. The smaller the areal unit the better for these types of analysis which is due in part to the MAUP as if the areal units are too large, then there could be unidentified crime hotspots nested within these larger areas. Therefore, the focus on crime hotspots analysis is to ensure that the areal unit of analysis is small enough to enable hotspots to be identified. However, unless analysis is carried out at each individual crime location, there will always be aggregation carried out and thus the MAUP will be of concern. It can be seen that police maximise their effectiveness when the focus is on micro-units of geography (Weisburd & Telep, 2014) which suggests the importance of using the correct areal unit for any analysis.

An argument against hotspots policing is the displacement aspect as if crime interventions are targeted in one area this can lead to crime moving to other nearby areas. However, there are studies which show a lack of supportive evidence for this (Weisburd & Telep, 2014). In order to see evidence of this, it is important then to look at hotspots in order to identify key spatial areas which could be targeted for interventions. The use of clustering methods to identify high crime areas and also lower crime areas would be important to identify if displacement has occurred i.e. a high crime area has become a medium crime area but the surrounding areas then becoming a high crime area.

Previous Hotspot Studies

Peter St. Jean (2007) studied random police beats in Chicago where he established that while an area was considered a high crime area, there were certain blocks (five blocks) within that area (of 59 blocks) in which most of the crimes occurred (60% of narcotics crimes, 53% of robberies, and 44% of assaults/batteries) (Bottoms, 2012). St. Jean's aim was to establish why hotspots existed within high crime areas in these particular blocks from the perspective of the offender (Bottoms, 2012). He found that in particular there were two crime hotspots which occurred at the busiest intersections in the area which would suggest that the convergence of a large number of people led to an increase in the crime rate (Bottoms, 2012). This would link with both the theories of Routine Activities Theory and Rational Choice Perspective. A lot of people would use these intersections in their daily commute to work/leisure or other pursuit in which case they would routinely be passing through these areas and offenders would know there were a large number of people in these areas who they could target. This suggests that the convergence of suitable targets, motivated offenders and lack of suitable guardians at these intersections (Routine Activities Theory) and the knowledge the motivated offenders had that these were busy intersections where there would be 'easy' targets led to the crime hotspots occurring at these intersections.

Eck and Weisburd (1995) and St. Jean (2007) believe it is very important in crime studies to look at micro units of place (Bernasco & Elffers, 2010). The Minneapolis study carried out by Sherman et al. (1989) also highlighted the importance of data aggregation. In areas which were deemed to be high crime neighbourhoods, approximately only twenty per cent of places within these neighbourhoods were crime hotspots (Sherman et al., 1989). The hotspots all seemed to be focused on main routes and were quite close to each other (Sherman et al., 1989). This suggests that even in high crime neighbourhoods, the vast majority of areas are actually relatively safe.

The Manchester study which was conducted looking at Manchester's 'gay village' highlights the issue of scale in hotspot analysis (Skeggs, Moran, Tyrer, & Binnie, 2004). This study looked at crime statistics from the 'Village' and the surrounding area and through these statistics, the 'Village' area was identified as a crime hotspot (Skeggs et al., 2004). However, there were many areas in the 'Village' which were not high crime areas and indeed, the central area was very safe and it was the area of the village closest to the boundary that was the least safe (Skeggs et al., 2004). This would suggest that in fact the 'Village' is actually a safe area and really it is the boundary area which is the hotspot area. Thus it is extremely important when carrying out hotspot analysis to try to ensure that the scale used is as small as possible to stop areas being labelled as hotspots when in fact it is a particular street within this area where all the crimes occur.

Main Place-Based Policing Initiatives in Glasgow (2011)

It is because of place-based policing that crime hotspot analysis is widely used so police can identify areas where many crimes are occurring and target resources and interventions here. I will briefly discuss three of the initiatives which cover this time period (2011).

Anti-Social Behaviour (ASB) Initiative

Although crime had begun to fall from 1991, certain crime types such as petty assaults and anti-social behaviour had continually risen (Audit Scotland, 2000). Due to this rise and the stresses this caused on the police force, Community Safety Partnerships were set up to tackle petty assaults and anti-social behaviour crime levels (Audit Scotland, 2000). These Community Safety Partnerships (CSP) were set up in 1999 and link police with the local authority, health boards or trusts, voluntary organisations and the fire service with the aim to reduce the crime problems by making areas safer (Audit Scotland, 2000). Through this multi-agency approach to promoting safety within communities, specific local problem-areas can be targeted to reduce crime (Audit Scotland, 2000). However, most CSPs did not carry out a full analysis of the area using available data so there is no real base level from which to show progress and identify the areas at which to target the interventions (Audit Scotland, 2000). The aim is long-term solutions to crime problems to ensure that the area remains safe (Audit Scotland, 2000). The targeting of interventions specific to local areas and spread across multi-agencies is useful (Audit Scotland, 2000) as most local agencies can identify where a major crime hotspot lies allowing them to target interventions there. However, they could benefit from a more formal analysis which could be carried out to, perhaps, identify other potential secondary hotspot areas which could also be targeted.

Violent (Knife Crime) Initiatives

Knife crime had been steadily increasing in the Strathclyde region since the late 1970's (Bleetman, Perry, Crawford, & Swann, 1997). 'Operation Blade' concerned a knife amnesty, intensive stop and search procedures, CCTV and metal detectors and a high profile media campaign highlighting the amnesty (Bleetman et al., 1997). This links with Rational Choice Perspective as this believes that the offenders will make a conscious decision not to carry a knife or carry out an assault based on their increased likelihood of being caught as they know that police are focussing on this crime type. This initiative was evaluated through looking at admissions to hospitals before and after 'Operation Blade' was introduced (Bleetman et al., 1997). This showed that the initiative initially appeared to be successful in reducing violent assaults which was a similar result to that shown by police data (Bleetman et al., 1997). However, within a year the number of violent assaults had begun to rise suggesting that this initiative was not fully effective in the long term (Bleetman et al., 1997).

To identify if reductions in crime had been made this could be carried out by using crime hotspots analysis to identify the main areas and focusing the resources specifically on these areas in particular. The only issue with using hotspot analysis for this crime type could be that some people may turn up at hospital with knife wounds and when these are reported to the police they are reported from the hospital with no other crime address given. Thus the results could be skewed towards hospitals appearing as knife crime hotspots where they are only treating the after-effects of knife crime. This is seen in the data which Bleetman et al. (1997) used, the actual location of the offence was not recorded for over half of the assaults processed.

This led on to the “No Knives Better Lives” (NKBL) Campaign designed by the Scottish Government as a long-term preventative approach to dealing with knife crime amongst young people (The Scottish Government, 2010). The aim was to deter young people from carrying a knife by showing them the consequences of carrying and using a knife (The Scottish Government, 2009b). By showing the penalties, the hope is that young people will rationally decide not to carry a knife as the consequences can be severe.

Crimes were identified as having wider costs to the public (than just being a victim) such as public health costs and policing costs and if these crimes could be reduced, this could enable services to be used for other purposes (Tanner, 2014). The Violence Reduction Unit (VRU) was established by Strathclyde Police in 2005 (Tanner, 2014) and is a collaborative approach with interventions used by the VRU focusing on enforcement and attitudinal change (Tanner, 2014). They often involve going into schools to try to stop young people particularly ‘at-risk’ from becoming offenders (Tanner, 2014) linking with prevention being better than reactive policing.

An initiative by the VRU was the Weapons and Public Space campaign which launched in 2010 (Tanner, 2014). This led to “670,000 stop and searches, including 12,000 visits to licensed premises… police seized 447 knives and charged 478 individuals with possession of a knife” (Tanner 2014, 4). This suggests the campaign was successful as over four hundred knives were taken off the streets. This could also lead to more hotspots being identified for this crime type within Glasgow during this period as the focus of the police has been on targeting this crime type.

Deterrent Initiative

CCTV is used in town centre initiatives for a number of reasons such as “deter criminals and disruptive groups from intimidating the public; to reduce organised crime especially where gangs of shoplifters, pickpockets and drug dealers carry out such activities in town and city centres; to detect anti-social and public order offences; to help convict offenders through the provision of high resolution images; to increase the public’s sense of safety; and to provide a greater sense of commercial security for the retail and business community” (Harris et al. 1998, 161). The introduction of CCTV cameras can also cause a hotspot to appear as people are suddenly caught shoplifting as they can be seen by the new CCTV camera. This means that the hotspot may have always existed but until the camera was placed, it did not appear to be a hotspot as not all the crimes which happened in this area were recorded. Also, hotspot analysis can be used if there is an area where a lot of crime happens, CCTV cameras can be introduced there which can stop people from offending in this area as their chance of being caught has increased greatly.

Ex-Police Analyst Interviews

I carried out semi-structured interviews with two ex-police analysts to provide an understanding of the type of crime hotspot analysis which is carried out by police analysts. The interviews were used to identify key aspects of crime hotspot analysis in order that it can be compared with other hotspot clustering methods. From the outset of the first interview it became apparent that hotspot analysis was used routinely by police analysts to identify areas of interest to operational officers. Hotspot analysis is considered throughout this section to refer to the Hotspot Analysis carried out in ArcGIS using Getis Ord Gi* as this is what is used by police. Throughout the course of the interviews there were four key areas that were of interest. These were (i) crime types used; (ii) how crime hotspots were defined; (iii) methods used to identify them; and (iv) why police use hotspots analysis.

Crime Types used in police analysis

Interviewee A explained that the main crime types that were used were any high-volume crimes as these were of particular importance as they were likely to have the biggest impact on the local communities. This meant that the most common crimes for this type of analysis are "Anti-Social Behaviour (ASB) disorder, vandalism and associated crimes such as house-breaking, burglary, robbery and violent crime". Prostitution would also be looked at, however, for the purposes of this thesis, sexual crimes were not part of the available dataset and will not be included in any analysis. It was also identified through interviewee A that police would only look at certain crime types that would be associated as crimes of place. This suggested that violent crime can be of particular importance in an urban setting but may be less of an issue in rural areas, leading the hotspot analysis to only be carried out for this crime in an urban setting. This highlights the importance of certain crimes being considered crimes of place. If crimes are always occurring in the same area this can provide evidence of Broken Windows Theory and Routine Activities Theory as people can associate these areas with crime and, therefore, believe that crime is "acceptable" in these areas.

Interviewee B again suggested that the crime types used were very dependent on the area being looked at. If the area was a city/town centre, then the most likely crimes to be used for hotspot analysis would be alcohol related crimes. Usually the crimes that would be used were "outside crimes, quite visible" which suggests that place and crime are strongly linked as certain types of crimes are more likely to happen in certain areas due to their nature.

Spatial units used in police analysis

Interviewee A also identified that crime hotspot analysis was more likely to only occur in urban areas as there were less clusters of crime in rural areas. This meant that for hotspot analysis, the main areas to focus on would be the city centre and town areas of Strathclyde. According to Interviewee A, hotspot analysis was mainly carried out on crime counts without converting to rates. Interviewee A believes this was due to the fact that the census is only carried out every ten years and, therefore, population data can be out of date, particularly if it is near to the next census being carried out.

Different time-scales were used by Interviewee A depending on the "specific tasking" and sometimes would consist of issues raised by the community or by officers themselves. This means that hotspot analysis is very subjective across divisions. Interviewee A also recognised that the time-scales that were used could be anything from weeks to financial years depending on the reasons for which the analysis was being carried out.

Interviewee A acknowledged that spatial scales used depended on who had asked for the analysis to be carried out as these could be at divisional level if the whole of Scotland was being compared, or could be at sub-divisions, local authority or multimember wards. The hotspot analysis would be carried out on recorded crime. Interviewee A also identified that incidence crimes can also be relevant as well particularly for ASB incidents. There are two main databases which are used by the police, the Crime Database and the Incidence Database. Also, within crimes there are detected and undetected crimes (detected means that a perpetrator has been identified). Interviewee B also identified that sometimes, the datasets which the analysis was to be carried out on had to be amended or were unsuitable in certain areas as the default region was too large for meaningful hotspot analysis to be carried out.

Methods used in police analysis

When asked about the particular methods used, both interviewees acknowledged that there were no specific statistical clustering techniques employed. They identified software GiS as being the main source of hotspot analysis techniques. This software enabled them to carry out the hotspot analysis using Getis Ord Gi* and Local Moran's I methods. This is why we use these methods and compare them to other statistical clustering methods. However, it was identified during the course of the interviews that these methods were chosen as they were inbuilt to the software and there had been no particular theoretical preference given to these methods over other clustering techniques. Default settings tended to be used which enabled buffers to be used which would use, for example, 1km^2 areas. Interviewee A suggested that this meant that sometimes manual interpretation would be used on the output to focus on smaller areas as sometimes Scotland wide maps which were produced would not provide adequate detail for local officers to use as there would just be a "sea of colour". Interviewee B stated that there were some "glitches" with the datasets as, for example, there was a lot of manual intervention required such as completed postcodes information fully to ensure they were in the correct format to enable the analysis to be carried out.

Throughout both interviews, the subjective nature of hotspot analysis was identified with both referring to this as an "art" that is mostly standardised but very open to adjustments. Training on the GiS tool was standard but there was no statistical knowledge employed to use other statistical methods such as k-means. Interviewee A also identified that sometimes local knowledge could be applied and this meant that the analyst could look at a spreadsheet and just identify that there was a pattern of higher crimes occurring in one area compared to another.

Why police use hotspots

Interviewee A also identified that there are four main reasons for hotspot identification to occur:

i) Community – there is a particular issue facing a part of the community who will raise it with the police, so there may be more notification to the police of certain crimes in an area
ii) Performance – there might be particular interest in one area where a target is not being met and there is analysis required to investigate which crimes are occurring in an area

iii) Senior managers – there could be interest/speculation in the media regarding crimes in other areas, and the senior managers may feel that their area would benefit from analysis being carried out into similar crime patterns
iv) Legislation – should a recent legislative change occur, it would be of interest to compare crimes before/after this came into effect.

Sometimes, crime hotspot analysis can be used to back up claims that officers hold which will enable the correct resources to be systematically targeted at affected areas.

Interviewee A believed that hotspot analysis was carried out on an ad-hoc basis as, although most of the time, local divisional analysts would create datasets and interpretation based on the numbers of crimes, this would not necessarily all be considered hotspot analysis. In the second interview, Interviewee B also identified that a further reason for carrying out hotspot analysis was in pilot studies to identify if crime rates had decreased as a result of preventative measures taken. Interviewee B identified that the main interest in hotspot analysis was identifying problematic areas which could be targeted by police officers on patrol when they were not responding to other calls. This means that it would be seen as being a preventative and reactive measure as it is once these areas are identified as being problematic (reactive) that more focus is placed on these areas to reduce the levels of crime in these areas and remove the crime hotspot from this area.

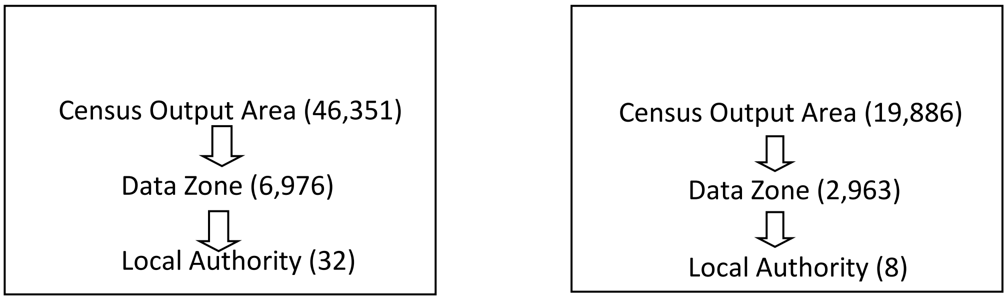
Although the way in which police analysts use hotspot analysis has remained constant, previously, it would be police officers who were interested in where the majority of crimes were occurring in their area. However, now it would tend to be a focus due to performance and targets. This means that it would be easily identified if an area suddenly had more crimes occurring within it.

The next chapter will look at ways of defining spatial units and types of crime data. I will then give an overview of the study area (Strathclyde).

Chapter 3 – Crime Data

Areal Units

Spatial datasets as described in Chapter 2 are split into aggregated areal units. This section will highlight the main type of areal units for Scotland and Strathclyde. Scotland is split into many different types of areal units, most of which are aggregated from smaller areal units. The smallest areal unit is postcode unit (e.g. G12 8QQ) which can then be aggregated to different areal units up to the largest areal unit, Local Authority (The Scottish Government, 2006). The most regularly used areal units are output areas (46,351), data zones (6,976) and local authorities (32) (The Scottish Government, 2011). This is summarised in Figure 3.1,



(a) Scotland areal splits (b) Strathclyde areal splits

Figure 3.1: Areal splits for Scotland and Strathclyde

Postcodes are very small areal units and are created by The Royal Mail to enable post to be delivered accurately and quickly (Scottish Neighbourhood Statistics, 2007). Postcodes can be further split into Postcode Units, Sectors, Districts and Areas each of which is nested within the previous area unit. Postcode areas are nested within postcode districts which are nested within postcode sectors and these are nested within postcode units. These are updated regularly due to new houses being built and other houses/commercial premises being demolished in order to keep the number of households within a postcode at around fifteen (Scottish Neighbourhood Statistics, 2007). As a result, they are rather inconsistent over time and will change quite considerably over time (Scottish Neighbourhood Statistics, 2007). Therefore, while postcodes would enable micro-analysis to be carried out, their longitudinally inconsistent nature leads to them not being selected for use in this thesis.

Output areas are used to give micro-scale analysis while still maintaining continuity over a number of years and also ensuring there are enough non-zero observations that hot spots can be detected. Output areas were created as an aggregation of postcodes (Office for National Statistics, n.d.). They contain an average of 50 households (250-375 people) with a minimum requirement of 20 households (at least 50 people) in each output area (Office for National Statistics, n.d.) As postcodes change much more frequently than output areas (i.e. when new houses are built, old derelict buildings knocked down) the output area level is as small an areal unit as is possible to maintain stability in the analysis. There is a total of 46,351 output areas across Scotland, of which there are 19,886 in the Strathclyde area. There are only a small number of output areas where no crimes were recorded for different crime types. If there are too many zero values this can cause issues for standard models and more complex techniques can be required. This means that to go to a smaller areal unit level would make it difficult to use standard models as there would be an increase in the number of zeros. Therefore, output areas were chosen as the smallest unit of analysis.

Output areas are then nested within data zones. In total, there are 6,976 data zones across Scotland. Of these 6,976 data zones, there are 2,963 that are located within the Strathclyde region which I will be studying. These data zones were created from information from the 2011 census and there were criteria identified to ensure data zones identified were appropriate (Scottish Neighbourhood Statistics, 2004). “The following criteria were taken into account in the definition of data zones, in approximate order of importance,

1) Approximate equality of population, between 500 and 1,000 people;

2) Compactness of shape;
3) Approximate homogeneity of social composition;
4) Existence, where possible, of some community of interest;
5) Accordance with other boundaries of local significance; and
6) Accordance with prominent features in the physical environment." (Flowerdew et al. 2004, 11)

All information produced in this thesis has been checked to ensure the data zone level enables confidentiality to be maintained. Local Authorities were created by the Local Government Boundary Commission for Scotland (Scottish Neighbourhood Statistics, 2007). They cover too large an area for meaningful hotspot analysis to be carried out and thus are not used for this thesis.

To identify crime hotspots, it is necessary to use an areal unit that is small enough to enable hotspots to be seen clearly. There are less confidentiality issues as there are between 500 and 1,000 people living within a data zone and 250 to 375 people living in an output area which means that there are fewer possibilities of individuals being identified compared to postcodes in which there are only fifteen households (Scottish Neighbourhood Statistics, 2007; The Scottish Government, 2006). Also, data zones and output areas are preferable as they were created to be as homogeneous as possible and to maintain a regular shape in line with local physical boundaries and communities with output areas fitting inside data zones (Flowerdew et al., 2004; The Scottish Government, 2006).

Sources

There are many different data sources for crime statistics. The four main sources are police call data, police recorded crime, victimisation studies and self-report studies. The next section details the differences between these.

Call Data

Police call data is the information taken down when someone makes an emergency call to the police services. All calls are recorded thus this is "unfiltered" data as every call is recorded despite some possibly not warranting police attention as no crime/offence is deemed to have occurred after investigation.

Police call data is occasionally preferred to police crime reports / recorded crime as the recorded crime reports can sometimes only provide locations and dates of offences whilst call data can sometimes give the location and the time of day of a call (Pierce, Spaar, & Briggs, 1984). There is no pre-selection of crimes to be recorded as all calls are recorded which includes crimes where the victim does not come forward but the crime is reported by a 'witness' in which case the call can be the only record of the offence (Sherman et al., 1989). However, the 'witness' can provide an incorrect description of the crime which leads to it being coded falsely (Barthe & Stitt, 2009). Another issue with using the call data to identify hotspots is that human error can mean that the wrong address is recorded. Particularly if the caller is not the victim, the address could be coded as the caller's address and not necessarily the address where the crime/offence took place (Buerger, Conn, & Petrosino, 1995).

Also, crimes can be reported after a period of time has passed and they can be recorded as occurring later than they actually did which can lead to issues if using the data to look for hotspots based on their time period within a day (Sherman et al., 1989). There can be certain locations which are more likely to report crimes (for example hospitals) even although the crime will not have occurred there. The crime can be recorded as happening there which leads to over-reporting in certain locations. This suggests the importance of looking at small-area level data as once a hotspot is identified, it can be checked to ensure that it is not the location of a hospital or other place where crimes did not occur but were recorded. As always there is no 'true' count of crime (Biderman & Reiss, 1967) but call data can be one of the best estimates. It is extremely difficult to gain access to call data, so while it may be preferable to reported crime data, it is recorded data that is regularly used.

Recorded Crime

This thesis uses recorded crime data as this was provided by Strathclyde Police. Recorded crime is the information logged by police once there has been verification that an actual crime/offence has taken place. These are filtered usually to individual types of crime that can be easily aggregated to other crime groupings. Recorded crimes are usually the official statistics produced by central police services.

One of the issues with recorded crime data is that 'official statistics' have no way of recording every single crime which is committed as often crimes are unreported (Coleman & Moynihan, 1999). Also, sometimes 'official statistics' can lead to certain crime types being targeted in an area such as speeding offences. This can mean that on one day many people are caught in the same area but throughout the rest of the year there are fewer crimes recorded in this area. This can lead to a slight distortion in hotspot analysis (from both under- and over-reporting) taken from 'official statistics' as this area may not be any worse than the street next to it for drivers speeding. However, because the police have chosen to conduct their speed checks in that street, it appears as a hotspot yet the street before and the street after are not recorded as such. Yet it can be suspected that if someone is caught speeding on one road, they are likely to be speeding on other roads nearby too.

There is also the concern that just because a crime is reported to the police it does not mean that it is necessarily recorded by them for reasons such as insufficient evidence to prove the crime/offence took place (Coleman & Moynihan, 1999). Also, crime is recorded in the wider social and political context because resource allocations are usually based on recorded crime data. This can lead to targeting certain crimes/offences in order to gain more resources to help tackle this crime.

Recorded crime can also be problematic as definitions in legislation can change leading some acts to go from being a minor offence to becoming a more major crime (Coleman & Moynihan, 1999). Also, new legislation can mean new crimes are defined which have previously not been recorded. This does not mean they have just appeared only that the act that defines them is now seen as being a crime (Coleman & Moynihan, 1999). This can also work in reverse as changes to legislation can lead to past crimes no longer being considered as crimes e.g. the laws banning homosexuality which existed in the UK until the 1960's. In the full dataset provided to me for this study, the years 1999 to 2013 were included but in the year 2012, there were a number of legislative changes. These changes caused the reclassification of several crimes/offences and new crime/offences such as the Offensive Behaviour at Football Act. To avoid an issue with the reclassification of crimes, a single year was chosen due to the changes in crime across years. Also, in order to see investigate the MAUP only one year needs to be studied. The year 2011 was of interest for this thesis due to it being the most recent census year. Data zones and output areas were reclassified in 2011 due to the census results and this also meant that population records were the most up to date.

Victimisation Studies

Victimisation studies are qualitative studies that are carried out asking if people have been a victim of a crime, usually over a 12-month period (but this can change depending on the study). This asks people to report if they have fallen victim to a crime and then records details about that crime.

There are some who argue that victimisation studies give a better view of crime problems (Walklate, 1989). This can cause differences in observations for particular crime types which are likely to be unreported in 'recorded police crime data' such as sexual offences or domestic abuse (Walklate, 1989). Indeed people can be less inclined to admit to being victims of such crimes due to the nature of the survey as people may report it to police at the time of the act but can be unwilling to admit such a personal issue to a stranger later (Coleman & Moynihan, 1999). Victimisation Studies can also be problematic as people may not always remember a crime if it did not cause a significant impact to them or they may not remember when it has occurred (Coleman & Moynihan, 1999). Another issue may be that people may not classify an act as a crime against them when it has been (Coleman & Moynihan, 1999). This means that there is still a problem with under-reporting even in victimisation studies.

Self-Report Studies

Self-report studies are also usually qualitative studies which are similar to victimisation studies but instead of asking people if they have been a victim of a crime, asks people to report if they have committed a crime. This asks people to answer honestly if they have committed any crimes or offences but this can be problematic in terms of reliability of answers.

Self-report studies have been used in the US and the UK with differing results (Coleman & Moynihan, 1999). They involve the offender admitting to crimes that they have committed (Coleman & Moynihan, 1999). These can help with identifying previously unrecorded crimes as an offender may admit to more crimes than they have been convicted for thus helping to eliminate some of the 'dark figure of crime' (Coleman & Moynihan, 1999). However, there may be concerns regarding the reliability of the offender given that by admitting to a crime, they may worry that they will receive a penalty for this (Coleman & Moynihan, 1999). They tend to mostly be used in the US as there is an increasing focus on the causes of crime based on the offender (Coleman & Moynihan, 1999). Whilst in the UK, they tend to be overlooked in favour of victim surveys as the emphasis in the UK is on the impact of crimes and those affected (Coleman & Moynihan, 1999). Self-report studies have highlighted that there are no key characteristics of offenders and indeed it is not just a small minority of the general population who offend (Coleman & Moynihan, 1999).

All four of these types of data: call data, recorded crime, victimisation studies and self-report studies measure very different aspects of crimes (Coleman & Moynihan, 1999). Call data is any offence reported by a witness/victim to the police, the focus of recorded crime is taken from the police about the offence, the emphasis of victimisation studies is on the impact of the offence itself and the motivation of self-report studies is on the offender (Coleman & Moynihan, 1999). Therefore, these datasets all likely include some under-reporting and all are from a different viewpoint (Coleman & Moynihan, 1999) but they could be combined to help gain a better picture of the 'dark figure of crime'. For my thesis, I was able to access recorded crime data provided by Police Scotland (formerly Strathclyde Police).

Patterns of Crime: Scotland

Figure 3.2 shows the recorded crimes in Scotland from 1998 to 2012 split by crime and offence categories. This suggests that there is an overall similar pattern in recorded crimes and offences over this time period, with crimes slowly decreasing and offences showing an upward trend. There is a substantial dip in 2008 and then an increase in 2009 for the total number of offences. This could be due to new legislations which caused more actions to be classed as offences. The vertical line shows the 2011-12 financial year which is the study year for this thesis. This shows that the study period is the middle of an overall decreasing crime trend while offences have risen slightly but are starting to decrease.

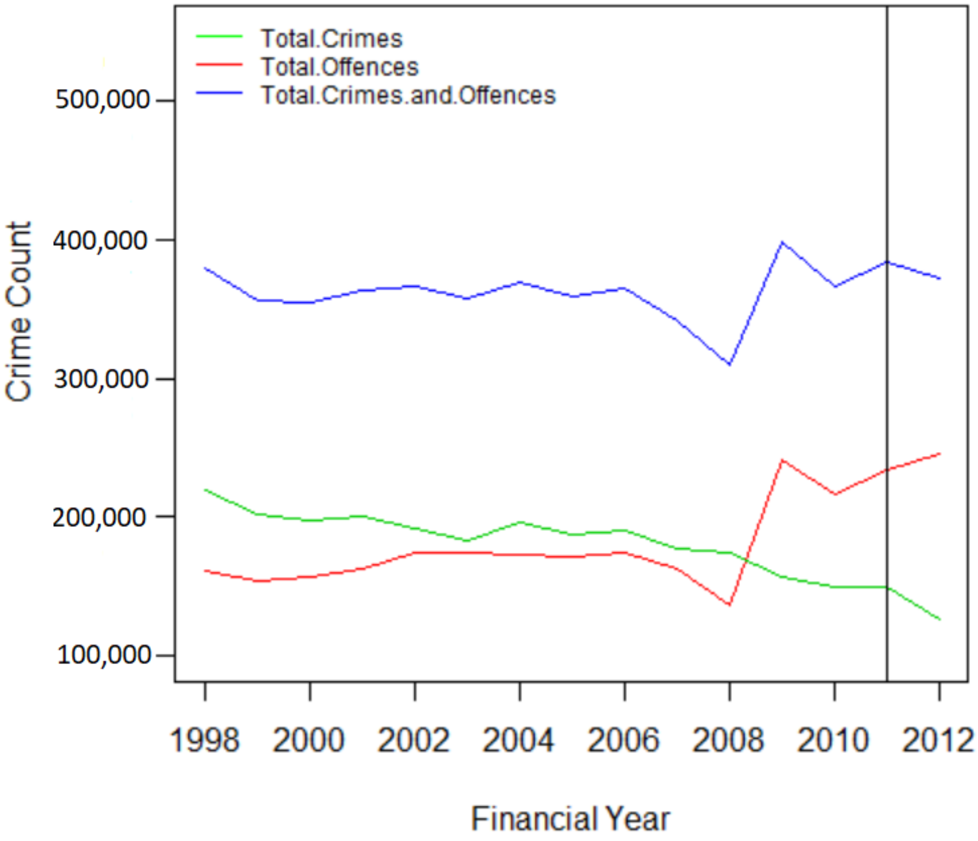


Figure 3.2: Recorded crimes in Scotland from 1998 to 2012

Study Area: Strathclyde

As I will be focussing only on the region of Strathclyde within Scotland, this section will give some background on the Strathclyde area and in particular Glasgow City as the most populated area within Strathclyde.

Prior to April 2013, there were eight regional police forces operating in Scotland. These were Central, Dumfries and Galloway, Fife, Grampian, Lothian and Borders, Northern, Strathclyde and Tayside (The Scottish Government, 2009a). Of particular interest for this thesis is the legacy Strathclyde Force which had the largest number of staff and share of the population and was only smaller in area size to Northern constabulary. The legacy Strathclyde Police Force region included Argyll and Bute, East Ayrshire, East Dunbartonshire, East Renfrewshire, Glasgow City, Inverclyde, North Ayrshire, North Lanarkshire, Renfrewshire, South Ayrshire, South Lanarkshire, and West Dunbartonshire.

The population in Glasgow City in 2011 was 593,245 (out of 5,295,403 across Scotland) so over 10% of the population of Scotland lived within Glasgow City (National Records Scotland, 2013a). Indeed, the Strathclyde region had 2,249,393 of Scotland's population within it (National Records Scotland, 2013a) so almost half the population live within the Strathclyde Police Force jurisdiction.

Within Glasgow, there were 409,801 people of working age and of this 19% were employment deprived as at 2011 (Scottish Index of Multiple Deprivation, 2015). This compares to only 13% (at 2011) of working age people being employment deprived across Scotland as a whole (Scottish Index of Multiple Deprivation, 2015) showing that Glasgow has a higher level of employment deprivation than Scotland on average. The percentage of the population aged 16 to 74 who were unemployed at the 2011 Census was 6.5% in Glasgow City which was slightly higher than the Scottish figure of 4.8% (The Scottish Government, 2018). The percentage of households who were not deprived in any way was 33.5% in Glasgow City compared to the Scottish figure of 40.1% (The Scottish Government, 2018) which seems to be considerably higher. The percentage of households who were deprived in 3 dimensions was 10.9% in Glasgow City and 6.4% across the whole of Scotland (The Scottish Government, 2018). This suggests that a larger proportion of the population live in deprived households in Glasgow City than across the whole of Scotland.

At a housing level, the number of homes which were socially rented (from council or other socially rented) was 36.7% at 2011 within Glasgow City compared to only 24.3% within Scotland as a whole (National Records Scotland, 2013b). While only 45.5% of homes were owned (owned outright, owned with a mortgage/loan, or shared ownership (part-owned/part-rented)) in Glasgow City compared to 61.9% across Scotland (National Records Scotland, 2013b).

Study Area Crime Data

The data which is used in this thesis were provided by Strathclyde Police (now Police Scotland) and covers all recorded crime which occurred within the area patrolled by the legacy Strathclyde Police Force in 2011. The dataset is all crimes and offences recorded by Strathclyde police within the 2011/12 financial year i.e. from April 2011 to March 2012 (referred to as the year 2011 for this thesis). The year 2011 data had already been provided by Dr Ellie Bates which had the 2001 data zone level attached and this only required to be updated to the 2011 data zone and 2011 output area aggregations which will be described in Chapter 4. The census of 2011 meant that population levels could be easily identified to calculate the crime rates as this was a census year.

Chapter 4 – Methodology

In this section I will discuss the following methods: k-means, finite mixture models, Local Moran's I and Getis-Ord Gi*. As has been previously discussed, if the number of areas is too large, then hotspots cannot be easily identified visually, and statistical methods are required. A cluster which has a high mean crime rate/count would be defined as a crime hotspot. Using cluster analysis will enable me to find groups of observations that are very similar to the other observations within the group but that are very dissimilar to observations contained within other groups.

In order to carry out different types of crime hotspot analysis, it is important to distinguish between counts and rates of crime because some methods need continuous data. k-means will therefore use rates while all other methods will use counts. Counts and rates are defined as follows for this thesis:

- counts of crime are the number of crimes/offences which occur within a certain area.
- rates of crime are the expected number of crimes per 100 population which are calculated by taking the counts of crime divided by the population number in a certain area multiplied by 100.

There can sometimes be issues when comparing counts and rates, however, in this instance output areas have approximately 250-375 people and data zones have approximately 500-1000 people in them. The way that these areal units are constructed is to ensure that they are fairly homogenous in population sizes. Therefore, it was not considered to be of concern when comparing k-means using rates and the other methods using counts. However, another consideration is that while in actual size, the populations may be considered fairly homogeneous, the actual make-up of these areas can be very mixed across race, age, economic backgrounds suggesting that taking these factors into account would be interesting further work.

Common Notation used in this chapter

x_i = crimes rates (or counts for k-means)
i, j index = areal units i.e. data zones / output areas
k = number of clusters
n = number of observations

w = spatial weight matrix
w_{ij} = entry associated with i and j-th entry in w
= 1 if areas i and j share a border
= 0 if areas i and j do not share a border

$\bar{x}^{(l)}$ = mean of areas in cluster l, also known as the centroid of l -th cluster

θ_m = cluster specific parameter vector for density function for cluster m in finite mixture models

π_m = prior probability of component m in finite mixture models

I_i = Local Moran's I for an area i in finite mixture models
z-statistic = $z_i = x_i - \bar{x}$

P_{il} = Posterior class probability for area i belonging to component/cluster l

H_m = set of area indices belonging to the m-th cluster

4.1 Cluster Methods (without spatial contiguity constraints)

4.1.1 k-means

One of the most popular partitional cluster methods is k-means clustering (MacQueen, 1967). Partitional clustering involves dividing the observations within the data into subsets which are non-overlapping i.e. each observation belongs to only one subset. k-means uses algorithms to minimise the within-cluster variance of the points and relative to this the between-cluster variance of the points should be maximised.

Definition

The equation for the within cluster sum of squares which is to be minimised is,

$$wcss = \sum_{l=1}^{k} \sum_{i \in H_l} (x_i - \bar{x}^{(l)})^2 \tag{4.1}$$

where,
x_i - value of x at area i
$\bar{x}^{(l)}$ - mean value of areas in cluster l (cluster centroid)
k - number of clusters
H_l - set of area indices belonging to the l th cluster

Ideally for k-means, we would look at all possible allocations to k clusters of the data. Due to this k-means is referred to as an NP-Hard problem because it is very difficult to enumerate all possible cluster allocations as the number of calculations required to do this becomes immense. This can be seen from Table 4.1 which shows the number of partitions required for a set number of observations, n, and number of clusters k.

Table 4.1: Number of partitions for selected observations and clusters (Everitt & Hothorn, 2011)

Number of Observations (n)	Number of Clusters (k)	Approximate number of possible partitions
15	3	$2x10^{6}$
20	4	$4x10^{10}$
25	8	$9x10^{18}$
100	5	10^{68}

In light of Table 4.1, it would seem that since the data with Strathclyde aggregated to n_1=2,963 data zones or n_2=19,886 output areas, there would be an extremely large number of calculations to be made. The actual number of calculations would be far greater than the 10^{68} from this table. Also, considering that the number of groups/clusters would possibly be greater than five, there is a need for an approximate algorithm to optimise the function. In order to apply k-means to the data, there are a number of algorithms which can be used which would not involve calculating all possible partitions.

Lloyd's Algorithm

The simplest version of k-means algorithm is Lloyd's algorithm (Lloyd, 1957).

Lloyd's algorithm is as follows for a given number of k clusters:

1) Initially select k random points/areas as cluster centroids
2) Assign each of the points/areas to the cluster corresponding to the closest centroid in terms of distance
3) Calculate the new centroid of each cluster based on the new assignments
4) Repeat 2. and 3. until clusters remain constant and the centroids do not change.

The main advantage of Lloyd's algorithm is that it can be used on large data sets.

Hartigan-Wong Algorithm

The algorithm used as default in R is the Hartigan-Wong algorithm (Hartigan & Wong, 1979). Hartigan-Wong method has a fast initial convergence. The main difference between Hartigan-Wong and Lloyd's algorithms are that Hartigan-Wong updates the cluster centroids once each point is moved to a new cluster, one at a time, while Lloyd's only updates the centroids once all the points have been grouped/re-grouped.

Hartigan-Wong algorithm is as follows for a given number of k clusters:

1) Initially select k random points/areas as cluster centroids
2) Assign each of the points/areas to the cluster corresponding to the closest centroid in terms of distance
3) Calculate the centroid of each cluster
4) For each point in turn, calculate the following

    a. given the current cluster centroids assign the point to the cluster with the closet centroid
    b. if point is moved to a different centroid, re-calculate the centroid for both the cluster it has left and the cluster it has joined.

5) Repeat 4) until any further changes would make the within cluster sum of squares larger and the between cluster sum of squares smaller.

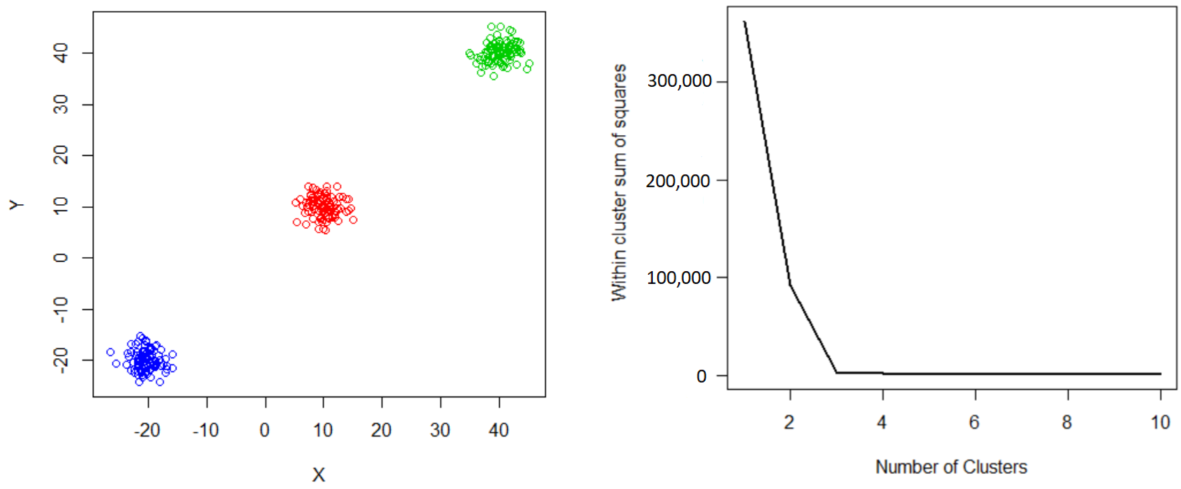
Selecting numbers of clusters, k

Before analysis is carried out, the correct number of clusters, k, and which cluster each observation belongs to is unknown. As data may have different optimal k values, k cannot be arbitrarily set for all data and a way of selecting k for each specific data is needed.

A selection option suggested by Jain (2010, p654) as "k-means is run independently for different values of k and the partition that appears the most meaningful to the domain is selected". This suggests running analysis and selecting the k that provides the most meaningful results. Another selection process that could be useful when looking at crime hotspots could be to fix k a priori to be meaningful for this analysis. This could be that k=3 which would lead to 3 clusters potentially being identified, Low, Medium and High crime clusters. Equally for k=5 this could lead there to be Low, Low-Medium, Medium, Medium-High, High crime clusters. However, this type of analysis suggests that for higher k values it would be hard to assign meaning to each of these cluster groupings.

Another way of identifying k would be to use an elbow plot to find a likely optimal value of k. An elbow plot is a line plot of the total within cluster sum of squares for all clusters as you increase the number of clusters. As the number of clusters rises, the within cluster sum of squares will always decrease, however, the aim is to see where increasing the number of clusters does not lead to a substantial decrease in the within cluster sum of squares i.e. where the bend (elbow) in the plot appears. Figure 4.1a shows data split into 3 distinct groups, the Blue group are centred at (-20,-20), the Green group are centred at (0,0) and the Red group are centred at (20,20). Figure 4.1b) then shows the elbow plot for this data. This shows the within cluster sum of squares for the data when split into 1 cluster up to 5 clusters. As can be seen from the plot, there appears to be an "elbow" or bend at both k=2 and k=3. However, after k=3, the difference in the within cluster sum of squares appears to be small as there is little further drop in the within group sum of squares. From this, it would appear that k=3 is the best way to partition this data and this agrees with the visual representation of the data.

For the analysis in this thesis, k was chosen by running the k-means algorithm several times for different k until increasing the number of k did not provide any substantial reduction in within groups variance. The elbow plot was used in visualising where this optimal k should lie.



(a) Three distinct clusters

(b) Elbow plot for three distinct clusters example

Figure 4.1: Three distinct clusters and the elbow plot

Strengths and Weaknesses

A strength of k-means, and one of the reasons that it is used so frequently, is that for large continuous data, it is computationally faster than many other clustering methods. It is simple and easy to use which means that it is a trusted method and can be applied to most data. This leads k-means to be an obvious choice when looking at different types of clustering methods for hotspot analysis as it is so frequently used (Andresen, Curman, & Linning, 2017; Jain, 2010; Kanungo et al., 2002; Morissette & Chartier, 2013; Zhang & Fang, 2013).

A potential issue with k-means is that it assumes that the underlying groups within the data will be spherical in shape. This can lead to issues when the data's true groups are non-spherical. However, many other cluster methods also make assumptions about the group shape and these limitations are yet to be fully overcome. Therefore, for this thesis, the group shape is assumed to be roughly spherical, when using k-means for hotspot analysis.

Another issue that exists with k-means is that each time k-means is run, it can potentially find different clusters as it will start with different initial conditions. In order to ensure that it is not local minima that are being identified and clusters formed based on this, it is beneficial to use multiple random starts and then select the solution with the best within cluster sum of squares.

Application to Crime Data

The software which is used for k-means is the "kmeans" function which is part of the basic package in R (R Core Team, 2016). k-means will be carried out on the crime rates as these are continuous and will use a number of random starts.

4.1.2 Finite Mixture Models

Definition

An increasingly popular alternative to k-means is finite mixture modelling (McLachlan & Peel, 2000). This differs from k-means clustering as finite mixture modelling is a model-based approach as opposed to an algorithmic approach. Finite mixture models have existed for many years but are used more frequently now due to the development of computing power which can process the computations involved quickly and efficiently. Finite mixture models assume that the data follows a weighted summation of probability densities and that each component density corresponds to a different cluster, with the aim being to estimate the parameters of the underlying probability density.

Maximum likelihood estimation cannot be directly used to estimate the model parameters in finite mixtures. We introduce missing data which are defined to be the (unknown) cluster memberships of each observation. The model parameters are then estimated by maximising the complete likelihood function and this can be done using the E-M Algorithm (Dempster, Laird, & Rubin, 1977). The likelihood function for a set of data (in this case crime counts) $\boldsymbol{x} = (x_1, ..., x_n)$ with k components (which are defined as the clusters) is,

$$L(\theta_1, ..., \theta_k, \pi_1, ..., \pi_k \mid \boldsymbol{x}) = \prod_{i=1}^{n} \sum_{m=1}^{k} \pi_m f_m(x_i \mid \theta_m) \qquad (4.2)$$

where,

x = data (in this case the crime counts)

$f_m()$ = the density of the m-th component (cluster)

θ_m = component specific parameter vector for the density function f

π_m = prior probability of an observation (area) belonging to the m-th component (cluster)

$0 < \pi_m \leq 1$ for m=1,…,k and $\sum_{m=1}^{k} \pi_m = 1$

Consider the vector of cluster memberships (missing data) for the i-th observation is defined to be $\boldsymbol{z}_i = (z_{i1}, ..., z_{ik})$ where z_{im} = 1 if x_i is from cluster m, or 0 otherwise. The complete likelihood of the parameters for x_i given z_i is $\prod_{m=1}^{k} \left[\pi_m f_m(x_i \mid \theta_m)\right]^{z_{im}}$ and specifically for a Poisson model $f_m(x_i \mid \lambda_m) = \frac{\lambda_m^{x_i}}{x_i!} \exp\{-\lambda_m\}$.

The complete data are then $\boldsymbol{y}_i = (x_i, \boldsymbol{z}_i)$. Each $\boldsymbol{z}_i$ is independently and identically distributed from a multinomial distribution with probabilities $(\pi_1, ..., \pi_k)$. We then have the complete-data likelihood as,

$$l(\theta_1, ..., \theta_k, \pi_1, ..., \pi_k, \boldsymbol{z}_1, ..., \boldsymbol{z}_n \mid x_1, .., x_n) = \sum_{i=1}^{n} \sum_{m=1}^{k} z_{im} \log(\pi_m f_m(x_i \mid \theta_m)) \quad (4.3)$$

$$l(\lambda_1, ..., \lambda_n, \pi_1, ..., \pi_n, \boldsymbol{z}_1, ..., \boldsymbol{z}_n \mid x_1, .., x_n) = \sum_{i=1}^{n} \sum_{m=1}^{k} z_{im} \log(\pi_m f_m(x_i \mid \lambda_1, ..., \lambda_m)) \quad (4.4)$$

$$l(\lambda_1,...,\lambda_k,\pi_1,...,\pi_k,\mathbf{z}_1,...,\mathbf{z}_n \mid x_1,..,x_n) = \sum_{i=1}^{n}\sum_{m=1}^{k} z_{im}\log\left(\pi_m\left(\frac{\lambda_m^{x_i}}{x_i!}\exp\{-\lambda_m\}\right)\right) \quad (4.5)$$

Estimation: E-M Algorithm

The E-M algorithm computes the maximum likelihood estimates of the model parameters for the data in the presence of missing data (Dempster et al., 1977). There are two steps to this iterative process, the (expectation) E-step and the (maximisation) M-step.

The E-step estimates the expectation of the missing data z_i which in the mixture model setting are the cluster memberships given the observations and the current parameter estimates. In the M-step the likelihood function is maximised given the missing data estimate from the previous E-step.

The E-step for iteration t is

$$\hat{z}_{im}^{(t)} \leftarrow \frac{\hat{\pi}_m^{(t-1)} f_m(x_i \mid \hat{\lambda}_m^{(t-1)})}{\sum_{j=1}^{k} \hat{\pi}_j^{(t-1)} f_j(x_i \mid \hat{\lambda}_j^{(t-1)})} \qquad \text{for} \quad \begin{matrix} i=1,...,n \\ j=1,...,k \end{matrix} \qquad (4.6)$$

where,

$\hat{z}_{im}^{(t)}$ = the t-th iteration estimate of the missing values (cluster group membership)

x_i = i-th data point (in this case the crime counts)

$f_m()$ = the distribution of the m-th component (cluster)

$\hat{\pi}_m^{(t-1)}$ = (t-1)th estimation of the prior probability of an observation (area) belonging to the m-th component (cluster)

$\hat{\lambda}_m^{(t-1)}$ = (t-1)th estimation of the mean and variance of the m-th cluster

The M-step for iteration t is

$$\hat{n}_m^{(t)} = \sum_{i=1}^{n} \hat{z}_{im}^{(t)} \qquad (4.7)$$

$$\hat{\pi}_m^{(t)} = \frac{\hat{n}_m^{(t)}}{n} \qquad (4.8)$$

$$\hat{\lambda}_m^{(t)} \leftarrow \frac{\sum_{i=1}^{n} \hat{z}_{im}^{(t)} x_i}{n_m^{(t)}} \qquad (4.9)$$

where n is the number of observations.

These steps are repeated until a pre-determined threshold is met or the maximum number of iterations is reached. The pre-determined threshold is decided by the researcher prior to the research commencing e.g. when the difference between the successive likelihoods becomes less than 0.05. At each iteration the likelihood function is guaranteed to increase leading to a guaranteed convergence for simple convex likelihood surfaces (which is unfortunately not the case for mixture models).

Selecting numbers of clusters, k

We would like to choose the model which has the largest likelihood value but it will continue to increase as the number of clusters increases. There needs to be a penalty included to discourage overfitting and we can do this using the Akaike Information Criteria or Bayesian Information Criteria. Both will penalise the likelihood differently and either will allow us to select the "best" number of clusters, k .

The Akaike Information Criterion (AIC) (Akaike, 1973) is a measure of relative quality of statistical models, in this case, the different numbers of clusters, for given data. It does not say that the model is a good fit only that it is a better fit relative to other models (i.e. there is the possibility that none of the models fit the data but the one selected will be the closest to fitting it). The "best" model relative to the other models fitted is the one in which the AIC value is lowest.

The AIC equation is,

$$AIC = 2q - 2\ln(L) \qquad (4.10)$$

where,

q is the number of estimated independent parameters in the model,

L is the maximised value of the likelihood function of the model M i.e. $L = p(x \mid \theta, M)$

The Bayesian Information Criterion (BIC) (Schwarz, 1978) is closely related to the AIC and the difference is BIC uses the following formulation,

$$BIC = q\ln(n) - 2\ln(L) \ , \qquad (4.11)$$

where

L is the maximised value of the likelihood function of the model M i.e. $L = p(x \mid \theta, M)$

θ are the parameter values that maximise the likelihood function
x are the observed data
n is the number of observations
q is the number of estimated independent parameters in the model.

Each mixture model with a different number of components (k) comprises a different model. The information criterion can be used to compare the models and select one with a corresponding selection of k.

Strengths and Weaknesses

As discussed in this section one of the key strengths of finite mixture models is the ability to adapt the model to fit the type of variable, which could be Poisson for count data or Normal for continuous data.

Finite mixture models make an assumption about the cluster distribution. If the cluster distribution and the true group distribution are not the same, then the model fitted can be incorrect which could lead to misleading results e.g. if the data came from a single t distribution but the mixture model fitted to the data was a mixture of normal distributions, a result would still be produced which would likely be incorrect, probably with k>1.

It is difficult to estimate the model if we don't have a large enough sample size. This will not be an issue with the crime data being used as the number of observations are 2,936 and 19,986 respectively. Another issue with using the E-M algorithm is that parameter estimation is slow. This is why k-means is more frequently used as it is faster. However, k-means will automatically assume that the groups are spherical while using the finite mixture models allows different models and shapes to be taken into account.

Application to Crime Data

The software which is used for the finite mixture modelling is a finite mixture package in R called flexmix (Gruen & Leisch, 2007, 2008; Leisch, 2004). For the crime data, the assumed distribution is a mixture of Poisson distributions i.e. each cluster comes from a different Poisson distribution (with a different mean). flexmix allocates the observations to different clusters using the E-M algorithm until the likelihood for a fixed number of clusters is greatest and changing the groupings does not increase the likelihood by a substantial amount (or until the maximum number of iterations has been reached).

4.2 Cluster Methods (with enforced spatial contiguity constraints)

In order to look at cluster methods which are spatially constrained, Local Indicators of Spatial Association (LISA) methods are used. Spatially contiguous means areas are clustered with other areas only if they share a border or are within a pre-specified distance from each other. Spatially non-contiguous means the areas that are clustered together do not need to be neighbours and there are no constraints placed on the areas being near each other. Two LISA methods are Local Moran's I and Getis Ord Gi* which can be calculated using ArcGIS/QGIS software. This study will utilise the 'Cluster and Outlier Analysis' package for the Anselin Local Moran's I method and the 'Hot Spot Analysis' package for the Getis-Ord Gi* method. To carry out LISA methods, the data was added to shapefiles for Strathclude at both output area and data zone level.

In order to use these methods, it is important for a spatial weights matrix w to be defined. The spatial weights matrix can be defined differently depending on how we wish to classify neighbours. It can be defined by distance or borders. A possible weights matrix can use either the Euclidean distance or Manhattan distance measures to identify areas as neighbours which are within a defined distance d from observation/area i. Another possible spatial weights matrix is to define it based on areas which have a direct border with each other i.e. if i and j share a border then they have a weight of 1 and 0 if they do not share a border. For the purposes of this thesis, the weight matrix will be defined using borders.

Both LISA methods used in this thesis (Local Moran's I and Getis Ord Gi*) have a number of strengths which are that they can highlight hotspots as they identify local minimums and 'coldspots' (Bates, 2014). Also, a level of significance is assigned to each area which enables them to be identified as statistical 'hotspots'. The neighbourhood matrix, w, is adjustable when using these methods which enables observations to be counted in different ways e.g. neighbourhoods can be defined in different ways.

Another strength of the LISA methods are that they can be readily visualised. The use of GIS software to carry out this analysis means that maps can be easily produced (Davis, 2012). Using maps can be an excellent reference to use as it gives an easily understood representation of complex statistical methods meaning greater visibility for where problem areas lie.

4.2.1 Anselin Local Moran's I (Cluster and Outlier Analysis)

The aim of the Local Moran's I statistic (Anselin, 1995) is to identify local spatial patterns in areal data. It was developed from the Global Moran's I which is a measure of the overall pattern of values of the variables being studied to see if the values change smoothly (positively autocorrelated) or not (negatively autocorrelated). Local Moran's I is a measure of how correlated a local observation is with its neighbours.

Definition

The use of Local Moran's I allows spatial patterns to be identified by looking at how the observations group in relation to the weights given to each observation in the data. We will calculate the relationships between areas neighbouring a single locale rather than the overall relationships. Global Moran's I, Local Moran's I and the associated z-statistic (for Local Moran's I) are given by the following equations (Anselin, 1995),

$$\text{Global Moran's I} = \frac{n}{S_o} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4.12}$$

where,

$$S_o = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$$

I_i = Local Moran's I

$$I_i = \frac{x_i - \bar{x}}{s_i^2} \sum_{j \neq i,\, j=1}^{n} w_{ij}(x_j - \bar{x}) \tag{4.13}$$

where,

w_{ij} = spatial weight matrix (with $w_{ii} = 0$)

$$s_i^2 = \frac{\sum_{j \neq i,\, j=1}^{n} w_{ij}\left(x_j - \bar{x}\right)^2}{n-1} \quad .$$

The z-statistic is,

$$z_{Ii} = \frac{I_i - \mathrm{E}[I_i]}{\sqrt{Var[I_i]}} \tag{4.14}$$

where $\mathrm{E}[I_i]$, the expected value, and $Var[I_i]$, the variance, are defined as,

$$\mathrm{E}[I_i] = -\frac{\sum_{j \neq i,\, j=1}^{n} w_{ij}}{n-1} \;, \qquad\qquad Var[I_i] = \mathrm{E}[I_i^2] - \mathrm{E}[I_i]^2$$

$$\mathrm{E}[I_i^2] = \frac{(n-b)\sum_{j \neq i,\, j=1}^{n} w_{ij}^2}{n-1} - \frac{(2b-n)\sum_{k=1, k \neq i}^{n} \sum_{h=1, h \neq i}^{n} w_{ik} w_{ih}}{(n-1)(n-2)} \;, \quad b = \frac{\sum_{j=1, i \neq j}^{n} (x_i - \bar{x})^4}{\left(\sum_{i=1, i \neq j}^{n} (x_i - \bar{x})^2\right)^2} \,.$$

Interpretation

The values for Local Moran's I can be either positive or negative and enables hot spots to be identified. Local Moran's I is positive when neighbouring areas have similar high or low values (forms a cluster) and is negative when the neighbouring areas have dissimilar values (area is an outlier). The idea of the cluster suggests that this set of areas is either a hotspot or a coldspot (depending on whether it is high or low).

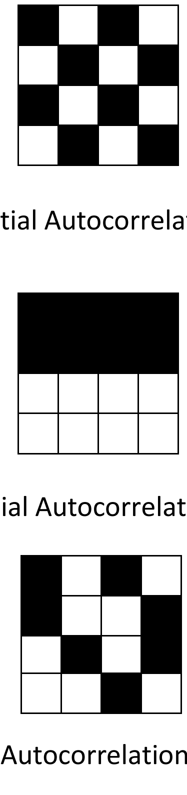
The output is in the form of z-scores (z-statistic) and COType (Cluster/Outlier Type) values which can show where the hotspots and coldspots are. Areas with a positive z-score have a similar value to the surrounding areas - either all high or all low values. The COType value associated will identify whether the areas are a hotspot or a coldspot, be 'HH' (High-High) means that it is a cluster of high values (in this case high crime counts) or 'LL' (Low-Low) would mean that it is a cluster of low values (low crime counts).

Areas with a negative z-score have different high or low values than the surrounding areas. The associated COType 'HL' (High-Low) means that the area has a high value while the surrounding areas have low values and a COType 'LH' (Low-High) would mean that it is a low value area surrounded by areas which have high values.

Figures 4.2(a) to (c) show high crime (white) and low crime areas (black). In Figure 4.2(a), there is no sign of positive autocorrelation as each high (white) crime spot is surrounded by a low (black) crime spot. These would be an example of H-L and L-H outputs. This shows negative spatial autocorrelation with a Global Moran's I of -1. Each of these areas would have negative Local Moran's I scores as each area of high crime is surrounded by a mix of areas with high and low crime counts.

In Figure 4.2(b), there is clear positive spatial autocorrelation, this shows that the top half is all low crime areas while the lower half is all high crime areas. This would have a Global Moran's I of +1 and would be an example of H-H and L-L outputs. Each of these areas would have positive Local Moran's I scores as each area of high crime is surrounded by other areas with high crimes counts and similarly for low crime count areas. The areas which lie horizontally in the middle (the border areas) would have mixed Local Moran's I scores as they are surrounded by both high crime areas and low crime areas.

In Figure 4.2(c), this shows areas of H-H, L-L, L-H and H-L mixed and there is no consistent pattern of spatial autocorrelation. High crime areas are scattered beside low crime areas as well as being close to other high crime areas and vice versa. Therefore, this type of pattern would show less of a pattern of spatial autocorrelation at a global level. At a local level, there is a low-low cluster in the bottom left of the plot and in the centre and a high cluster in the top left. Both of these would have positive Local Moran's I values as they are surrounded by areas which are similar in value to them.



(a) Negative Global Spatial Autocorrelation: Moran's I close to -1

(b) Positive Global Spatial Autocorrelation: Moran's I close to +1

(c) Mixed Spatial Autocorrelation: -1< Moran's I < +1

Figure 4.2: Spatial autocorrelation examples for Moran's I

Strengths and Weaknesses

One of the strengths of the Local Moran's I method is that it can account for local variations. This makes it more intuitive to understand. Therefore, by looking at the local level it stops spatial autocorrelation being masked at a global level. There might be no autocorrelation at a global level but if smaller areas are looked at, there could be some local autocorrelation which is important to identify.

This strength also then leads to one of the main weaknesses of this technique as it is very complex to correctly interpret (M. A. Andresen, 2015). It involves extra training and expertise being developed which leads to its lack of popularity with analysts. Quite often this is the reason why k-means clustering is carried out as it is easier to understand and interpret.

Also, it is impacted by the multiple testing problem, which is that the possibility of a type 1 error occurring automatically increases as the number of areal units being measured increases. A type 1 error is caused by wrongly concluding the clusters found are not a result of a random process and this can vary depending on the level of significance chosen. Bates (2014) discusses that a possible solution to this would be correcting the z-score using Bonferonni methods which would standardise the results and reduce the increasing type 1 error possibility. However, the Bonferonni methods when applied to previous examples were found to be too conservative and produced hotspot areas which were identified by practitioners as being too cautious, therefore Bonferonni use in crime hotpot analysis is limited (Bates, 2014).

Application to Crime Data

The Cluster and Outlier Analysis package enables hotspots and coldspots and spatial outliers to be detected based on the Local Moran's I statistic. For this thesis, Local Moran's I was run on the counts of crime using the default settings with a weight matrix based on a value of 1 if the areas border each other and 0 otherwise. The output from running the Local Moran's I Cluster Analysis in ArcGIS is a map highlighting the areas which are high-high, low-low, low-high and high-low clusters. This therefore, splits the data into different clusters and a map can be produced showing where the H-H, L-L, L-H and H-L areas lie. E.g. each H-H observation (area) can be a different cluster e.g. medium-high crime area, high crime area, and really high crime area.

4.2.2 Getis-Ord Gi* (Hot Spot Analysis)

The aim of the Getis-Ord Gi* statistic (Getis & Ord, 1992) is to compare the local averages to the global averages and checks if the locally identified hotspots are still relevant at the global level when looking at the full data as a whole. Getis Ord Gi* is known as hot spot analysis in ArcGIS.

Definition

The Getis Ord Gi* statistic is a local measure. It is a z-score and and looks at each area in the setting of its surrounding areas. Gi* calculates the sum of all the values for all areas, I, and its surrounding neighbours (j's) and calculates what proportion the local sum (i) is of the total sum for all areas. When the local sum is very different from the expected local sum, this results in statistically significant z-scores as the difference is too large to be due to random chance. In order for an area to be identified as a hotspot, the area must be a high value surrounded mostly by other high values.

The Gi* statistic is,

$$G_i{*} = \frac{\sum_{j=1}^{n} w_{ij} x_j - \bar{x} \sum_{j=1}^{n} w_{ij}}{S \sqrt{\left[\frac{\left(n \sum_{j=1}^{n} w_{ij}^2 - \left(\sum_{i=1}^{n} w_{ij} \right)^2 \right)}{n-1} \right]}} \tag{4.15}$$

where,

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \left(\bar{x}\right)^2}$$

w_{ij} = spatial weight matrix with the additional constraint that $w_{ii} = 1$ for all i

x_j = values (crime count) at location (area) j

The Null Hypothesis (H₀) associated with the Getis Ord Gi* statistic is,

H_0: there is no difference between the local mean level for an area and its neighbours and the population mean for this number of areas

H_A : there is a difference between the local mean level for an area and its neighbours and the population mean for this number of areas.

Figure 4.3(a) shows point i and its neighbours j's. Dark blue represents very high counts and blue represents high counts. Point i has a positive association with its neighbours j as the point i is dark blue and the surrounding areas (j's) are also a blue shade.

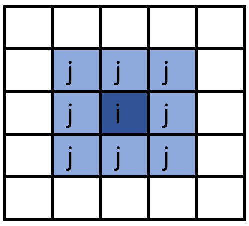


Figure 4.3(a) Getis Ord Gi* positive association (positive z-scores)

Figure 4.3(b) shows point i and its neighbours j's. Again, dark blue represents very high counts and in this case yellow represents low counts. Point i has a negative association with its neighbours j as the point i is dark blue and the surrounding areas (j's) are yellow suggesting they are lower crime areas.

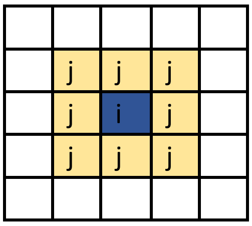


Figure 4.3(b) Getis Ord Gi* negative association (negative z-scores)

As the Gi* statistic is in the form of a z-score, the null and alternative hypotheses become,

$$H_0 : Gi^* \leq 0$$
$$H_A : Gi^* > 0$$

H_0 would refer to areas which are either non-significant or a coldspot and H_A would refer to areas which are a hotspot. By identifying if the local pattern of crime is similar or different to what is generally observed across the data as a whole, these coldspots and hotspots can be identified. Areas which have a high positive z-score and a low p-value suggest there is a cluster of high values with a higher z-score implying a more intense cluster. While areas which have a low negative z-score and low p-value suggest there is a cluster of low values with a lower z-score implying a more intense cluster. Thus for this thesis, to identify the hotspots, areas with a high positive z-score and low p-value would be selected.

Strengths and Weaknesses

The main weakness of the Gi* statistic is that it is heavily affected by outliers, as it becomes skewed if there are observations which do not lie clearly within a cluster and this can cause the clusters to be formed incorrectly. Gi* is also impacted by the multiple testing issue outlined in the Local Moran's I section.

As discussed in Local Moran's I, Gi* is also able to identify hotspots and coldspots. Also, similarly to Local Moran's I, output is in the form of maps which is easily understood and very visual.

Application to Crime Data

The Cluster and Outlier Analysis package enables hot and cold spots to be detected based on the Getis-Ord Gi*. This tool looks at the neighbouring features of the area to detect where clusters lie. This is not able to detect spatial outliers but if the identification of spatial outliers is important, the Local Moran's I statistical method can be used (see above). For this thesis, Getis-Ord Gi* was run on the counts of crime using the default settings with a weight matrix based on a value of 1 if the areas border each other and 0 otherwise. Getis-Ord Gi* will enable me to identify areas near each other which are similar and can be clustered together with areas which have a high crime count being considered hotspots.

4.3 Adjusted Rand Index

The Rand index or Rand measure (Rand, 1971) is a measure of similarity between data clusterings. It can compare clustering results from two different cluster methods on the same data. The adjusted Rand index (Hubert & Arabie, 1985) comes from the Rand index but assumes that the expected values if the clusterings came from two random clusterings would be 0.. The ARI for all the observations, if these are split into different clusters using different methods e.g. one partition comes from k-means and the other from Getis Ord Gi*, is number which shows how similar the two cluster groupings are. This will enable me to compare the four clustering methods at the output area level and the data zone level.

If the data X={x_1,x_2,…, x_n}, are split into two different partitions, E={e_1,e_2,…, e_S} and F={f_1,f_2,…, f_R} such that $\bigcup_{s=1}^{S} e_s = X = \bigcup_{r=1}^{R} f_r$ and $e_s \bigcap e_t = \phi = f_r \bigcap f_u$ where $1 \leq s \neq t \leq S$ and $1 \leq r \neq u \leq R$
then there are four inputs to the Rand Index. These involve counting the number of pairs of observations in the data which are:

i) a - the number of pairs of observations in the same cluster in partition E and the same cluster in partition F
ii) b - the number of pairs of observations in the same cluster in partition E and different clusters in partition F
iii) c - the number of pairs of observations in different clusters in partition E and the same cluster in partition F
iv) d - the number of pairs of observations in different clusters in partition E and different clusters in partition F

The Rand Index (RI) is then calculated as,

0 ≤ RI ≤ 1

$$RI = \frac{a+d}{a+b+c+d} \tag{4.16}$$

where $0 \leq RI \leq 1$

Then 'a' and 'd' can be thought of as being in agreement as the groupings match (either same 'a' or different 'd' groupings in both) while 'b' and 'c' are disagreements as the groupings don't match. If the two partitions matched exactly then RI would be equal to 1.

The contingency table counts all the instances of an observation occurring in each combination of clusters. An example for two clusterings, **E** and **F**, can be seen in Table 4.2.

Table 4.2: ARI Contingency Table for Clustering **E** and Clustering **F**

	F1	F2	F3	F4	Sums
E1	n_{11}	n_{12}	n_{13}	n_{14}	a_1
E2	n_{21}	n_{22}	n_{23}	n_{24}	a_2
E3	n_{31}	n_{32}	n_{33}	n_{34}	a_3
E4	n_{41}	n_{42}	n_{43}	n_{44}	a_4
	b_1	b_2	b_3	b_4	

where,

n_{sr} is the number of times an observation occurs in cluster s of E and cluster r of F

a_s is the s-th cluster row sums

b_r is the r-th cluster column sums.

If the two partitions i.e. clustering 1 and clustering 2, matched exactly then the ARI would be equal to 1.

The Adjusted Rand Index (ARI) is calculated as,

$$ARI = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] / \binom{n}{2}} \qquad (4.17)$$

where $ARI \leq 1$

Investigation of MAUP

The clustering methods will be applied at both the output area and data zone level to compare the results. I will look at MAUP by identifying the cluster methods which show similar cluster patterns between the data split at both output area and data zone levels. If there is a genuine MAUP issue, there will be a difference in the clustering output.

The cluster groupings for each method at each areal unit level will be taken. Maps are produced for each method at each areal unit levels using the ArcGIS software. The cluster groupings shown in the maps will be discussed and compared visually. If cluster is at the data zone level, for each area we know the output areas which lie below the data zone. We can then assign each output area to the same cluster that the parent data zone belongs to. This enables us to compare on the output area level, the clusterings which were done at the output area and data zone levels. The adjusted Rand index will be used to calculate a value comparing each cluster grouping with each other. First at the output area level, then at the data zone areal level, and then across the output area and data zone areal levels. The results of these analyses will be discussed in the next chapter.

Chapter 5 - Results

Software Used

Microsoft Access is used to aggregate the data to data zone and output areas from the original longitude and latitude co-ordinates.
R is the software used to provide the clustering for k-means (standard built-in cluster package) and finite mixture modelling (flexmix package).
ArcMAP is the software used to provide the clustering using Local Moran's I and Getis Ord Gi* Statistic (Spatial Statistics Toolbox -> Mapping Clusters Toolset).
Area: Strathclyde, Data from Police Scotland (formerly Strathclyde Police) and National Records Scotland

Data Used and Methods

The analysis for this thesis, was carried out on two different areal resolutions. These were 'All Crimes' in 2011 first split at (2011) data zone level and then split at the (2011) output area level. The output areas are nested within the data zones and these are all nested within the Strathclyde region which is the study area for this thesis. The clustering methods carried out on each areal data were k-means, finite mixture modelling, Local Moran's I, and Getis Ord Gi*. Adjusted Rand Index is calculated for the pairs of clusterings at output area level and the data zone level. For the clusters at data zone level, as we know the output areas which lie within each data zone, we can then assign each output area to the same cluster that the parent data zone belongs to. This enable us to compare on the output area level, the cluterings which were done at both the output area and data zone levels. The adjusted Rand index can be calculated for this comparison.

The term 'all crimes' is used in this thesis to refer to the total number of recorded crimes and offences variable in the data. The category 'crimes of a sexual nature' was excluded as due to the sensitive nature of these crimes, it was not permissible to get access to this data from Police Scotland. As such 'all crimes' relates to all crimes and offences excluding any crimes categorised by Police Scotland as 'crimes of a sexual nature'.

The data were available for the financial year 2011/12 with 2001 data zone aggregation thanks to Dr Ellie Bates who produced the 2011 data for her PhD thesis. This covered the time period from April 2011 to March 2012 and for the purposes of this thesis, will be referred to as the year 2011. The year 2011 is therefore the time period of analysis with the all crimes variable as the target. An advantage of analysing this year was the ability to get up to date data zone and output area splits (as there was a reclassification of output areas and data zones in 2011) and corresponding population levels for these areas from the 2011 census results.

The obtained data were aggregated at the 2001 data zone level and the individual crime points had to be aggregated to 2011 output areas and 2011 data zones. The maps for both 2011 output areas and data zones were obtained from the National Records Scotland website which produced the census statistics for this year. The map of the output areas for the whole of Scotland was created and the 2001 data zone map for Strathclyde was then added. The outlying output areas which did not lie underneath the Strathclyde data zone map were removed. This left the map of output areas of Strathclyde which meant the individual crime locations could then aggregated to the output area which they lay within. This provided data on all recorded crimes in Strathclyde aggregated at output areas. These steps were then repeated to aggregate to the 2011 data zone. In total, there were 19,886 output areas (Figure 5.1) in total for the Strathclyde region, compared to 2,963 data zones (Figure 5.2).

Table 5.1: Summary of areal unit counts within Strathclyde

Areal Unit	Count
Output areas	19,886
Data zones	2,963

In Figure 5.1, it can be seen that there are larger and smaller areas, the smaller areas appear to be in the centre of the map and correspond to the urban, city area, while the larger areas appear to be in the surrounding areas which are more rural areas. This is the same with Figure 5.2. The dark areas seen in the centre of both Figure 5.1 and Figure 5.2 show that there are a large number of small areal units at both the data zone and output area levels. This shows the Glasgow city centre area where there are a lot of data zones and even more output areas covering this small geographic space. When comparing Figure 5.1 and Figure 5.2, it can be seen that there are many more output areas than data zones and that usually several output areas (from Figure 5.1) could be combined to form one data zone (in Figure 5.3).

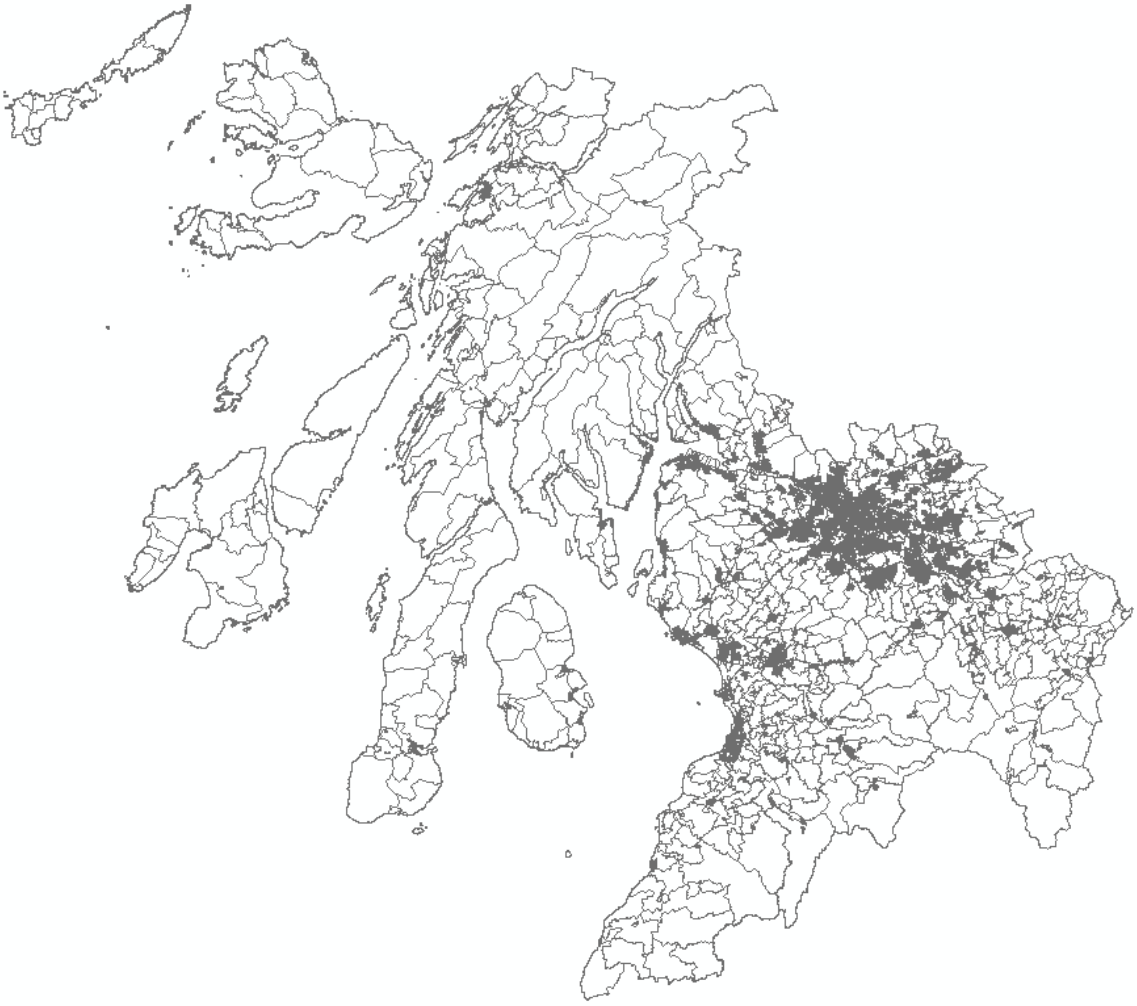


Figure 5.1: Strathclyde output areas for 2011

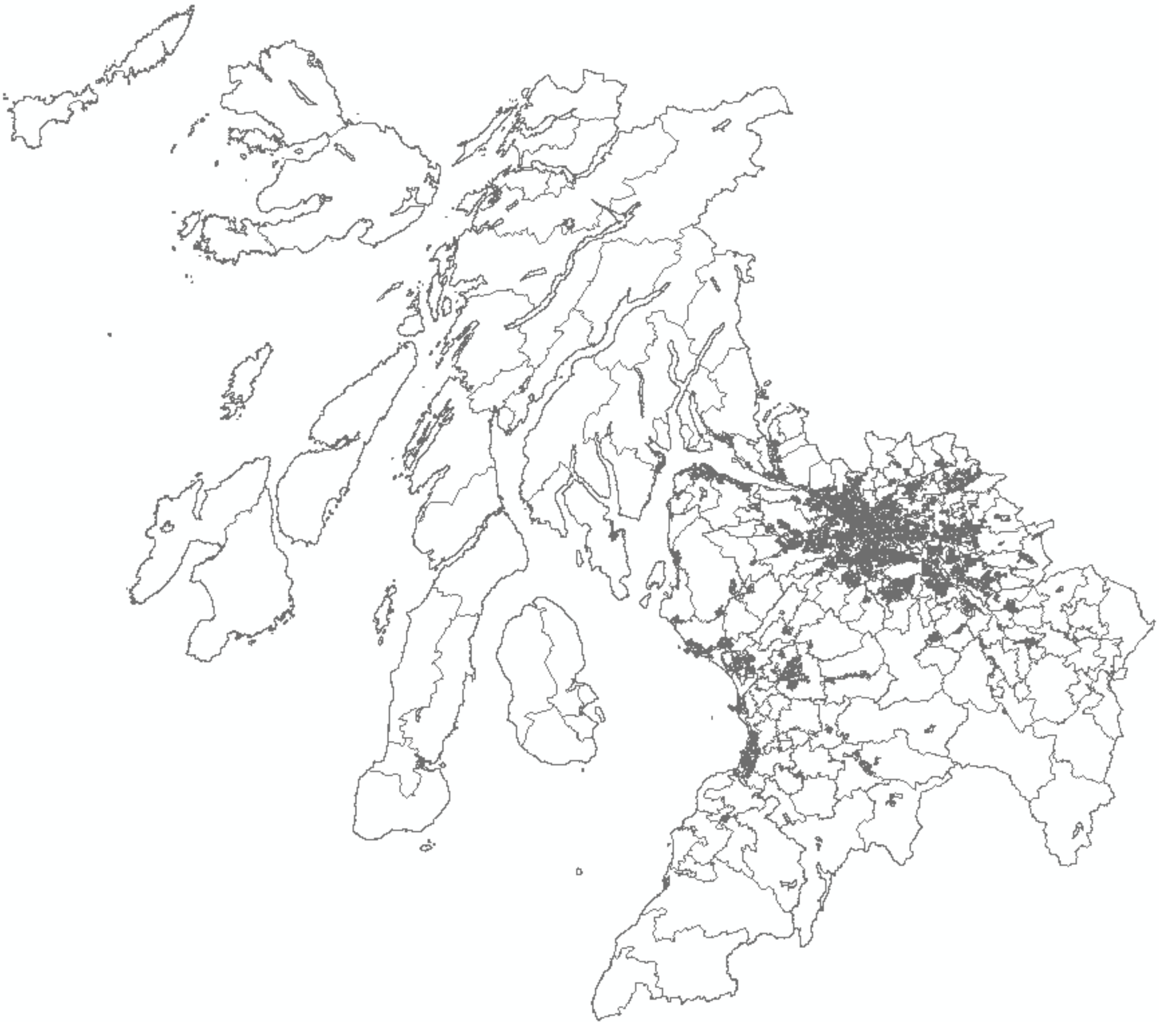


Figure 5.2: Strathclyde data zones for 2011

Figure 5.3 shows one of the data zones selected, S01010444 and the underlying output areas within it. The data zone boundary is shown by the light blue thicker line with the thinner black lines showing the output area boundaries The underlying output areas are S00112888, S00112893, S00116850, S00112912, S00112911, S00112892, S00112891, S00112890. This plot shows that this one data zone has eight output areas nested within it. Not all of the output areas lie directly within the data zones as there are sometimes slight overlaps with the next data zone but these are usually very slight differences so don't concern us too much. There are limitations with this as there could potentially be crimes that are recorded in the wrong data zone or output area or the population levels could be slightly incorrect if the output area does not lie directly within the data zone. However, due to the differences being very small as most output areas lie directly within the data zones, I do not believe this is a huge concern.

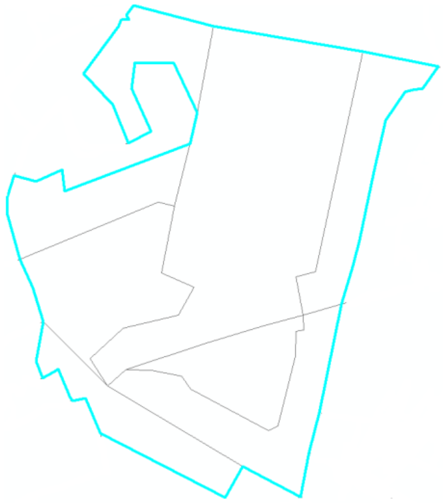
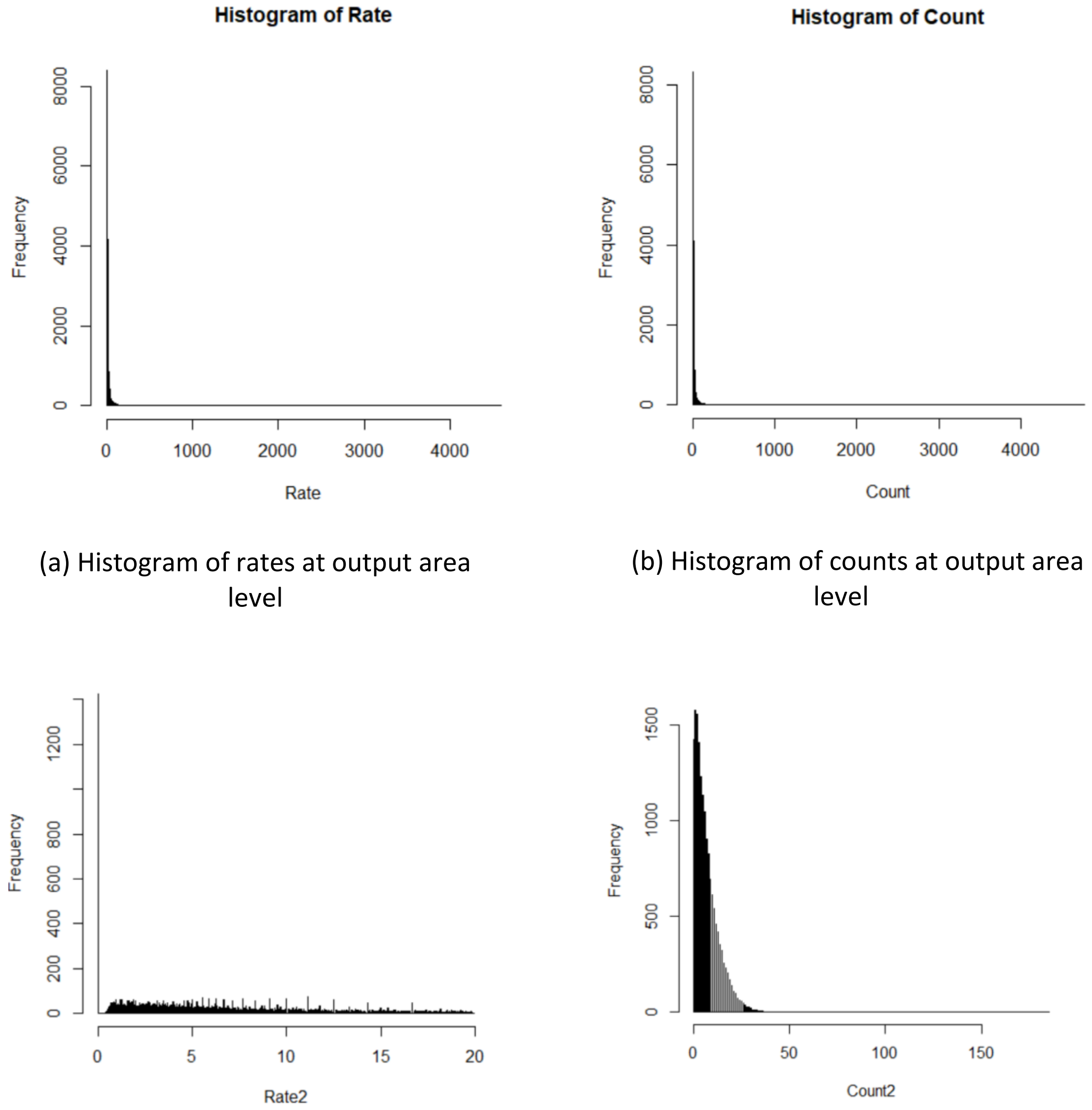


Figure 5.3: Data zone S01010444 with underlying output areas

Crime Counts vs Rates

The crime counts were taken to be the total crimes which occurred in each data zone or output area. In order to create the crime rates, the census output from Scottish Neighbourhood Statistics was used to get the population levels for each data zone and output area. The crime rates were total crimes in an area per 100 population. Usually rates per population are calculated per 10,000 or 100,000 but in this case per 100 population was chosen because there are only approximately 500 people per data zone so a rate of 1000 or more would not make sense. As output areas are even smaller areas then there are even lower population levels (approximately 100-200 people) in output areas. Therefore, the average populations were in the hundreds and thus per 100 seemed a reasonable selection. Figure 5.4 shows histograms of both rates (a) and counts (b) and then rates and counts with outliers removed in (c) and (d) for output area crimes in Strathclyde. These show that the distribution of the data are skewed to the right with a few outliers appearing to lie further to the right of the graph. This is further seen in Table 5.2 where it can be seen that there is a difference between the mean and median values as the mean is skewed by outliers.



(a) Histogram of rates at output area level

(b) Histogram of counts at output area level

(c) Histogram of rates at output area level removing outliers

(d) Histogram of counts at output area level removing outliers

Figure 5.4: Histograms at the output area level in Strathclyde for rates and counts of crime

Table 5.2 shows that the minimum value in an output area for both rates and counts is 0. The maximum values are 4,581 crimes per 100 people and 4,764 crimes respectively. The rate is very high as this output area. This suggests that there may be outliers in the data. This is further seen as the mean and median values are different. The median for the rates is 6.5 crimes per 100 people and the mean for rates is 18.94 crimes per 100 people. This suggests that there are a few outliers but not too many and most areas would appear to be low crime areas based on these statistics. The median for the counts is 7 crimes and the mean for counts is 19.31 crimes. This reinforces that there are a few outliers with most output areas being low crime.

Table 5.2: Summary statistics at the output area level

Variable	Output area rate	Output area count
Minimum	0	0
Median	6.50	7.00
Mean	18.94	19.31
Maximum	4,581	4,764

Table 5.3 shows the total number of crimes and offences which occurred in 2011 within the Strathclyde region (excluding crime of a sexual nature). It shows that the majority of crimes which occurred were actually officially classed as offences (234,396). For the purposes of this thesis, all crimes and offences will be referred to as crimes. In total there are 384,083 crimes which will be aggregated to output area and data zone levels which will be analysed in this thesis.

In total there were 12 output areas which had over 1000 instances of crimes occurring in them for 2011. The output area with the highest crime count had 4,764 crimes occur within it in 2011.

Table 5.3 Crime counts per crime category

Crime Category	Description	Count
Group 1	Non-sexual Crimes of Violence	5,173
Group 3	Crimes of Dishonesty	72,418
Group 4	Fire-raising and Vandalism	35,812
Group 5	Other Crimes	36,284
Group 6	Miscellaneous Offences	113,929
Group 7	Motor Vehicle Offences	120,467
	Total Crimes and Offences	384,083

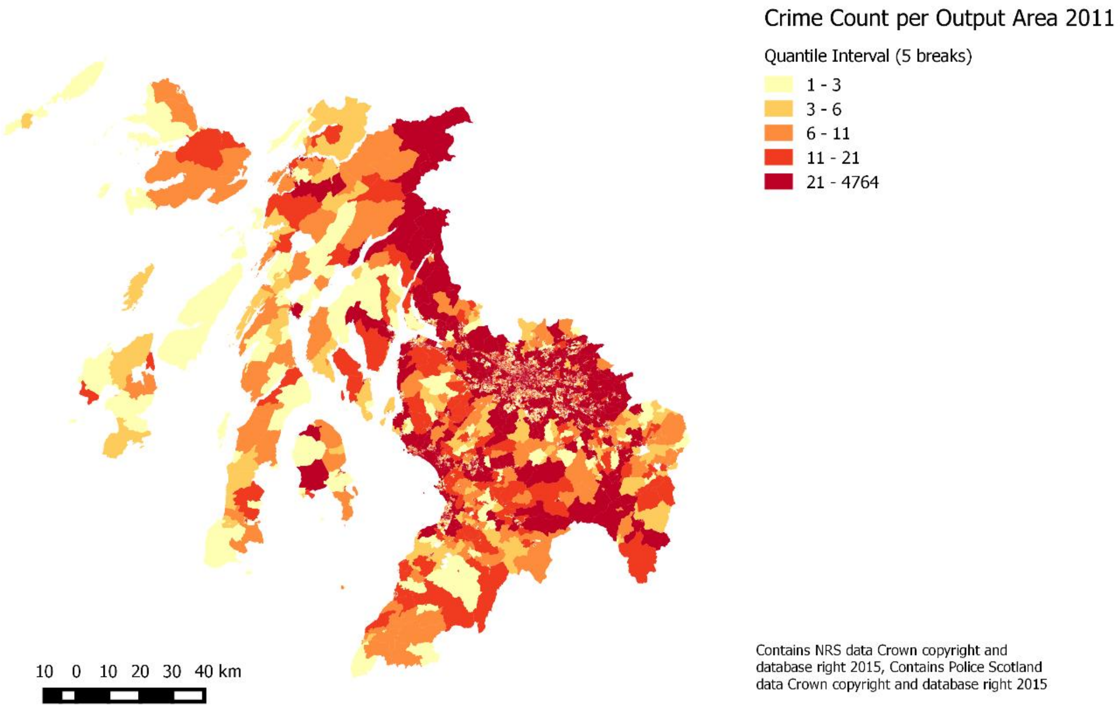
Choropleth Hotspot Maps

The police analysts can sometimes use Choropleth mapping which colours each areal unit in a map according to certain colour coding for quick analysis and then 'Cluster and Outlier Analysis' or 'Hotspot Analysis' toolsets for further investigation. Figures 5.5, 5.6 and 5.7 show the data at output area level, split into 5 categories based on different criteria. As can be seen, depending on the criteria chosen, the categories visualised varies greatly. The criteria chosen for the initial exploratory analysis were quantile intervals, equal intervals and Jenks intervals within the Choropleth mapping option in ArcGIS and these are described in Table 5.4.

Table 5.4 Summary of the Choropleth quantile, equal and Jenk's options

Name	Description
Quantile	Divides the attributes into categories with equal numbers of features i.e. equal number of output areas in each category
Equal	Equal sized sub-ranges of the attribute value i.e. equal range of crime counts in each category
Jenks	Natural breaks - splits data in to natural groups by minimising the average deviation from the mean, while maximising the deviation from the means of the other groups. i.e. put areas with similar crime counts together which makes groups similar to each other and dissimilar from other groups

Both the quantile and the equal interval methods (Figure 5.5 and 5.6) do not really tell us much about where the hotspots lie. They only show that the count of crimes in each output areas varies greatly, from 1 to 4,764. However, Jenk's method can be considered similar to the k-means method as it aims to reduce the difference in the values (crime counts) within the groups and maximise the differences in values between the groups. The mapped Jenk's intervals can be seen in Figure 5.7.


Crime Count per Output Area 2011
Quantile Interval (5 breaks)
1 - 3
3 - 6
6 - 11
11 - 21
21 - 4764
10 0 10 20 30 40 km

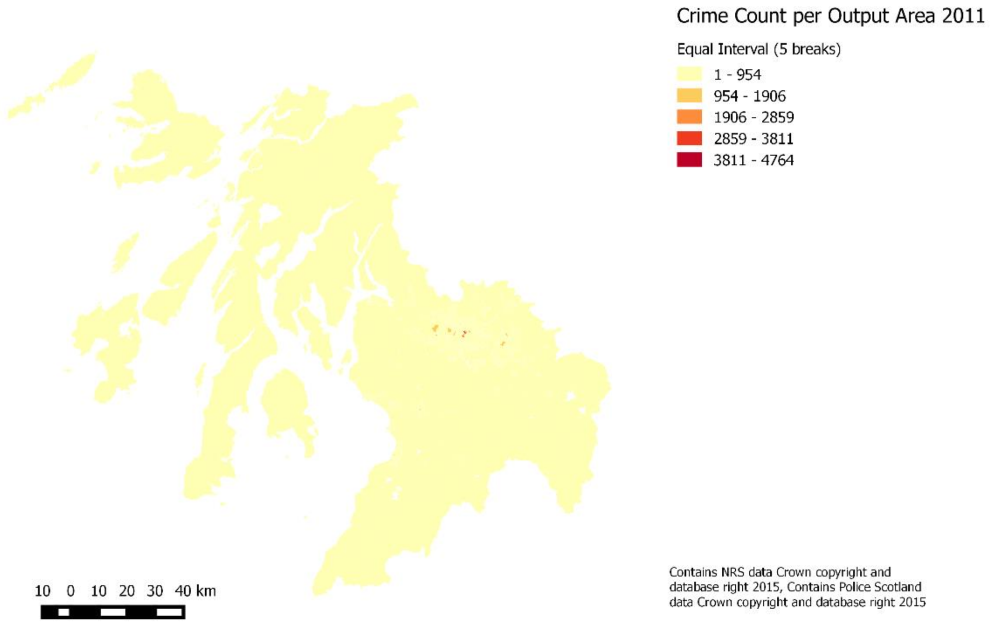
Contains NRS data Crown copyright and database right 2015, Contains Police Scotland data Crown copyright and database right 2015

Figure 5.5: Map of Strathclyde splitting by quantile intervals (output areas)

Using the quantile method, as seen in Figure 5.5 shows that there is a great deal of variation in the crime counts in the output areas within each category. There are a mix of high, medium and low crime areas all over the map which makes it hard to identify high and low crimes. The yellow or low crime category only has a range of 3 values, while the red or high crime category ranges from 21 counts per output area to 4,764 crime in an output area. The difference between 21 crimes and 4,764 is so great that it would not seem very intuitive to have areas with those crime counts in the same category. This suggests that a quantile map is not helpful in this situation as it is very difficult to interpret useful results from this.

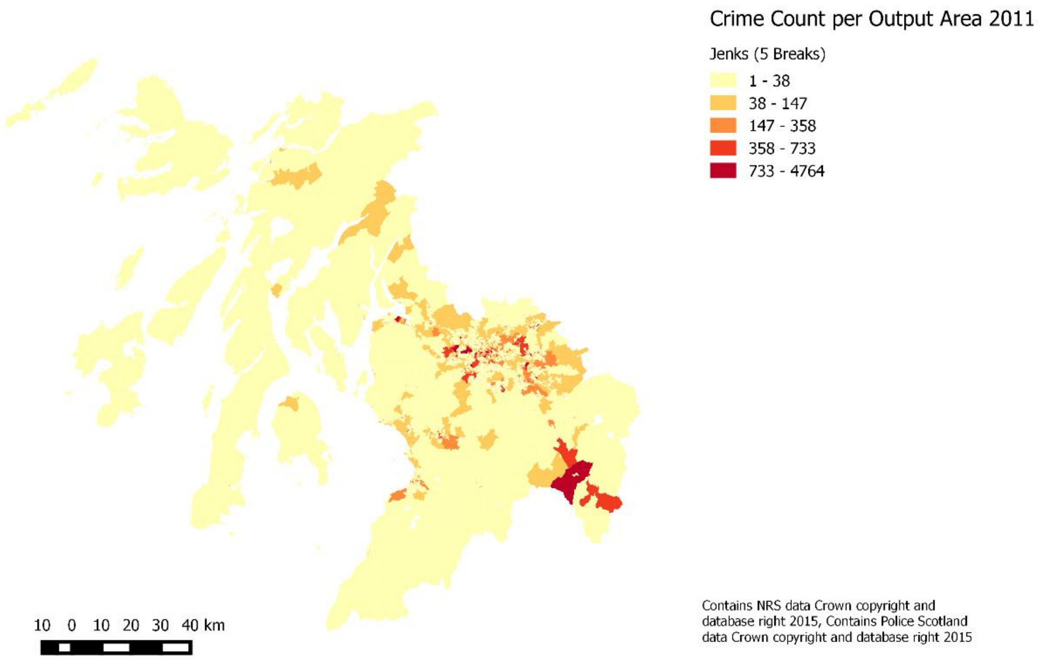
Using the equal method, as seen in Figure 5.6 shows a very different map to the quantile method. The output areas are mostly in the yellow or low crime category. There appears to only be a few in the medium (orange) or high (red) crime categories and these seem to be based in the centre of the map. This could highlight that there is only a very small number of areas that can be though of as being high crime.

Using the Jenk’s method, as seen in Figure 5.7, this again shows a very different map to the equal and quantile maps. The areas are mostly low crime areas (yellow) which is similar to the equal method map.   There are a mix of high and medium crime areas in the centre of the map corresponding to the Glasgow city centre area. There also appears to be another area with high crime at the bottom right of the map. It would be interesting to see if these hotspots are seen using any of the cluster methods. This suggests that a Jenk’s map can be useful when looking for hotspots in this situation as it is relatively easy to interpret results from this and it is similar to the k-means clustering method.



Contains NRS data Crown copyright and database right 2015, Contains Police Scotland data Crown copyright and database right 2015

Figure 5.6: Map of Strathclyde splitting by equal attribute intervals (output areas)



Contains NRS data Crown copyright and database right 2015, Contains Police Scotland data Crown copyright and database right 2015

Figure 5.7: Map of Strathclyde splitting by Jenk's intervals (output areas)

Methods

Table 5.5 provides a summary of the methods, variables used and whether the clusters are assumed to be spatially contiguous.

Table 5.5: Quick Reference Methods Guide

Method	Variable Used	Assumption of Spatial Contiguity
k-means	2011 Crime Rate	No
finite mixture modelling	2011 Crime Count	No
Local Moran's I	2011 Crime Count	Yes
Getis Ord Gi*	2011 Crime	Yes

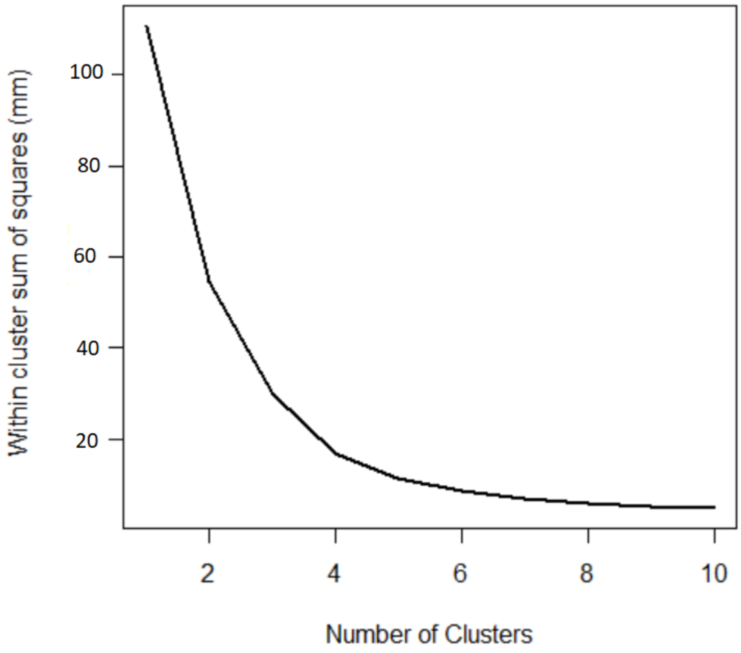
Output Areas Analysis

In this section, I will present the results for each of the four cluster/hotspot methods applied at the output area level of Strathclyde. The data were the crime counts and rates for 2011 aggregated to output area level which meant there were 19,886 observations in the data.

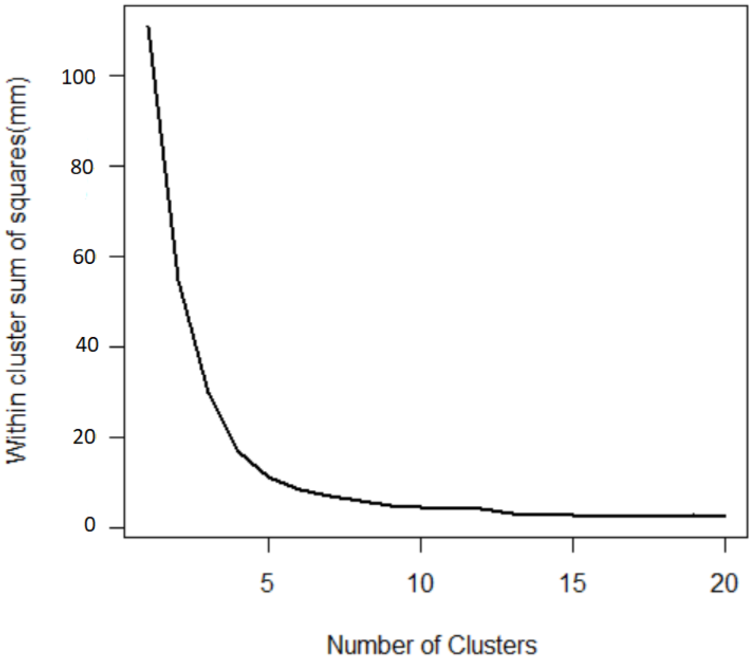
Methods with no assumption of spatial contiguity (OA)

k-means

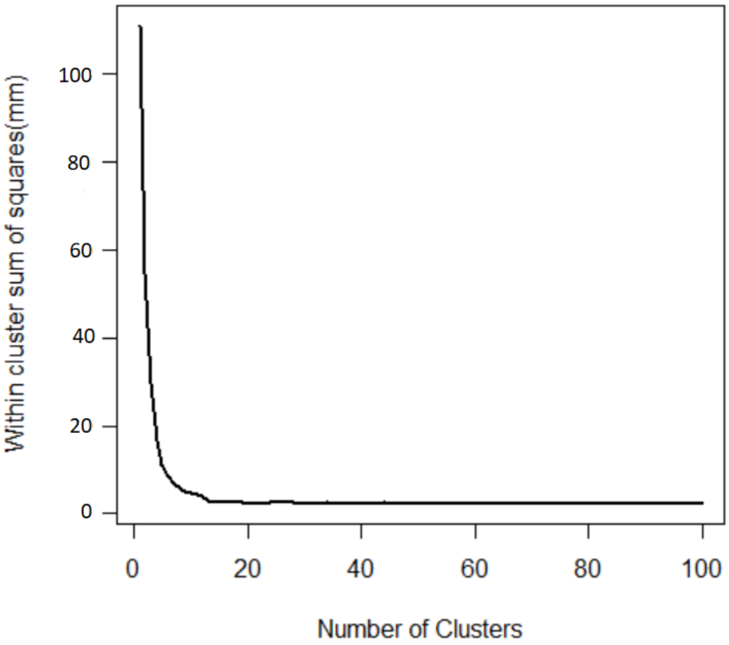
k-means was applied to the crime rates variable using the “k-means Output Area code” found in Appendix A. The within cluster sum of squares was calculated for the number of clusters from 1 to 10, from 1 to 20 and from 1 to 100. The elbow plots for the within cluster sum of squares vs the number of clusters can be seen in Figure 5.8.



(a) Elbow plot for max k=10



(b) Elbow plot for max k=20



(c) Elbow plot for max k=100

Figure 5.8: Elbow plots for k-means for max k=10, 20, and 100 at output area level

From Figure 5.8 (a), it can be subjective when identifying where the “elbow” appears to lie. For the researcher to choose k in this case, it can appear k=5 or k=6 could be the best fit. Therefore, it was decided to increase k to see if this would impact the elbow plot at all and the within group sum of squares for k-means when k=20 and k=100 was displayed in figure 5.8 (b) and (c).

As can be seen from the Figure 5.8 (b) and (c) elbow plots, there appears to be no difference from k=20 onwards suggesting that using a k of this size would be excessive and again the “elbow” appears to be near k=5 or k=6. Therefore, the best option subjectively appears to be k=6 as this is where it appears there is a substantial drop in the within cluster sum of squares. As this is subjective it is useful to look at the differences in the within cluster sum of squares to identify where the largest difference lies. Therefore, within the first 10 cluster groupings the optimal cluster number appears to be k=6. Once k=6 is selected as the optimal k, k-means is re-run on the data, the resultant groupings are shown in Table 5.6, ordered by the cluster means.

Table 5.6 shows the majority of output areas (91.21%) are very low crime areas in 2011 with a mean crime rate of 8 per 100 people. Less than 1% of the output areas appear as medium or higher crime areas. The highest crime cluster has a mean number of 3,258 crimes per 100 people for the 4 output areas contained within it. Given that there are only 4 output areas within this cluster, it would suggest these are outliers. When the high rates were checked these were found to lie mostly in the city centre area where the crime counts were highest and the population levels were low. Figure 5.9 shows the 2011 data in the six clusters.

Table 5.6: Descriptive statistics for clusters for k-means 6 cluster solution at output area level

Cluster Group	1	2	3	4	5	6
Count of Output Areas	18,139	1,410	258	58	17	4
Mean of Crime Rates in Output Areas	8.41	68.28	235.38	567.44	1244.85	3258.00
Cluster Type	Very Low crime	Low crime	Low-Medium crime	Medium	Medium-High crime	High crime
Percentage of Strathclyde Output Areas	91.21%	7.09%	1.30%	0.29%	0.09%	0.02%

In Figure 5.9, there are a mix of high and low crime areas next to each other, and if the clusters were required to be spatially contiguous, then there would be a far greater number of spatial clusters. However, it can be seen that the majority of the output areas are low crime areas. There are a large number of small areas concentrated in the middle of this region. This represents Glasgow city centre and Figure 5.10 shows a zoomed in version of this area. It can be seen that a large number of output areas lie within this relatively small spatial area.

A lot of the very low crime, low crime and low-medium crime groups can be seen in Figure 5.10 to lie on the outskirts of Glasgow city centre. It can be seen that there are many areas which are spatially contiguous to each other that are low crime areas (seen by the dark blue areas). There are, however, a few areas which have an abrupt change in crime rates as there are very low crime areas with neighbouring medium-high crime areas.

Figure 5.10 also shows mostly the medium to high crime clusters (peach colour) lie near the Glasgow city centre area of the map. This is not surprising given that a lot of crime literature identifies city centre areas as having high potential for criminal activity. It can also be seen that a lot of the high crime areas are spatially contiguous to either medium or medium-high crime areas. This suggests that crimes are occurring near to each other. k-means analysis shows that the vast majority of output areas across Strathclyde are low crime areas and the higher crime cluster lie in the Glasgow area.

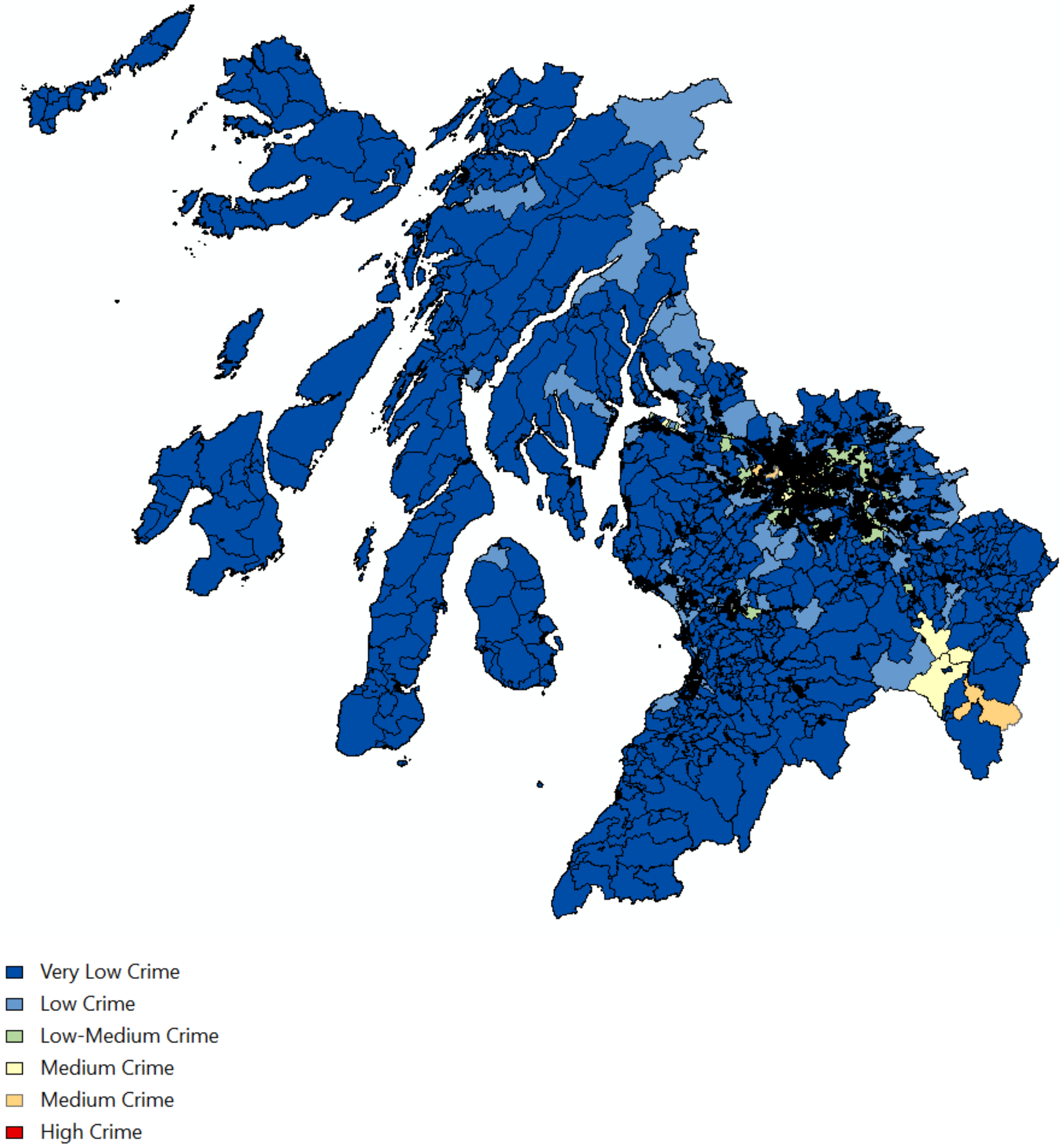


Figure 5.9: Map of clusters in Strathclyde for k-means 6 cluster solution at output area level

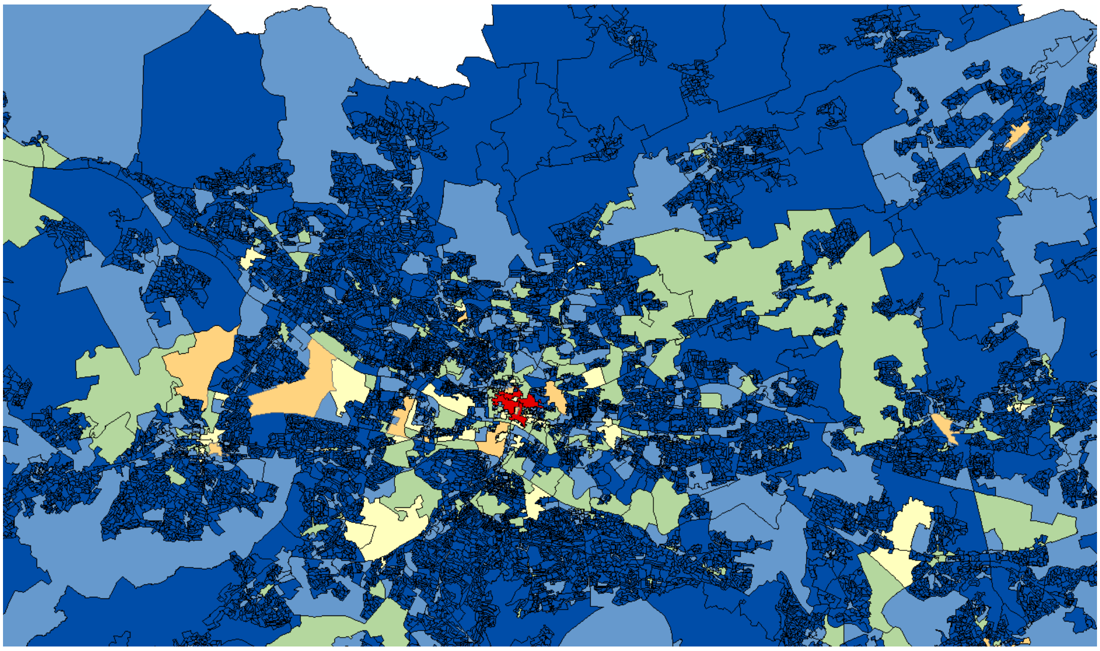


Figure 5.10: Map of clusters in Glasgow city centre for k-means 6 cluster solution at output area level

Finite Mixture Modelling

To carry out finite mixture model-based clustering on crime counts, I looked at the histogram of the data in Figure 5.4(b). From this it looks as if all data lies near the lower end of the scale with only a few outlier points lying above this suggesting there is likely only one or two clusters in the data e.g. low-medium crime vs high crime. This can be investigated by using the "flexmix" package in R to carry out finite mixture model-based clustering. Using a population offset was explored but since data are aggregated at output area level, this was not deemed necessary for this analysis as due to the construction of output areas most output areas have very similar population levels.

The mixture component distribution family was chosen to be Poisson as the data are counts and can never be negative. The full R coding used can be found in Appendix A. The maximum possible number of clusters was set to be 20 and then another was run using 10 as the maximum. In both cases, the AIC/BIC indicator suggested that the "best" number of clusters for the data were 2 clusters as seen in Figure 5.11. The two distinct clusters in the data corresponding to low and high crime areas can be seen in Table 5.7.

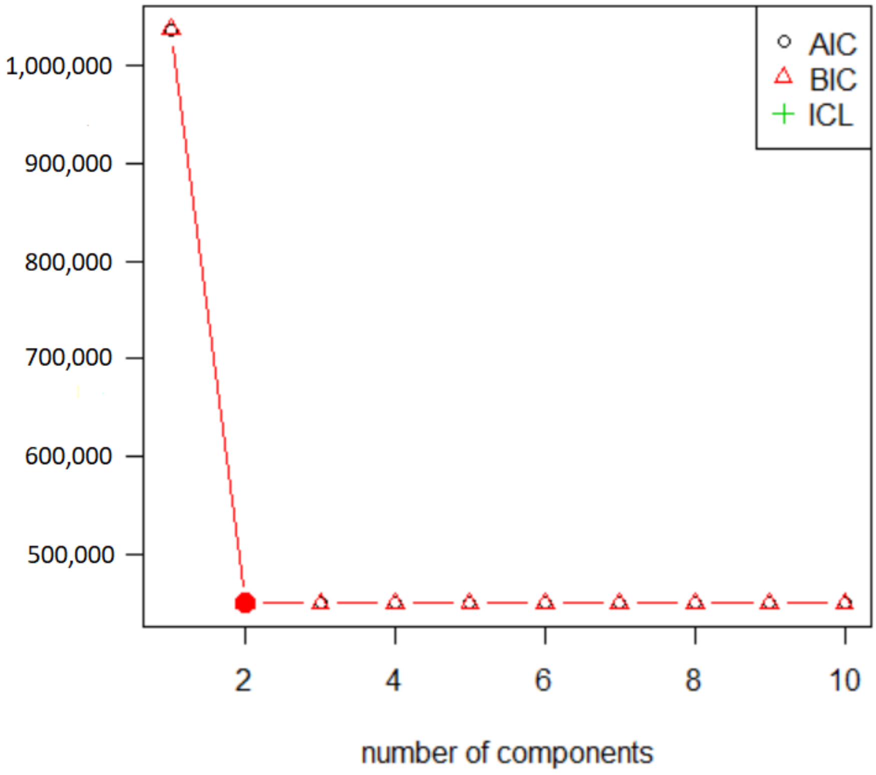


Figure 5.11: AIC/BIC/ICL plot for k=1 to k=10 using finite mixture models at output area level

Table 5.7: Descriptive statistics for clusters for finite mixture models 2 cluster solution for output areas

Cluster Group	1	2
Count of Output Areas	18,846	1,040
Mean of Crime Rates in Output Areas	10.44	179.96
Cluster Type	Low Crime	High Crime
Percentage of Strathclyde Output Areas	94.77%	5.23%

The output provides evidence of two clusters with the mean crime counts in each cluster as 10.44 and 179.96 respectively. This would suggest there is a low crime area cluster and a high crime cluster. It also highlights the differences with k-means clustering as this suggests that there are 6 distinct clusters in the data. The majority of output areas are considered low crime areas which is similar to the k-means cluster output as the majority of output areas there were low crime areas 91.21% for k-means and 94.77% for finite mixture models

The clusters are again not spatially contiguous as the high crime areas (red) in Figure 5.12 can be seen to be mainly in Glasgow city centre with a few outliers. Had spatial contiguity constraints been imposed on the data, there would have been a far greater number of spatial clusters.

Zooming in on Glasgow city centre (Figure 5.13), it can be seen that the majority of the high crime (red) output areas lie within this region.

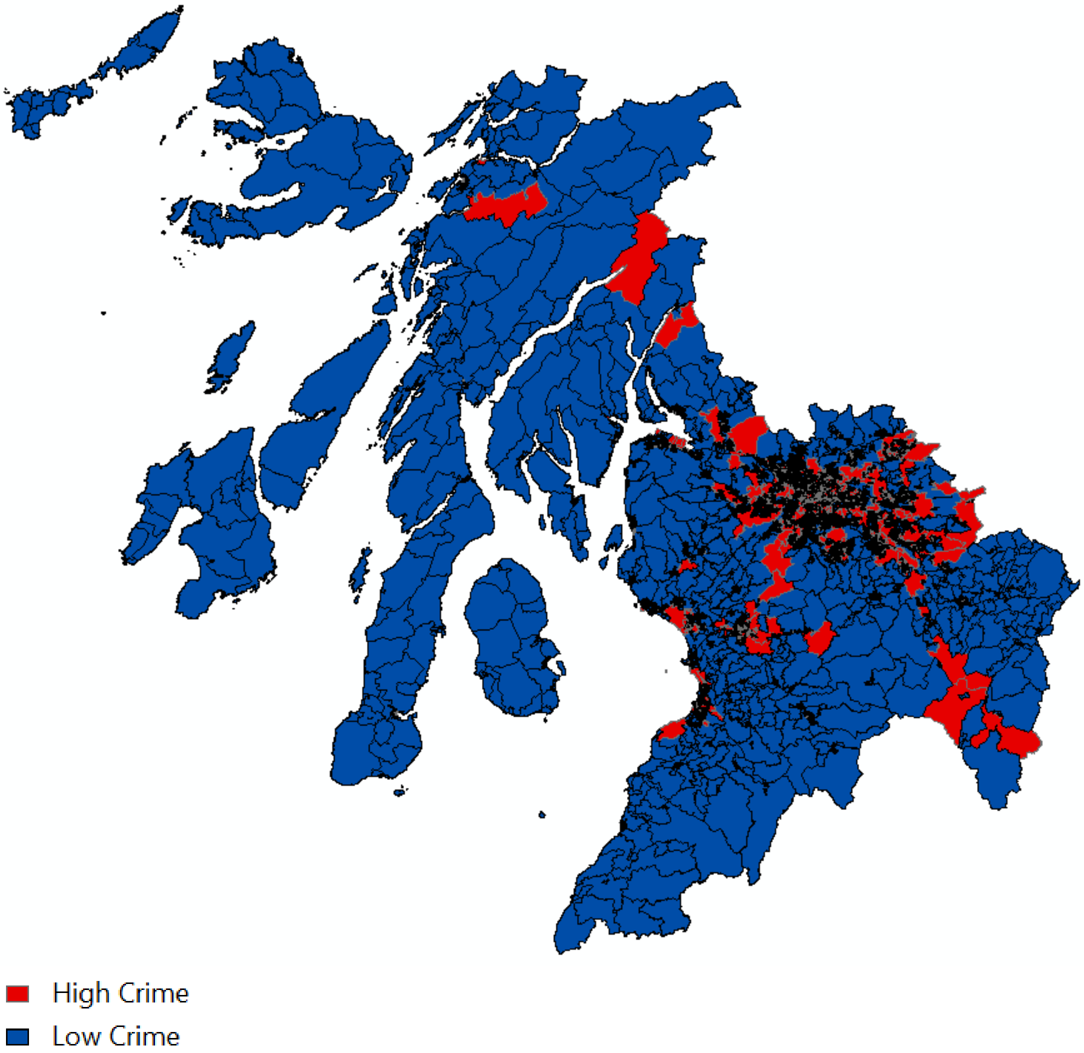


Figure 5.12: Map of clusters in Strathclyde for finite mixture models 2 cluster solution at output area level

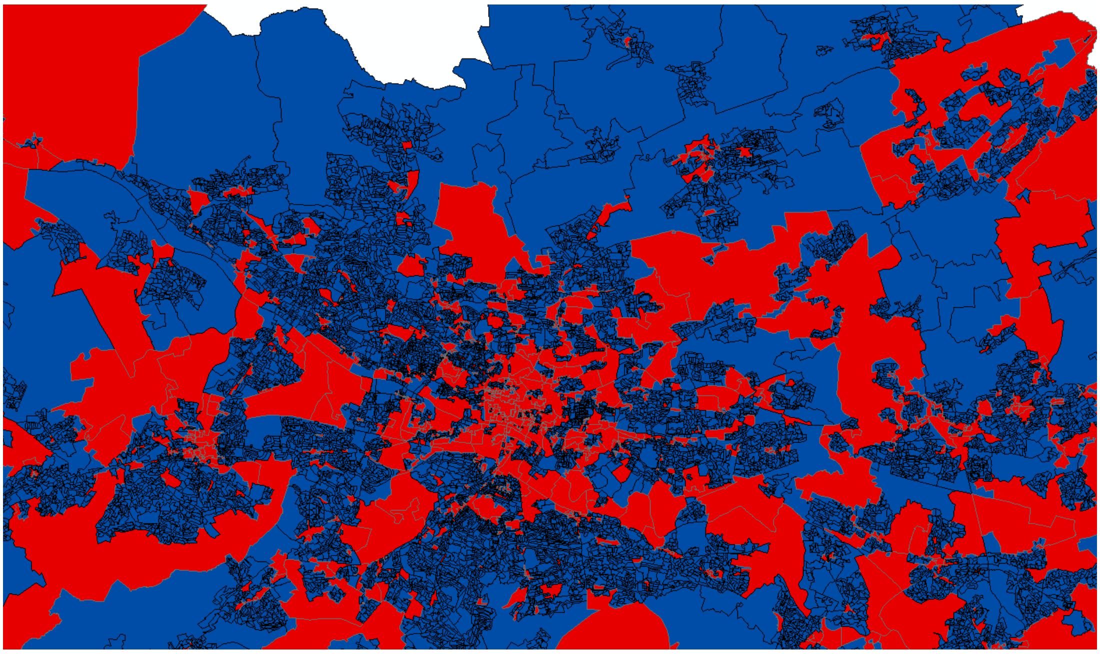


Figure 5.13: Map of clusters in Glasgow city centre for finite mixture models 2 cluster solution at output area level

Methods with assumption of spatial contiguity constraints (OA)

Local Moran's I

Local Moran's I can look at spatially contiguous clustering as it identifies high crime areas next to other high crime areas and similarly with low crime areas. This is carried out using ArcMAP software within ArcGIS and loading the 'Cluster and Outlier Analysis' toolset on the crime counts. The 'Cluster and Outlier Analysis' toolset is run using the default settings with weight matrix of 1 if the areas border each other and 0 otherwise. This is the same way police analysts use this software as described by the ex-police analysts.

Table 5.8 highlights the difference in this method compared to both k-means and finite mixture modelling. The lowest crime cluster here has only 6.7% of output areas belonging to it, unlike the previous methods which had over 90% of output areas lying in this cluster. This is because Local Moran's I look for low crime areas to be neighbours (contiguous) to other low crime areas before it will put these in the same cluster. As can be seen from Table 5.8, most of the output areas (81.01%) lie in the non-significant cluster, suggesting that these are neither high or low crime areas in relation to their neighbours and can be thought of as the medium crime category. This medium crime cluster has an average of 14.3 crimes per output area which is almost 5 times as many crimes on average in the low crime cluster. The highest crime cluster has just over 5% of output areas belonging to it and has an average of 133 crimes per output areas which is a great deal more crimes than any of the other clusters. Figure 5.11 shows the map of these clusters.

Table 5.8: Descriptive statistics for clusters for Local Moran's I for output areas

Count of Output Areas	1,333	1,339	16,110	94	1,010
Mean of Crime Counts in Output Areas	3.50	7.65	14.30	42.44	133.37
Cluster Type	Low-Low Cluster	Low-High Outlier	Not Significant	High-Low Outliers	High-High Cluster
Percentage of Strathclyde Output Areas	6.70%	6.73%	81.01%	0.47%	5.08%

Figure 5.14 shows the Local Moran's I output for the crime counts for 2011 output areas. As can be seen from the clustering map, this provides different clustering analysis to the previous k-means and finite mixture modelling results. There still appears to be a high crime cluster area in the Glasgow city centre area similar to the other clustering methods. However, this method identifies if this is an outlier output area or if this is a cluster of areas which are spatially contiguous and would form a larger cluster. As can be seen from the plot, the vast majority of areas are part of the "Not Significant" cluster which suggests that these areas have neither high or low crime counts in relation to their neighbours. The areas of particular interest are the High-High cluster groups(red) and the High-Low outlier groups (peach) as this will show hotspots.

Figure 5.14 shows that there are a number of H-H cluster groups in the centre of the map (Glasgow city centre) which will be explored in Figure 5.15 in more detail. Also, of interest are the areas which are seen in the dark blue which are L-L clusters of output areas. These show the potential for coldspots to be identified in this area as they are low crime areas surrounded by other low areas. For the purposes of this thesis, the H-H crime areas are one cluster, H-L is another cluster, L-H is another and L-L are another cluster. This would show the data split into 5 clusters (with the majority of output areas lying in the non-significant category with neither high nor low crime counts).

Figure 5.15 shows that the majority of the High-High (red) crime clusters lie within the Glasgow city centre area. These seem to be surrounded by dome Low-High (light blue) crime areas too. There are also a number of non-significant output areas (yellow) which suggest that these are neither high or low crime areas in relation to their neighbours.

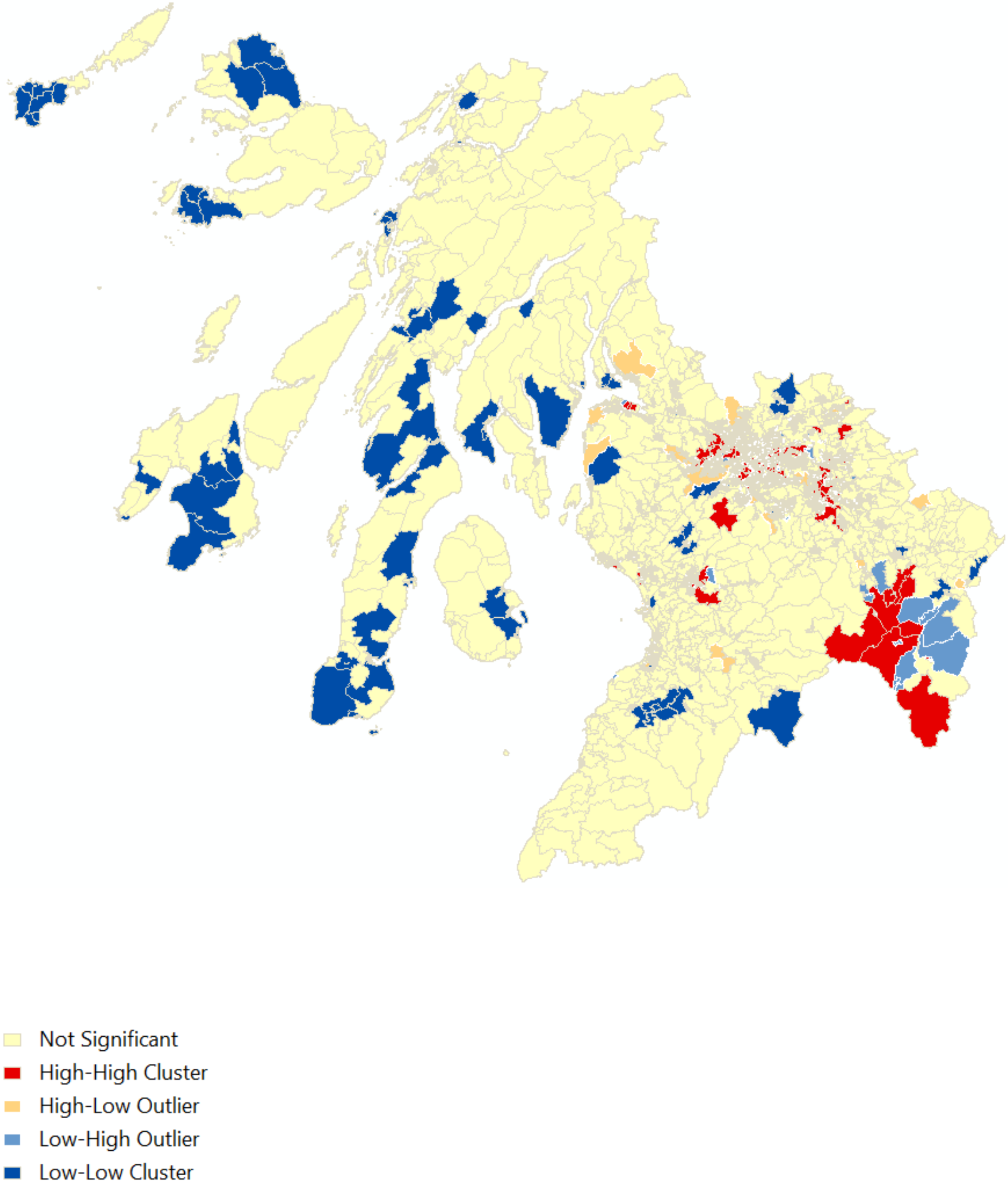


Figure 5.14: Map of clusters in Strathclyde for Local Moran's I at output area level

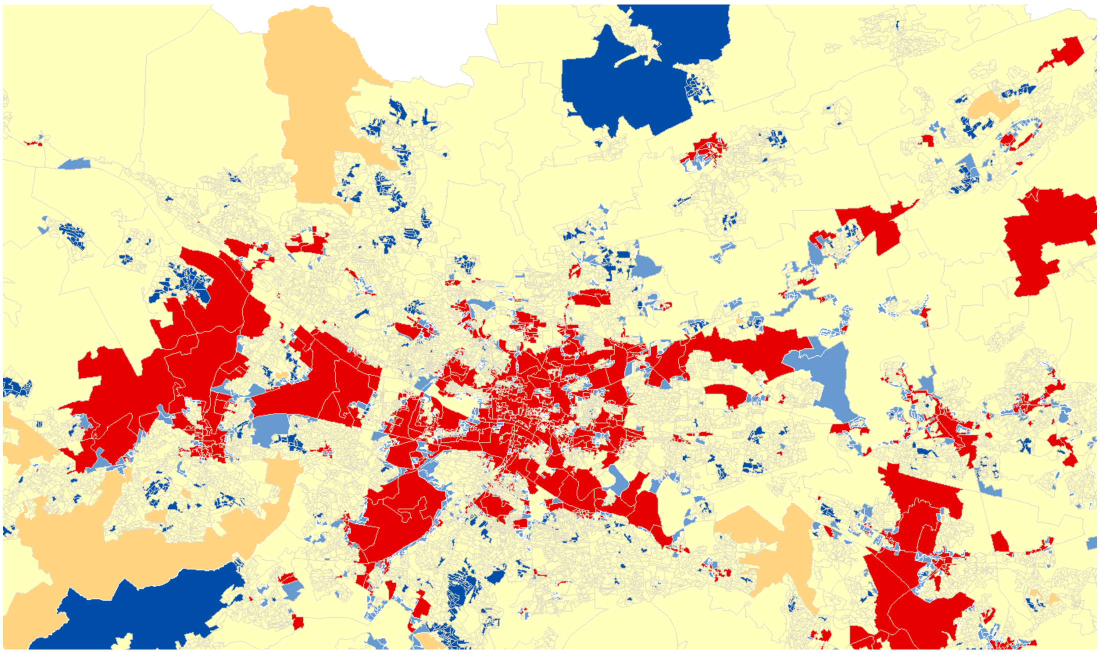


Figure 5.15: Map of clusters in Glasgow city centre for Local Moran's I at output area level

Getis Ord Gi*

Similarly to Local Moran's I, Getis Ord Gi* looks at spatially contiguous clustering as it looks at nearby areas to identify high and low clusters. The hotspots analysis option in ArcMAP within ArcGIS produced the map shown in Figure 5.16 and the descriptive statistics for the clusters are seen in Table 5.9. The default options were used with weight matrix of 1 if the areas border each other and 0 otherwise.

Table 5.9: Descriptive statistics for clusters for Getis Ord Gi* for output areas

Count of Output Areas	18,311	309	349	917
Mean of Crime Counts in Output Areas	12.93	39.86	48.44	128.66
Cluster Type	Low crime (Not Significant)	Low-Medium Crime (Hot Spot 90% significance)	Medium Crime (Hot Spot 95% significance)	High Crime (Hot Spot 99% significance)
Percentage of Strathclyde Output Areas	92.08%	1.55%	1.76%	4.61%

Table 5.9 shows that, similarly to k-means (91.21%) and finite mixture modelling (94.77%), the majority of output areas are in the lowest crime cluster (92.08%). The output areas range from low crime areas with only 13 crimes on average to high crime areas which have 129 crimes per output area on average. Figure 5.16 can be used to see if these correspond to similar areas as k-means and finite mixture modelling.

Figure 5.16 shows both hot and cold spots. Coldspots are areas which are clustered together which have low crime counts. There are no apparent cold spots in the area of Strathclyde as there are only not significant areas and hotspots with 90%, 95%, and 99% significance. It appears as though most of the output areas are not significant in the outlying areas which means they are neither significantly high crime areas nor low crime areas and could be considered in this case to be low or low-medium crime areas as the mean for this cluster is only 13 crimes on average per output area which is relatively low compared to the other cluster means.

As there is clearly a hotspot with 99% confidence level in Glasgow city centre (red) with a mixture of coldspots and non-significant output areas surrounding it. There also appears to be a hotspot with 99% confidence in the bottom right-hand side of the map. As this is so significant in this map, it is of interest to look at the other maps from k-means, finite mixture modelling and Local Moran's I analysis to see if they identified this area as a hotspot as well. Both k-means and finite mixture models highlighted this area as a low crime area surrounded by some high crime areas and thus did not highlight the whole area as a hotspot. Local Moran's I identified this as a High-High crime area similar to Getis Ord Gi*. This highlights the affect methods can have on the clusters identified. Through the non-spatially contiguous methods, only part of this area was identified as a high crime area or hotspot. While using methods with spatial contiguity constraints, there is a larger area identified as a hotspots (slightly larger an area with Gi* then Local Moran's I).

Figure 5.17 shows the map zoomed in to Glasgow city centre output areas and this shows almost all of the hotspot with 99% significance group lies in this area. There are a lot of non-significant areas there as well.

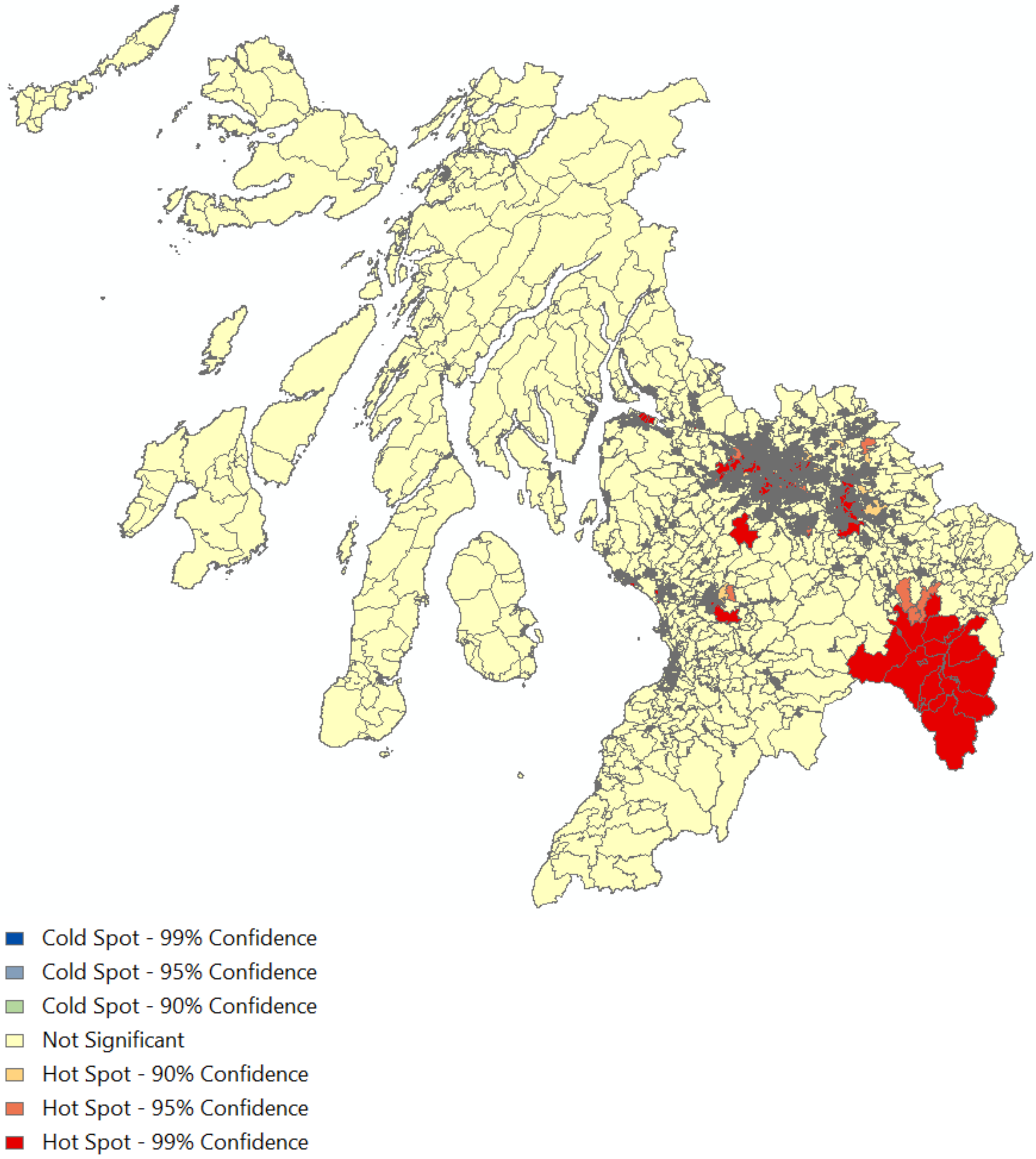

Cold Spot - 99% Confidence
Cold Spot - 95% Confidence
Cold Spot - 90% Confidence
Not Significant
Hot Spot - 90% Confidence
Hot Spot - 95% Confidence
Hot Spot - 99% Confidence

Figure 5.16: Map of clusters in Strathclyde for Getis Ord Gi* at output area level

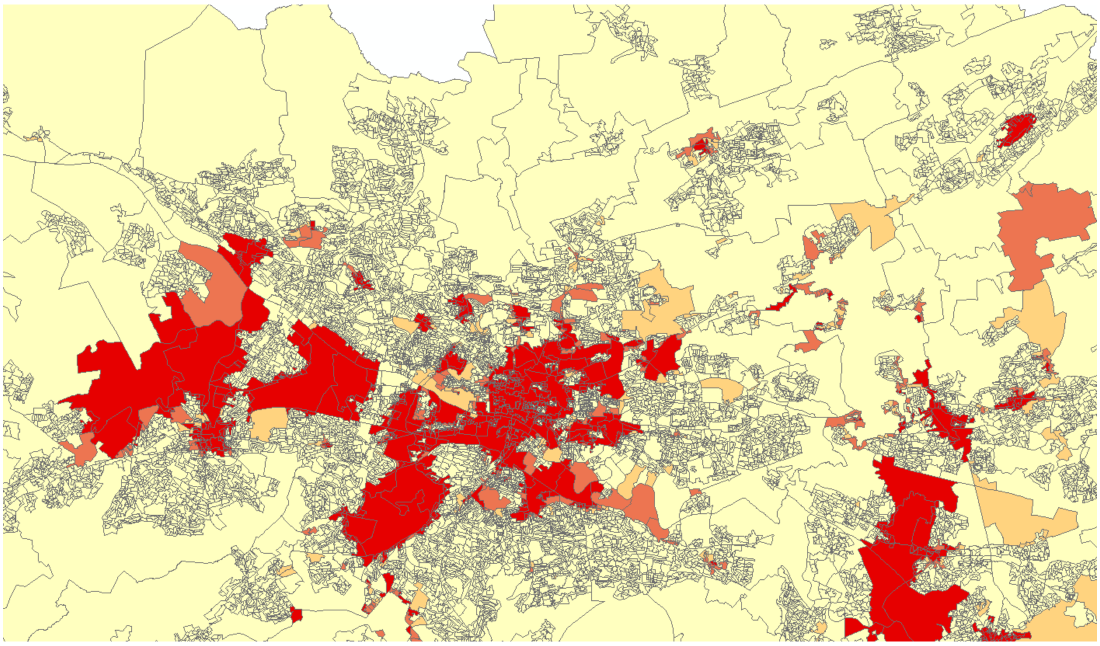
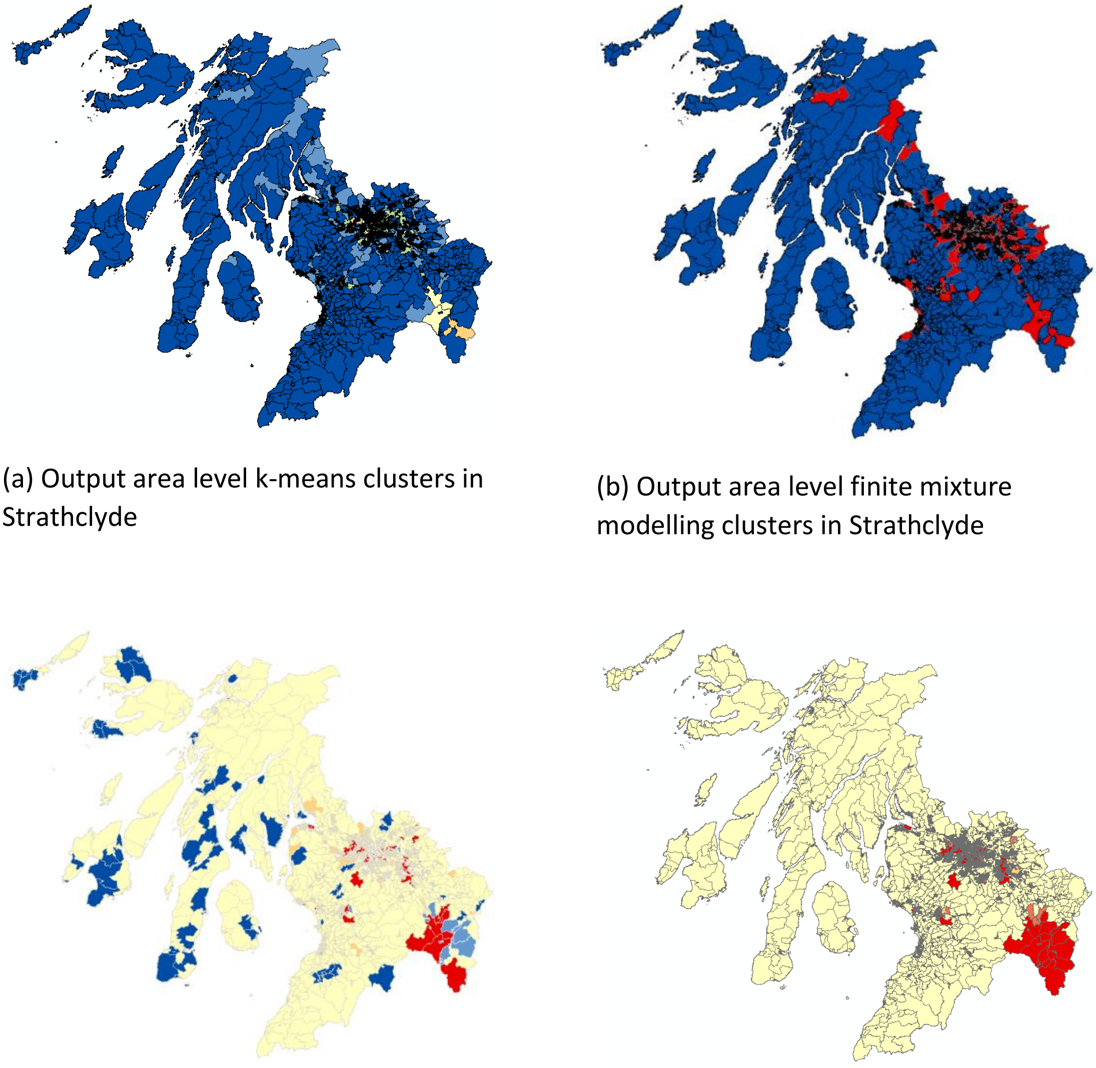


Figure 5.17: Map of clusters in Glasgow city centre for Getis Ord Gi* at output area level

Therefore, for the output area level analysis, it would appear that each method produces different results. However, there are some common themes such as Glasgow city centre is highlighted as having the most hotspots within it for each method used. These are looked at in greater detail in the next comparison sections. This will allow us to see if the maps appear to be consistent at both the output area and data zone levels. The ARI will be calculated to assess the cluster groupings at each method.

Strathclyde Comparison Maps (OA)

Figure 5.18 gives an overview of the maps produced by each of the clustering methods at the output area level. Figures 5.18(a) and (b) show that the dark blue (low) crime areas appears to be similar in both plots and lie on the surrounding areas of Strathclyde including the Islands. In Figures 5.18(c) and (d), the yellow areas show the non-significant category cluster which is similar in both maps. However, (c) has some more light blue areas suggesting there are a few low-medium crime areas bordering high crime areas (L-H outlier).



(a) Output area level k-means clusters in Strathclyde

(b) Output area level finite mixture modelling clusters in Strathclyde

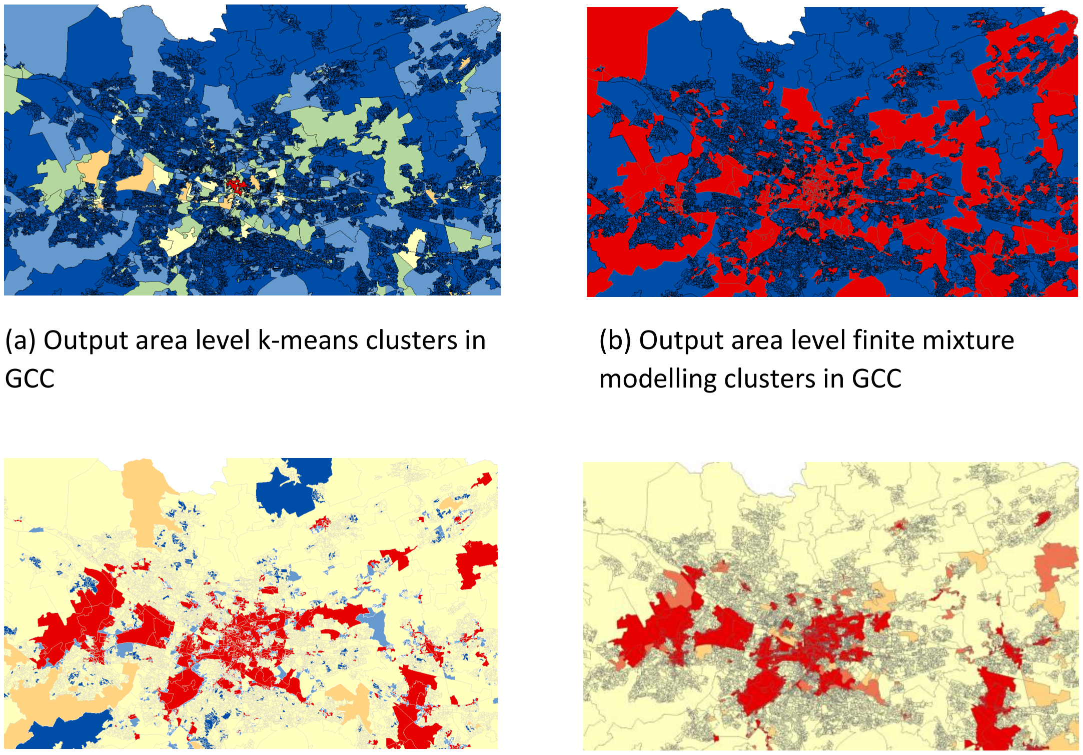
(c) Output area level Local Moran's I clusters in Strathclyde

(d) Output area level Getis Ord Gi* clusters in Strathclyde

Figure 5.18: Clusters for each method at output area level in Strathclyde

All graphs in Figure 5.18 highlight the bottom right area as a medium to high crime area in Figure 5.18(a) this is seen as peach, in (b), (c) and (d) this is seen as red. The Gi* map (d) highlights all of the output areas in this area as being a high crime cluster (red area). However, the Local Moran's I map only highlights some of them as being a High-Low outlier (peach) and the other nearby output areasas being Low-Low cluster suggesting there is one high crime output area surrounded by neighbouring low crime areas. Local Moran's I output is also similar to the k-means and finite mixture modelling maps as they both show the same output areas near the top of the cluster as being medium to high crime areas while the surrounding areas are low.

Glasgow City Centre (GCC) Comparison Maps (OA)



(a) Output area level k-means clusters in GCC

(b) Output area level finite mixture modelling clusters in GCC

(c) Output area level Local Moran's I clusters in GCC

(d) Output area level Getis Ord Gi* clusters in GCC

Figure 5.19: Clusters for each method at output area level in Glasgow City Centre area

It is useful to take a closer look at the Glasgow city centre (GCC) area as it is hard to see in the Strathclyde map due to having a large number of output areas in a small geographical space. Figure 5.19(a) shows that the 4 output areas in the very high crime cluster (pink) all lie together in the centre of the GCC area which is expected given the majority of high crime areas appear to lie in this small space for each method. As Figure 5.19(b) shows, there are a lot of output areas in this area which belong to the high crime cluster from the finite mixture modelling analysis. This shows that when the data is split into only 2 clusters, GCC area appears to be where most high crime areas lie (red areas).

From Figures 5.19(c) and (d), it appears as that there are a few output areas which are highlighted in both maps, these are seen as red in both maps. These can be seen across the centre of the maps and near the bottom right of the maps. These areas are highlighted as part of the High-High cluster group (high) crime cluster by Local Moran's I and as part of the high crime hotspot cluster (with 99% significance) by Gi*. This shows that both of these methods appear to have similar groupings in the clusters.

Adjusted Rand Index (OA)

The adjusted Rand index was calculated for the clusters identified by each of these methods at the output area level and the values can be seen in Table 5.10. The two methods which produce the most similar cluster groups are k-means and finite mixture models with an ARI of 0.676. The next closet are Local Moran's I and Getis Ord Gi* with an ARI of 0.477. These are both highlighted in pale green in the Table 5.10. This suggests that the methods with no assumption of spatially contiguity were similar to each other as were the two methods with spatial contiguity constraints. When comparing the clusterings produced by k-means to the clusterings produced by the two methods with spatial contiguity constraints, it can be seen that the ARI value dropped to 0.134 and 0.287 respectively. This suggests that these methods do not produce very similar clusterings. When finite mixture modelling is compared to the Local Moran's I and Getis Ord GI* cluster groupings, these again are very low values of less than 0.31 suggesting that these are quite different cluster outputs.

This implies that there is some variability in the methods as they produce very different results. The differences could be due to the numbers of clusters being different in most of the methods so further work could look at having the same number of clusters for each method and identifying if this would produce more similarities between groups for the different methods.

Table 5.10: Adjusted rand index for clusters at output area level

Output Areas	k-means	finite mixture modelling	Local Moran's I	Getis Ord Gi*
k-means	1.000	0.676	0.134	0.287
finite mixture modelling	0.676	1.000	0.116	0.305
Local Moran's I	0.134	0.116	1.000	0.477
Getis Ord Gi*	0.287	0.305	0.477	1.000

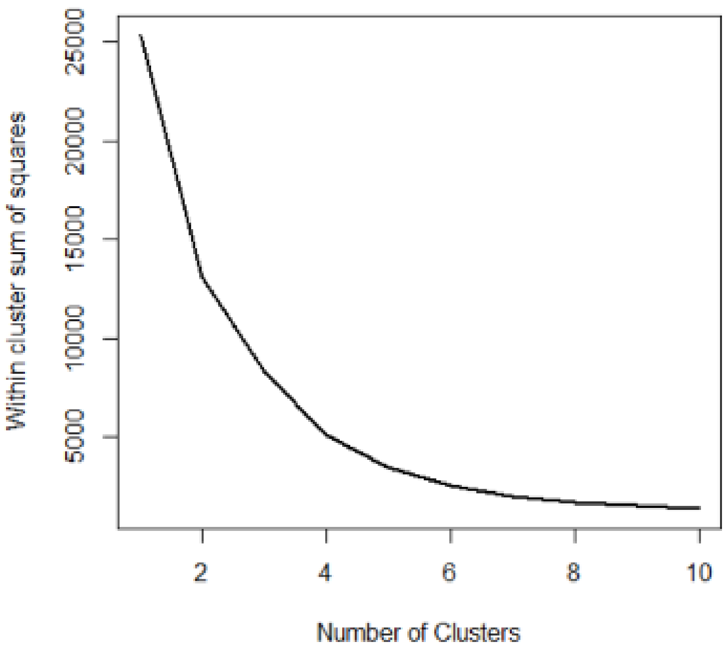
Data Zones Analysis

In this section, I will present the results for each of the four cluster/hotspot methods applied at the data zone areal level of Strathclyde. The data were the crime counts and rates for 2011 aggregated to data zone level which meant there were 2,963 observations in the data.

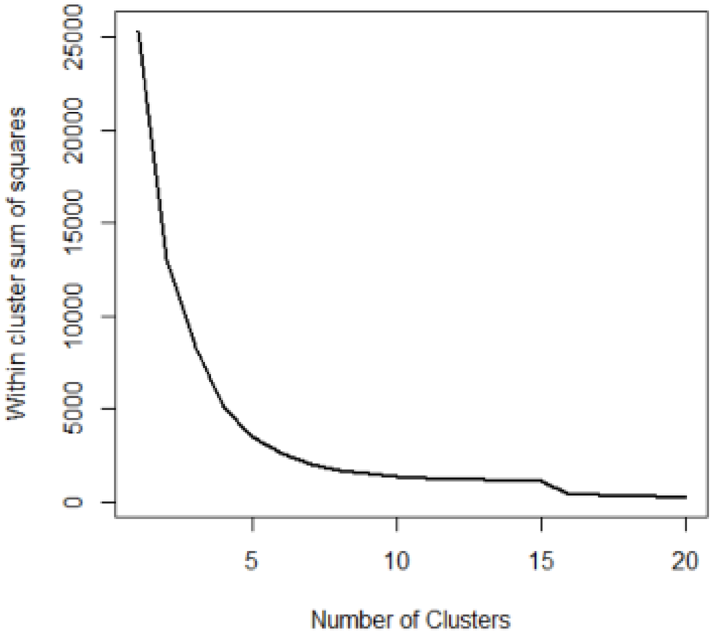
Methods with no assumption of spatial contiguity (DZ)

k-means

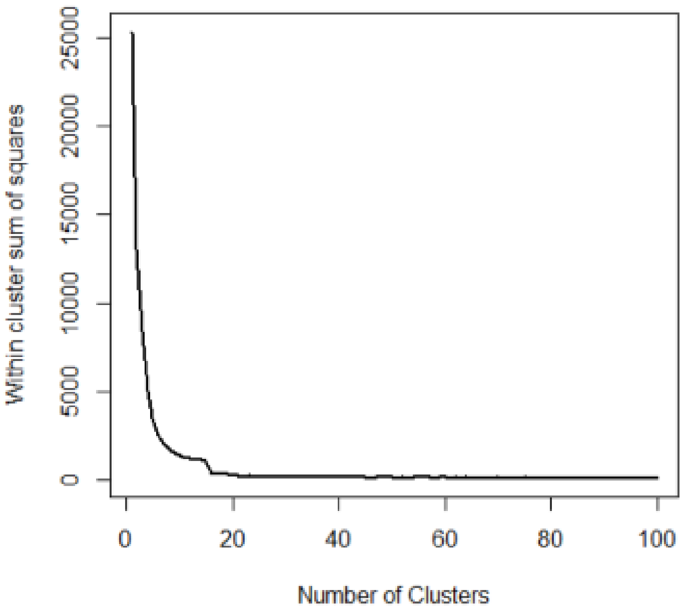
k-means was applied to the crime rates variable using the “k-means Data Zone code” found in Appendix A. The within cluster sum of squares was calculated for the number of clusters from 1 to 10, from 1 to 20 and from 1 to 100. The elbow plots for the within cluster sum of squares vs the number of clusters can be seen in Figure 5.20,



(a) Elbow plot for max k=10



(b) Elbow plot for max k=20



(c) Elbow plot for max k=100

Figure 5.20: Elbow plots for k-means for max k=10, 20, and 100 for data zones

From Figure 5.20 (a), it can be subjective when identifying where the "elbow" appears to lie. For the researcher to choose k in this case, it can appear k=5 or k=6 could be the best fit. Therefore, it was decided to increase k to see if this would impact the elbow plot at all and the within cluster sum of squares for k-means when k=20 and k=100 was displayed in figure 5.20 (b) and (c).

As can be seen from the Figure 5.20 (b) and (c) elbow plots, there appears to be no difference from k=20 onwards suggesting that using a k of this size would be excessive and there appears to be two "elbows", one at k=5 and one at k=16. Therefore, the best option appears to be k=5 as this is where it appears there is a substantial drop in the within cluster sum of squares. Therefore, within the first 10 cluster groupings the optimal cluster number appears to be k=5. Once k=5 is selected as the optimal k, k-means is re-run on the data, the groupings are as shown in Table 5.11.

Table 5.11 shows the majority of data zones are low (1,109) or medium-low (1,117) crime areas in 2011 with a mean crime rate of 2 per 100 people or 4 per 100 people. Less than 4% of the data zones appear as medium-high or high crime areas. None of the cluster types appear to have significantly high crime rates per 100 people as the high crime cluster has 4 data zones and a mean crime rate of 40 crimes per 100 people. Figure 5.21 shows the 2011 data split in to five clusters discovered by k-means.

Table 5.11: Descriptive statistics for clusters for k-means 5 cluster solution for data zone level

Cluster Group	1	2	3	4	5
Count of Data Zones	1,109	1,117	627	106	4
Mean of Crime Rates in Data Zones	2.11	4.35	6.99	12.35	40.28
Cluster Type	Low crime	Low-Medium crime	Medium	Medium-High crime	High crime
Percentage of Strathclyde Data Zones	37.43%	37.70%	21.16%	3.58%	0.13%

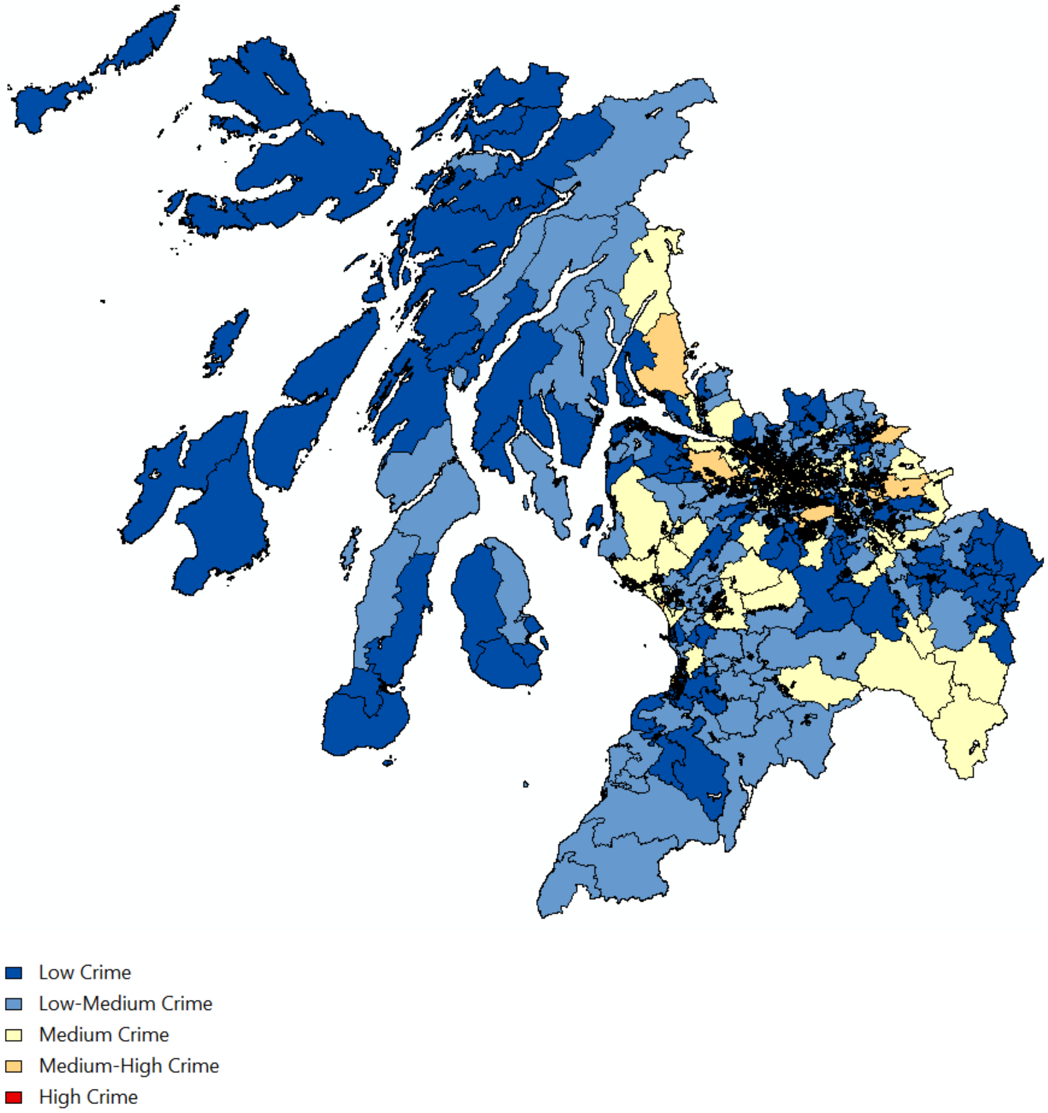


Figure 5.21: Map of clusters in Strathclyde for k-means 5 cluster solution at data zone level

As can be seen from Figure 5.21, most of the low and medium-low crime groups (dark and light blue) are around the outskirts of the map and the islands area of Strathclyde. This is again similar to the output area level maps produced by k-means. Also, the majority of the medium-high and high crime groups appear to be in the Glasgow city centre area as can be seen in Figure 5.22.

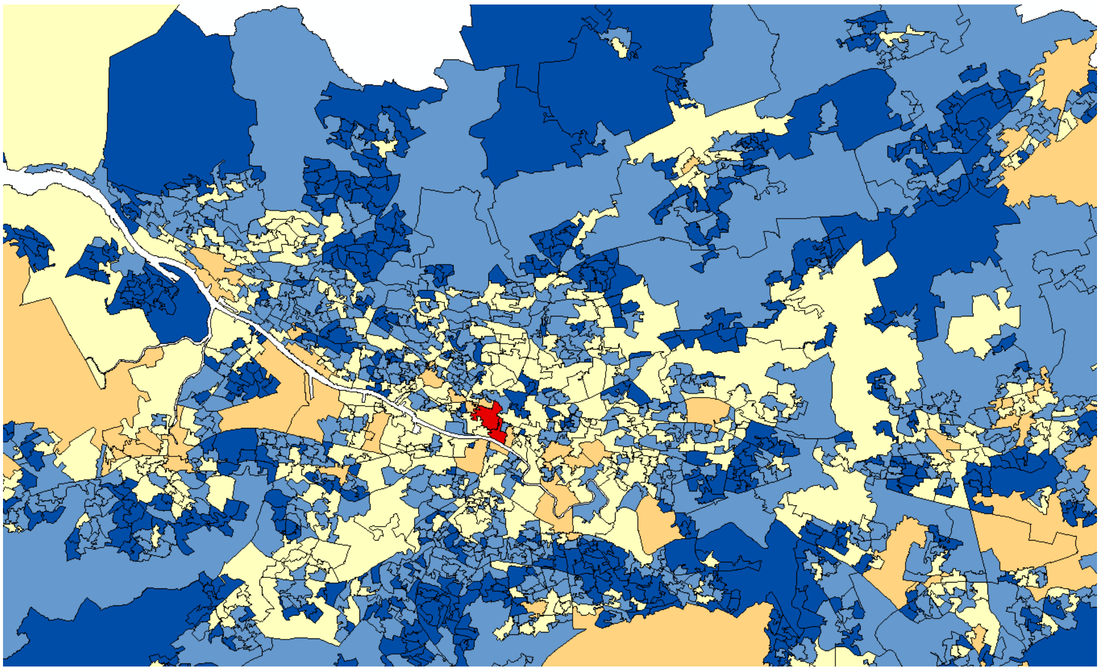


Figure 5.22: Map of clusters in Glasgow city centre for k-means 5 cluster solution at data zone level

In Figure 5.22, it can be seen that two of the high crime data zones lie within the Glasgow city centre area (seen as the red in the centre of the figure). This appears similar to the k-means analysis at output area (and finite mixture models, Local Morans's I and Getis Ord Gi*), as most of the high crime groups were in the Glasgow city centre for each method used at the output area level.

Finite Mixture Modelling

To carry out finite mixture model-based clustering, I again utilised the "flexmix" package in R. The mixture component distribution family was again chosen to be Poisson as the data are counts and can never be negative. The model was run on the counts of crime data at the data zone areal level. The full R coding used can be found in Appendix A. The maximum number of clusters was set to 10. The AIC/BIC criteria were then plotted to identify the "best" number of clusters for the data. The plot for AIC/BIC can be seen in Figure 5.23. this shows that after k=2 and k=3 the AIC/BIC values both decrease and thereafter, there do not appear to be any substantial decreases suggesting that 3 clusters would be the optimum value. The output can be seen in Table 5.12.

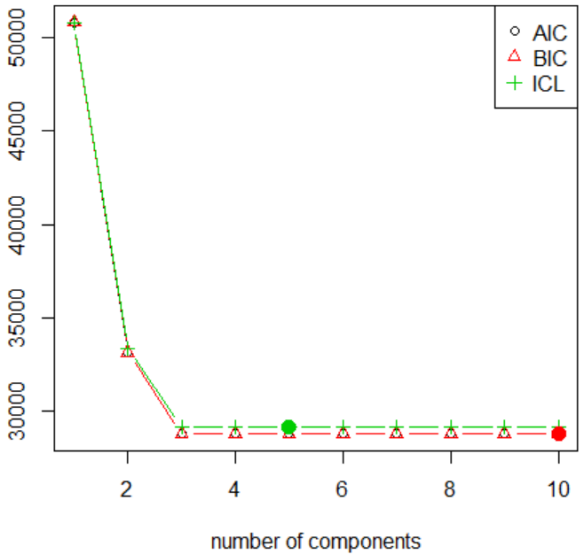


Figure 5.23: AIC/BIC/ICL plot for k=1 to k=10 clusters using finite mixture models at data zone level

Table 5.12: Descriptive statistics for clusters for finite mixture models 3 cluster solution at data zone level

Cluster Group	1	2	3
Count of Data Zones	1,229	1,454	280
Mean of Crime Counts in Data Zones	16.16	38.65	78.94
Cluster Type	Low Crime	Medium Crime	High Crime
Percentage of Strathclyde Data Zones	41.48%	49.07%	9.45%

The output provides evidence of three clusters with the mean crime counts in each cluster as 16.16, 38.65 and 78.94 respectively. This would suggest there is a low crime area cluster, a medium crime area cluster and a high crime area cluster. It also highlights that both the low and medium crime clusters have similar mean crime counts. This shows the differences with k-means clustering which split the data into 5 distinct clusters at the data zone level. However, the majority of data zones are considered low crime areas which is similar to the k-means cluster output.

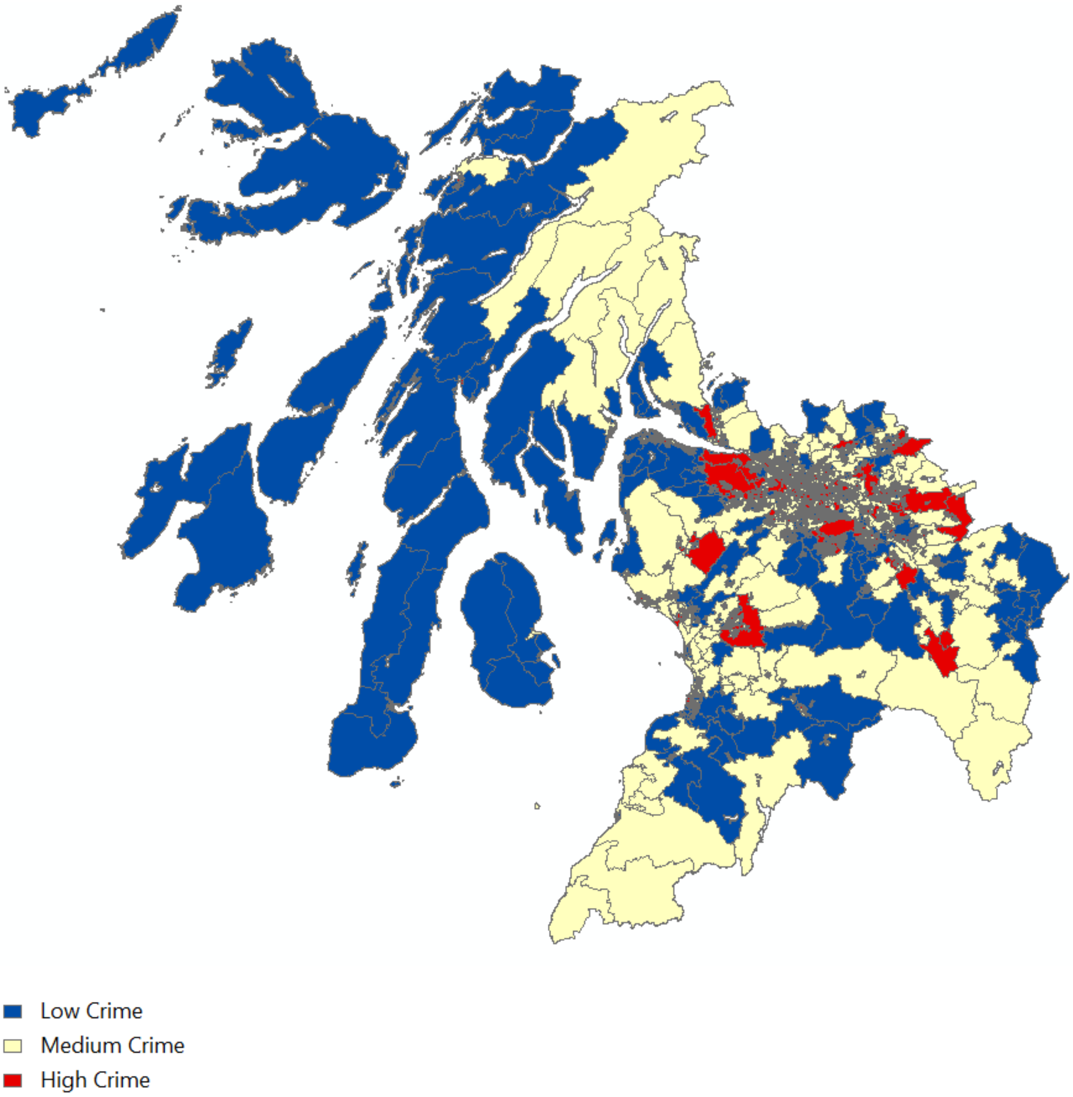


Figure 5.24: Map of cluster groupings in Strathclyde for finite mixture models 3 cluster solution at data zone level

From Figure 5.24, it appears as though the majority of high crime areas are in Glasgow city centre and the low crime areas are on the outskirts and the islands. This is similar to the majority of other clustering methods at both the output area and data zone areal levels.

Zooming in on Glasgow city centre (Figure 5.25), seems to have a lot of medium crime areas (yellow) which is not surprising given most of the data zones are either low or medium crime areas. There is also a mixture of low and high crime data zones as well. This is very similar to the hotspot clusters identified at the output area level and at k-means at the data zone areal level where the majority of the high crime cluster (hotspots) were in the Glasgow city centre region.

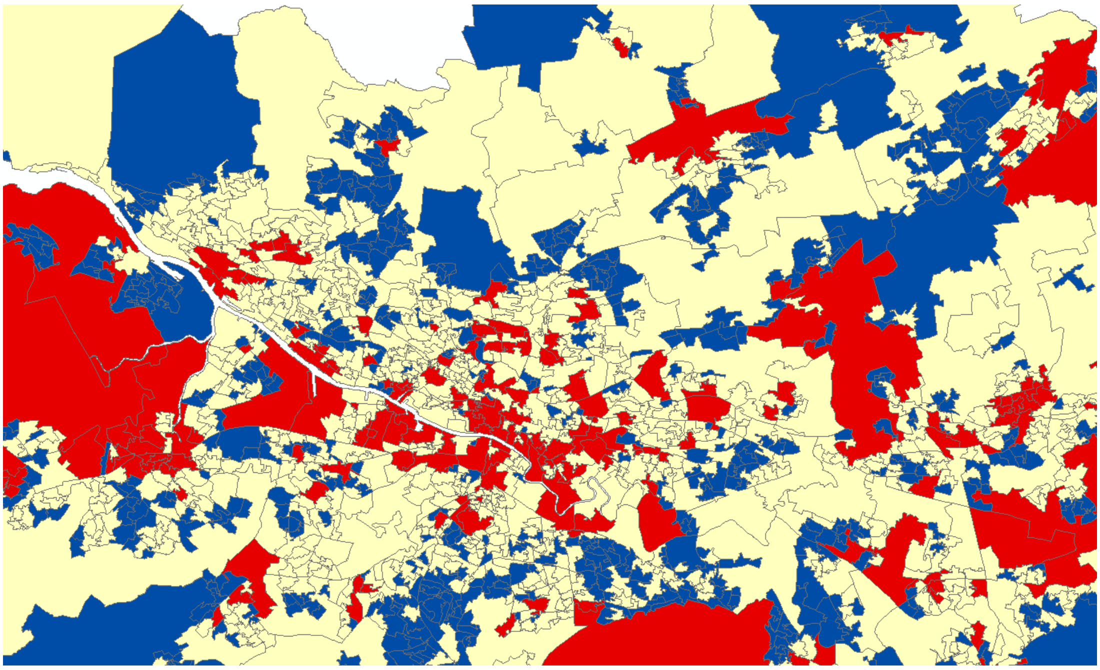


Figure 5.25: Map of cluster groupings in Glasgow city centre for finite mixture modelling 3 cluster solution at data zone level

Methods with assumption of spatial contiguity constraints (DZ)

Local Moran's I

Local Moran's I was carried out using ArcMAP software within ArcGIS and loading the 'Cluster and Outlier Analysis' toolset on the crime counts. The 'Cluster and Outlier Analysis' toolset is run using the default settings with weight matrix of 1 if the areas border each other and 0 otherwise.

Table 5.13: Descriptive statistics for clusters for Local Moran's I cluster solution at data zone level

Cluster Group	1	2	3	4	5
Count of Data Zones	230	102	2287	26	318
Mean of Crime Counts in Data Zones	15.01	24.58	31.20	49.85	61.48
Cluster Type	Low crime	Low-Medium crime	Medium	Medium-High crime	High crime
Percentage of Strathclyde Data Zones	7.76%	3.44%	77.19%	0.88%	10.73%

Table 5.13 shows that there is a difference between this method and the other cluster methods at the data zone level. Most data zones lie in the low crime cluster for all other

methods (>90%) but for Local Moran's I, most lie in the non-significant cluster (77.19%). The mean crime counts for each cluster vary from 15 crimes (low) to 61 crimes (high) which shows that almost 4 times as many crimes occur in the high crime cluster output areas as in the low crime cluster output areas. This table also shows that there are over 10% of the data zones in the high crime category. This is quite a large number of output areas to be in the high crime cluster compared to k-means where less than 1% of output areas were in this cluster. But this does show a similarity to finite mixture modelling where over 9% of the output areas were in the high crime cluster.

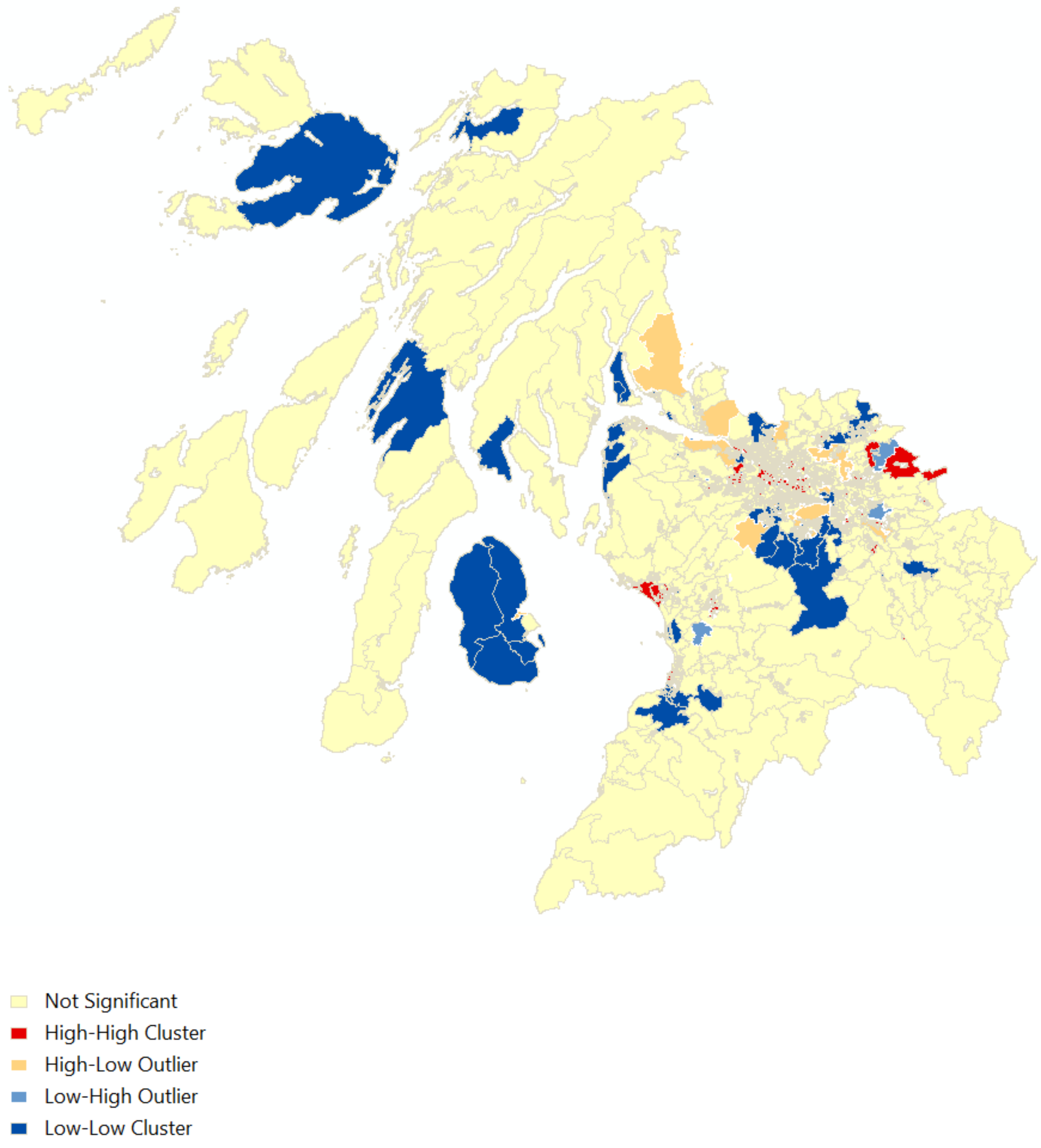


Figure 5.26: Map of clusters in Strathclyde for Local Moran's I at data zone level

Figure 5.26 shows the Local Moran's I output for the crime counts for 2011 data zone areas. The majority of the data zone areas are not significant, similar to the map at output area level for Local Moran's I. It appears from this map that there are a number of coldspots in the islands data zone areas (similar to both k-means and finite mixture model maps). The majority of the high crime cluster areas are in the Glasgow city centre area and Figure 5.27 shows the zoomed in version of this. The red areas are seen to be the High-High cluster which are output areas which are neighboured by other high crime output areas. The peach areas are medium-high crime areas as these are areas which are high crime areas surrounded by low crime areas.

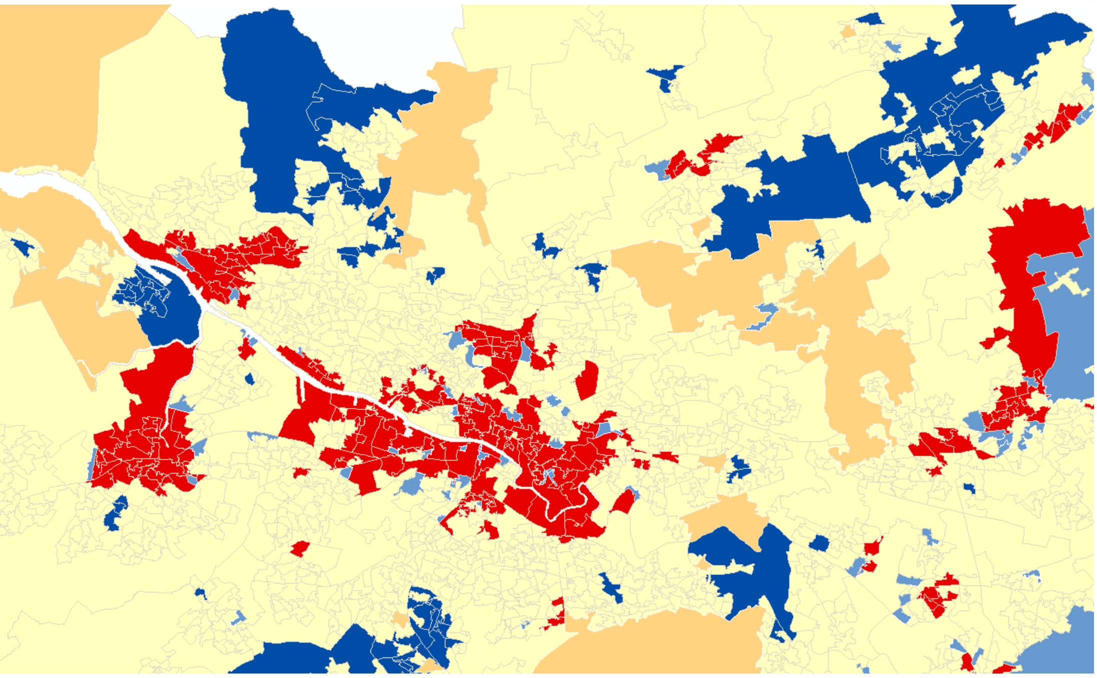


Figure 5.27: Map of clusters in Glasgow city centre for Local Moran's I at data zone level

Getis Ord Gi*

The hotspots analysis option in ArcMAP within ArcGIS produced the map shown in Figure 5.28. The default options were used with weight matrix of 1 if the areas border each other and 0 otherwise using the counts of crime at the data zone level.

Table 5.14 shows that most of the data zones lie in the medium crime cluster (76.75%). This is in contrast to k-means (21.16%) and finite mixture modelling (49.07%), but it is similar to Local Moran's I where (77.19%) of data zones lie within this medium cluster. The medium crime clusters for all methods at the data zone level have similar crimes on average from 30 to 39. The output areas range from low crime areas with only 11 crimes per output area on average to high crime areas which have 67 crimes per output area on average. Figure 5.28 can be used to see if these correspond to similar areas as Local Moran's I.

Table 5.14: Descriptive statistics for clusters for Getis Ord Gi* at data zone level

Cluster Group	1	2	3	4	5	6	7
Count of Data Zones	27	91	95	2274	118	151	207
Mean of Crime Counts in Data Zones	11.30	15.16	15.56	29.78	48.24	50.46	67.44
Cluster Type	Very Low Crime	Low crime	Low-Medium crime	Medium	Medium-High Crime	High Crime	Very High Crime
Percentage of Strathclyde Data Zones	0.91%	3.07%	3.21%	76.75%	3.98%	5.10%	6.99%

Figure 5.28 shows both hot and cold spots and most data zones are not significant. There is clearly a hotspot with 99% confidence level in Glasgow city centre (red) with a mixture of coldspots and non-significant output areas surrounding it shown in Figure 5.28. The hotspot with 99% confidence that was identified at the bottom right-hand side of the map at the output area level, is now showing as not significant at the data zone level. Both k-means (medium-high) and finite mixture models (medium) highlighted this area as a medium crime area surrounded by low crime areas. Local Moran's I identified this as a not significant area similar to Getis Ord Gi*. This again highlights the effect methods can have on the clusters identified. Through the non-spatially contiguous methods, this area was identified as a partly a low crime area surrounded by medium and high crime areas at the output area level while it was a medium crime area at data zone level. Through using the methods with spatial contiguity constraints, this area was identified as a high crime area or hotspot at the output area level while showing as not significant at the data zone areal level.

There are not-significant areas showing north of the river (seen as the white line in Figure 5.29) while south of the river there appears to be many high crime hotspots data zone areas. This is similar to the output area map produced at Gi* and also similar to the other data zone maps produced using k-means, finite mixture modelling and Local Moran's I.

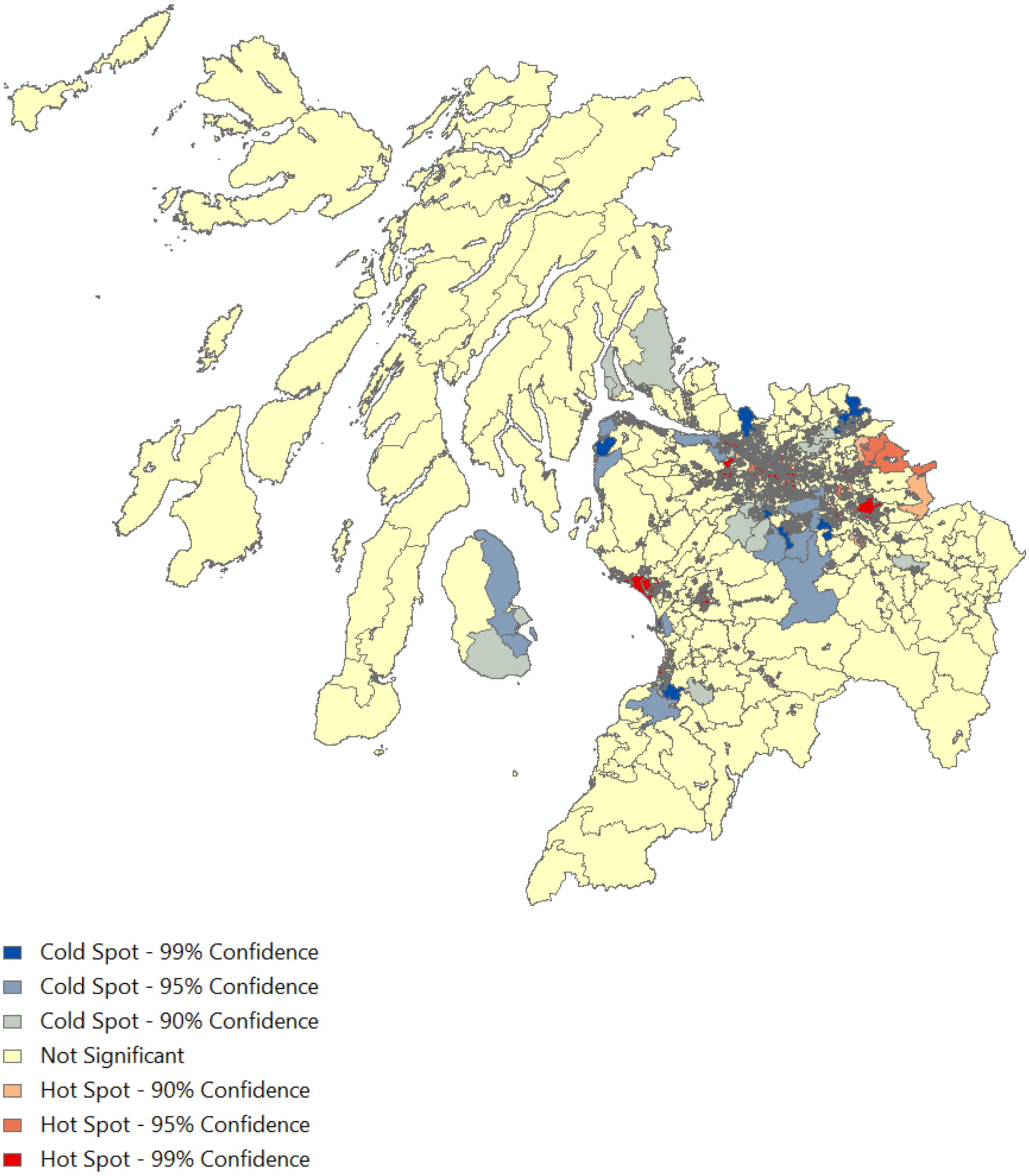

Cold Spot - 99% Confidence
Cold Spot - 95% Confidence
Cold Spot - 90% Confidence
Not Significant
Hot Spot - 90% Confidence
Hot Spot - 95% Confidence
Hot Spot - 99% Confidence

Figure 5.28: Map of clusters in Strathclyde for Getis Ord Gi* at data zone level

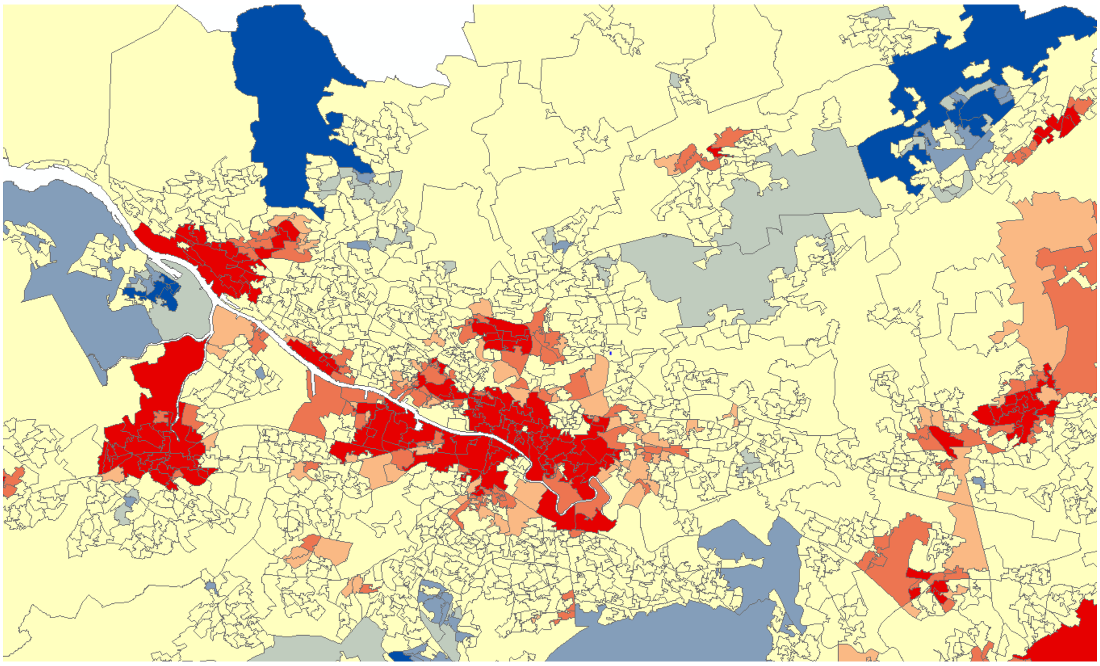
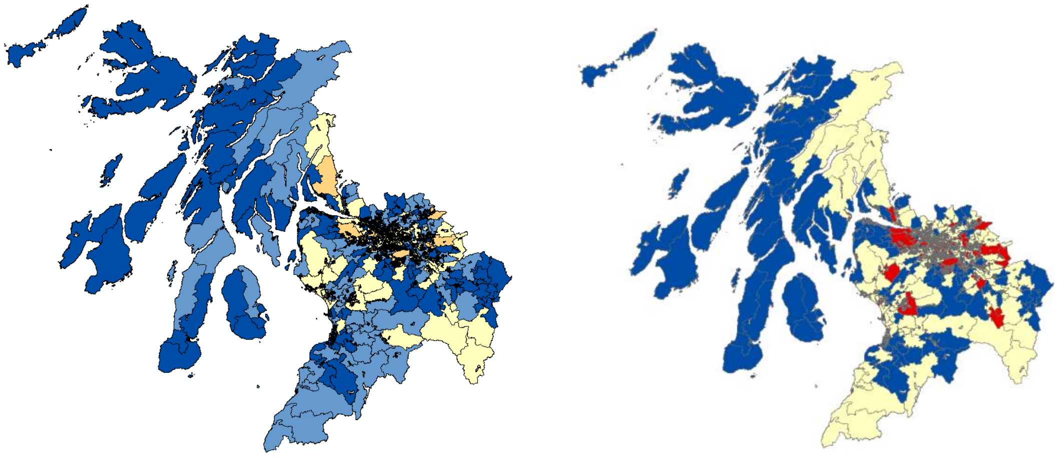


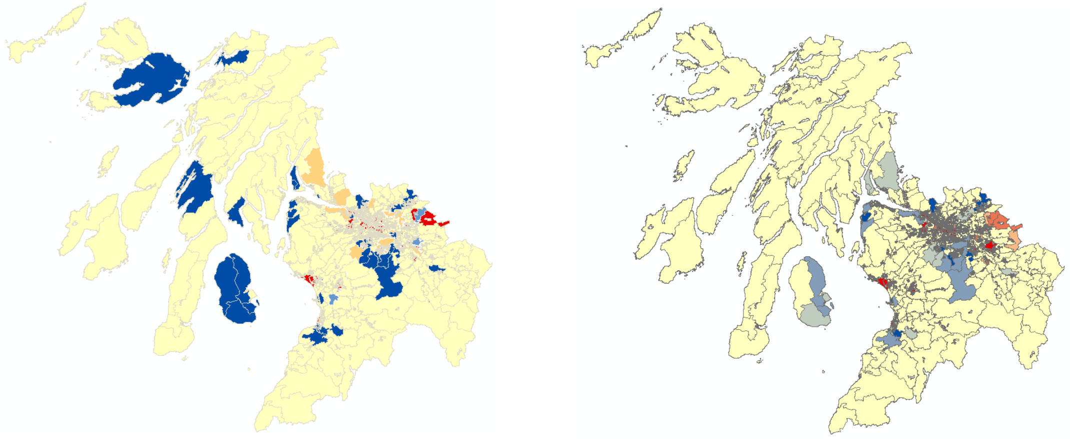
Figure 5.29: Map of clusters in Glasgow city centre for Getis Ord Gi* at data zone level

Strathclyde Comparison Maps (DZ)



(a) Data zone level k-means clusters in Strathclyde

(b) Data zone level finite mixture modelling clusters in Strathclyde



(c) Data zone level Local Moran's I clusters in Strathclyde

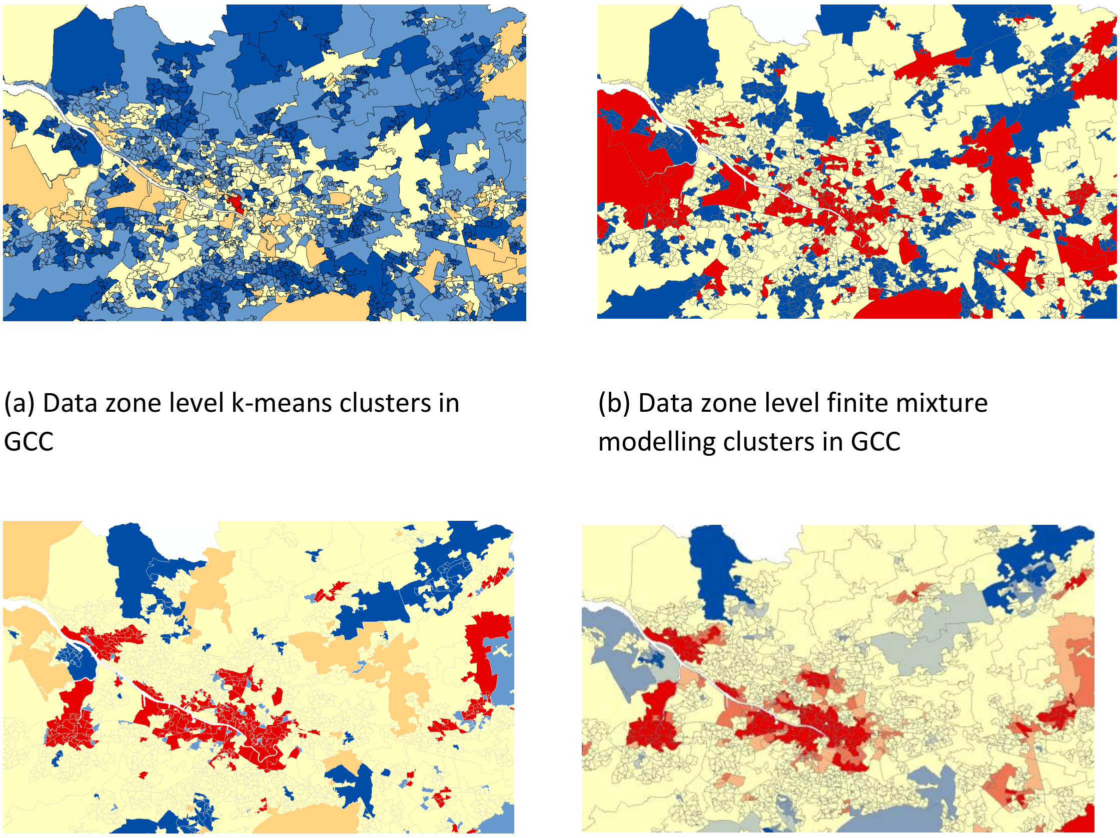
(d) Data zone level Getis Ord Gi* clusters in Strathclyde

Figure 5.30: Clusters for each method at data zone level in Strathclyde

Both Figures 5.30(a) and (b) appear to be very similar as they both have low crime then low-medium crime (dark and light blue areas) in the outskirts and islands of Strathclyde. There is a yellow area at the bottom right of Figure 5.30(a) which is highlighted by k-means output as being a medium crime area. This is seen in Figure 5.30(b) as also a medium crime area, although as k-means (a) has 5 clusters and finite mixture modelling (b) only has 3 clusters, this area is 'hidden' almost in (b) as it is the same crime cluster (medium) as its surrounding areas. This suggests that the interpretation of clusters can vary depending on the methods used. This area shows as non-significant for both Local Moran's I and Getis Ord Gi*, which are also classed as the medium crime clusters. But these are similar to the finite mixture modelling output as this area is 'hidden' as it is the same cluster as it's surrounding neighbours in each of the methods (bar k-means).

In Figure 5.30(c) there are a few dark blue areas (Low-Low Clusters) in the outskirts of Strathclyde similar to the other maps which also have these as being low crime areas (dark blue / blue) in (a) and (b) or not-significant (yellow) in (d). Most of the high crime hotspots lie within the centre of Figure 5.30(d). The majority of the surrounding areas are not significant suggesting these are neither high nor low crime areas and this is similar to the output of each of the other different cluster methods. This all highlights that there are some differences in the clusters identified using each method on the output area level data.

Glasgow City Centre (GCC) Comparison Maps (DZ)



(a) Data zone level k-means clusters in GCC

(b) Data zone level finite mixture modelling clusters in GCC

(c) Data zone level Local Moran's I clusters in GCC

(d) Data zone level Getis Ord Gi* clusters in GCC

Figure 5.31: Clusters for each method at data zone level in Glasgow City Centre area

Figures 5.31(a) there are medium-high (peach) mean crime clusters in Glasgow city centre. It can be seen at the centre of the map lies the high (red) crime cluster. However, there are also a great deal of low (blue) and medium (yellow) mean crime areas as well. This is similar to Figure 5.31(b) as there is a mix of low (blue), medium (yellow) and high (red) crime areas in this small area with finite mixture modelling results too. Figure 5.31(c) shows there are a great deal of non-significant areas still in GCC, but there are also some High-Low outliers (peach) on the outskirts of the centre. This is surprising as these tended to be seen as low or medium crime areas in the k-means and finite mixture modelling results. These are even seen as low crime areas (blue) in the Gi* results (d). In Figure 5.31(d) there are lots of red, high crime areas (hotspots with 99% significance) right in the middle of GCC. These seem to be in the same areas as the High-High clusters group (red) for Local Moran's I method (c) and the medium-high and high crime areas for k-means and finite mixture modelling suggesting these areas are seen as high crime in all methods outputs.

Adjusted Rand Index (DZ)

The adjusted Rand index was calculated for the clusters identified by each of these methods at the data zone areal level and the values can be seen in Table 5.15. Both the spatially non-contiguous and spatially contiguous (Local Moran's I and Getis Ord Gi*) show that they are the most similar to each other. At the data zone level, it is the spatially contiguous methods which produce the most similar clusterings as they have an ARI of 0.75 which is very close to 1. k-means and finite mixture models produce similar cluster groups with an ARI of 0.451. However, this is lower than the ARI for the output areas suggesting that there are more dissimilarities in these methods at the data zone level. Also, none of the other pairs of cluster methods appear to produce similar cluster groups as are all close to 0 (all are less than 0.1) suggesting that the clusters are due to random clusterings. This implies that the different methods produce very different results. Again, this could be due to the numbers of clusters being different in most of the methods.

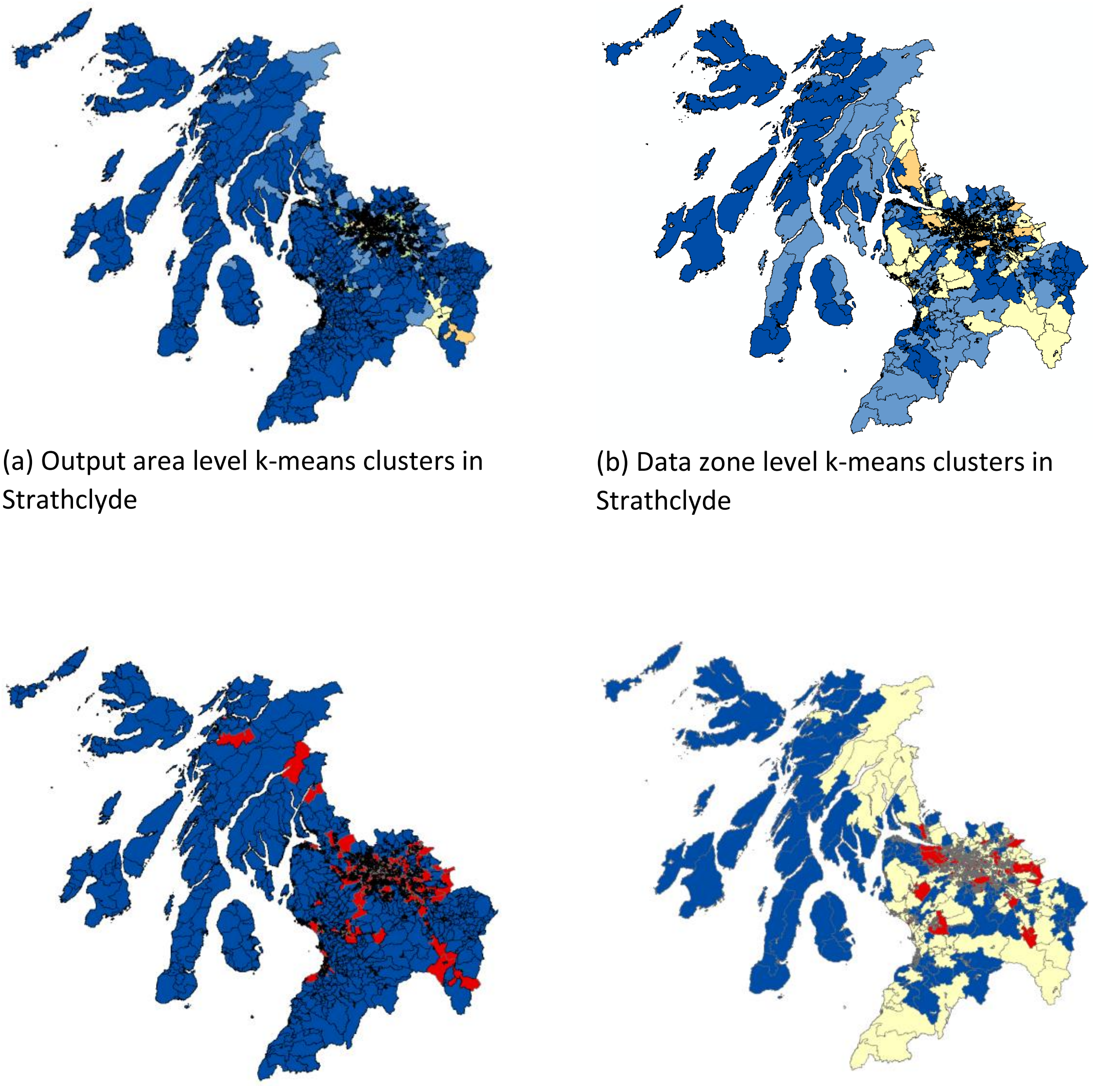
Table 5.15: Adjusted Rand index for cluster groupings at data zone level

Data Zones	k-means	finite mixture modelling	Local Moran's I	Getis Ord Gi*
k-means	1.000	0.451	0.053	0.077
finite mixture modelling	0.451	1.000	0.068	0.092
Local Moran's I	0.053	0.068	1.000	0.750
Getis Ord Gi*	0.077	0.092	0.750	1.000

Overall Comparison

Strathclyde Comparison Maps

Figure 5.32 shows the maps side-by-side for each of the clustering methods at output area level and data zone level for Strathclyde. This enables a comparison between the areal units to be made.

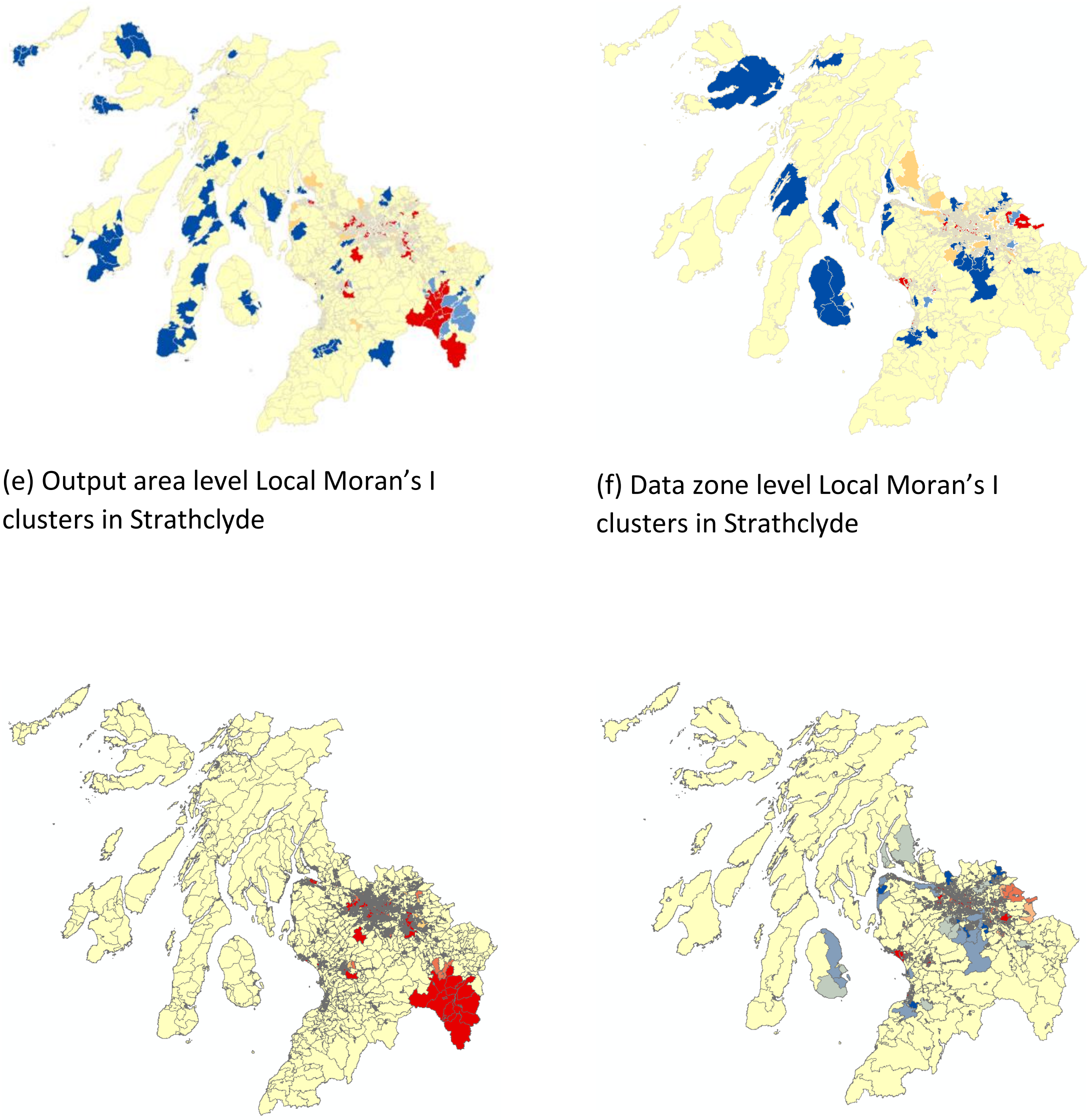


(a) Output area level k-means clusters in Strathclyde

(b) Data zone level k-means clusters in Strathclyde

(c) Output area level finite mixture modelling clusters in Strathclyde

(d) Data zone level finite mixture modelling clusters in Strathclyde



(e) Output area level Local Moran's I clusters in Strathclyde

(f) Data zone level Local Moran's I clusters in Strathclyde

(g) Output area level Getis Ord Gi* clusters in Strathclyde

(h) Data zone level Getis Ord Gi* clusters in Strathclyde

Figure 5.32: Comparison of cluster methods across output area and data zone levels for Strathclyde

In Figures 5.32(a) and (b), the clusterings produced by k-means at the output area level appear similar to the clusterings produced at the data zone level. The low (dark blue) and low-medium (light blue) areas around the outskirts of the map look to be similar. There does appear to be a larger spatial area identified as low-medium crime areas than low crime areas at the data zone level than the output area level and there are more medium (yellow) crime areas at the data zone level.

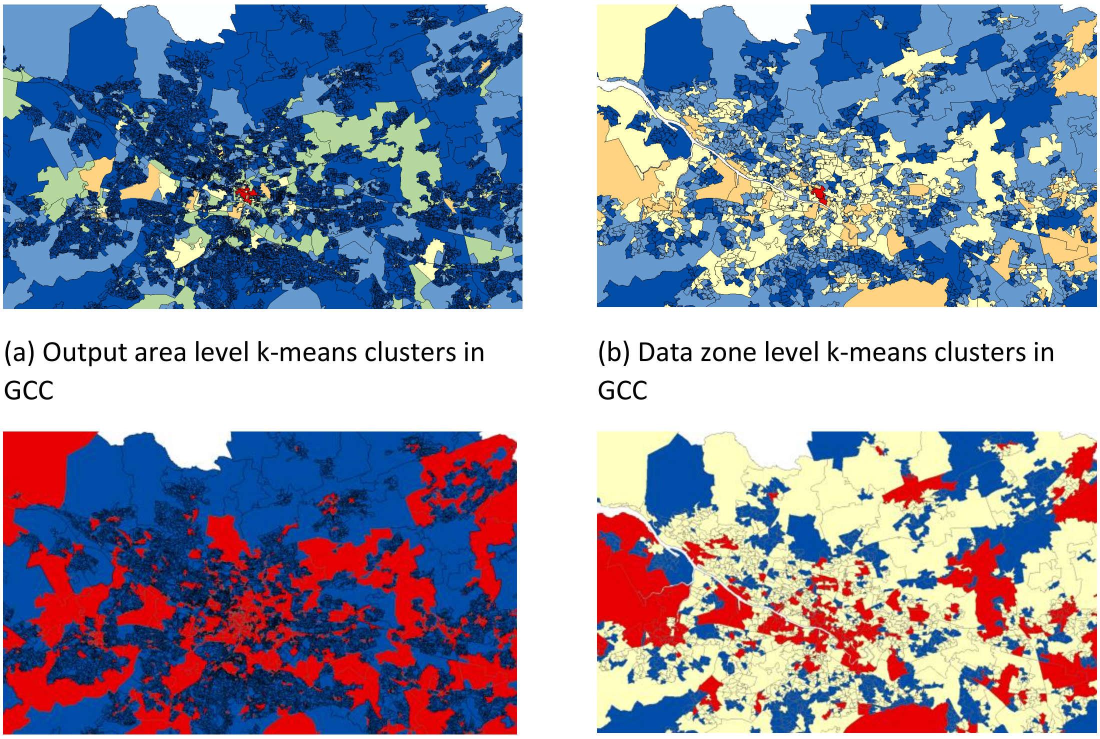
In Figures 5.32(c) and (d), the finite mixture modelling clustering outputs appear similar as the low (blue) crime cluster is on the outskirts of Strathclyde on both maps. However, there are a few high (red) areas seen at the output area level around the outskirts which appear to be 'hidden' at the data zone level.

Figures 5.32(e) and (f) which show the Local Moran's I clusterings, appear to be mostly similar as both the output areas and data zones appear to be in the mostly non-significant medium (yellow) crime cluster. There are also some similarities in the low (blue) crime clusterings identified at both the output area and data zone levels. However, there is a large medium-high (peach) crime area at the bottom of (e) and this is part of the not-significant cluster and is not highlighted separately at the data zone level (f).

Figures 5.32(g) and (h) show the clustering for Getis Ord Gi* at the output area and data zone levels and similarly to Local Moran's I, this appears to be quite similar as most belong to the not-significant (medium (yellow)) cluster. However, there appears to be a hotspot identified at the output area level in the bottom right of Strathclyde (seen as the red area) and this area appears as not0signficant at the data zone level which is similar to Local Moran's I.

Glasgow City Comparison

Figure 5.33 shows the maps side-by-side for each of the clustering methods at output area level and data zone level for Glasgow city centre. This enables a comparison between the areal units to be made.

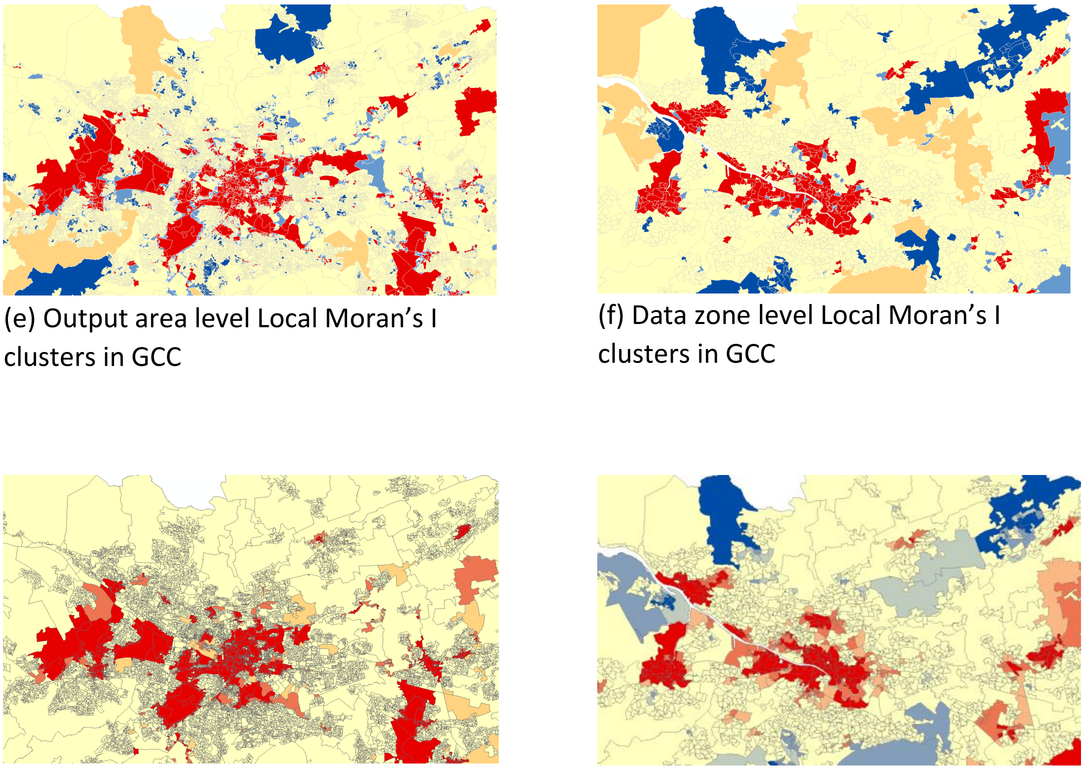


(a) Output area level k-means clusters in GCC

(b) Data zone level k-means clusters in GCC

(c) Output area level finite mixture modelling clusters in GCC

(d) Data zone level finite mixture modelling clusters in GCC



(e) Output area level Local Moran's I clusters in GCC

(f) Data zone level Local Moran's I clusters in GCC

(g) Output area level Getis Ord Gi* clusters in GCC

(h) Data zone level Getis Ord Gi*clusters in GCC

Figure 5.33: Comparison of cluster methods across output area and data zone levels for Glasgow City Centre

The k-means clusterings (seen in Figures 5.33(a) and (b)) appear to be quite different as there are more medium (yellow) crime areas at data zone level while these are low (dark blue) or low-medium (light blue) crime areas at the output area level. There are some similarities in the low-medium (light blue) areas and there does appear to be a similarity near the centre where a medium-high (yellow/peach) area can be seen at the output area level and a larger area can be seen covering the same output area at the data zone level. For finite mixture modelling, the clusters at output area (c) and data zone (d) level appeared to be quite different as the high (red) crime areas don't appear to lie in the same areas. (c) and (d) also show some similarities in the low (blue) crime areas between the areal levels.

For Local Moran's I clusterings in Figure 5.33(e) and (f), these appear relatively similar as the high (red) cluster appears to lie in the centre of both the output area (e) and data zone (f) maps. The outlying areas are mostly not-significant (yellow) in both maps but there does seem to be differences in where the low (blue) and medium-high (yellow/peach) clusters lie suggesting there are differences across output areas and data zone levels. This is similar to Figures 5.33(g) and (h) which show the Getis Ord Gi* clusterings as the areas seem to be mostly not-significant for both output areas (g) and data zones (h). Also, there are no coldspots (significantly low crime areas) at the output area level but there are some which can be seen at the data zone level suggesting some differences. However, the high (red) cluster area seems to be concentrated at the centre of the maps, although the output area map (g) shows there to a few at the right and at the bottom of the city centre area while the data zone map (h) shows there to be more high crime areas in the top left and centre left of the map.

Therefore, for both Strathclyde and Glasgow city centre, there appear to be some similarities amongst the output area and data zone clusterings for each method. However, there are a few hotspots seen at the output area level that are not seen at the data zone level suggesting the MAUP is prevalent as these hotspots are not identified at the data zone areal level.

Adjusted Rand Index

The underlying output areas for each datazone had the cluster groupings for the data zone mapped to it. This then meant that the cluster groupings created at both the output area and the data zone levels could be compared at the output area level. I.e. each underlying output area took the value of the corresponding data zone. There were 733 output areas (out of 19,886 in total) which did not lie directly within a data zone. These were left with no corresponding data zone cluster group. In order for the clusterings being compared to be the same length, these output areas were removed at both the output area level and the data zone level. This left 19,153 output areas where the clusterings could be compared across data zone and output area levels.

The adjusted Rand index was calculated for the clusters identified by all of these methods at the output area level for both data zone and output areal clusters and the values can be seen in Table 5.16. None of these methods appear similar as all values are close to 0. The method with the most similarity in clustering groups between output area and data zone areal levels is Gi* and Local Moran's I with ARI's close to 0.1. This suggests the influence of the MAUP as there appears to be no similarities between the clusters at output area and the clusters identified at the data zone area level. This shows the importance of identifying the areal units prior to running analysis as the areal units chosen have an impact on the clusters identified.

Table 5.16: Adjusted Rand Index for output areas and data zones (all at output area level)

All	k-means (data zones)	finite mixture modelling (data zones)	Local Moran's I (data zones)	Getis Ord Gi* (data zones)
k-means (output areas)	0.042	0.035	0.070	0.097
finite mixture modelling (output areas)	0.026	0.024	0.070	0.066
Local Moran's I (output areas)	0.041	0.066	0.138	0.157
Getis Ord Gi* (output areas)	0.034	0.038	0.122	0.145

It is of interest to look at a contingency table for two of the clusterings in to investigate why these are so low. Table 5.17 shows this for k-means clusterings compared between output areas and data zones and Table 5.18 shows this for Local Moran's I compared between the output areas and data zones.

Table 5.17: Contingency table comparing for clusterings produced at output area and data zone areal levels for k-means

k-means						
	DZ1	DZ2	DZ3	DZ4	DZ5	Total Rows
OA1	6448	7078	3578	429	9	17542
OA2	112	398	619	182	4	1315
OA3	7	49	114	52	5	227
OA4	0	5	25	20	2	52
OA5	0	1	4	7	2	14
OA6	0	0	1	0	2	3
Total Columns	6567	7531	4341	690	24	19153

Table 5.17 shows that only about a third of the output areas have the same clustering at the output areal level as the data zone areal level. Most output areas lie in cluster 1 at output area level but cluster 2 at data zone level (7,078). This highlights that there are not many similarities between clusterings at the two areal levels using k-means.

Table 5.18: Contingency table comparing for clusterings produced at output area and data zone areal levels for Local Moran's I

Local Moran's I						
	DZ1	DZ2	DZ3	DZ4	DZ5	Total Rows
OA1	370	19	858	11	2	1260
OA2	16	108	849	16	322	1311
OA3	866	460	12508	138	1590	15562
OA4	14	4	65	3	1	87
OA5	3	34	426	2	468	933
Total Columns	1269	625	14706	170	2383	19153

Table 5.18 shows that only about two thirds of the output areas have the same clustering at the output areal level as the data zone areal level (mainly in cluster 2). This suggests that there are some similarities between clusterings at the two areal levels using Local Moran's I. The ARI for comparing Local Moran's I clusterings is higher than the ARI for comparing k-means clusterings which is expected but the value is still very low (0.13). This suggests another way of comparing the clusterings at output area and data zone level could be used.

Therefore, while visually the maps appear to highlight similar areas as hotspots (high crime clusters), when the cluster groupings are compared, there are not many similarities in the cluster groupings identified. The maps may also look similar as when we look at the maps we might look at both high and medium-high areas and see them as being similar clustering but when these are compared using ARI, these are two separate clusters. The difference between the cluster groupings compared at output area and data zone level (Table 5.16) highlights the MAUP as if MAUP did not cause an issue, the cluster groupings would be most similar and perhaps closer to the value of 1. The results of ARI separately at both the output area and data zone areal levels (Table 5.10 and 5.15) show that different methods can have very different clustering results with some similarities seen in the clusterings identified by the two spatially contiguous methods and also the two spatially non-contiguous methods.

Chapter 6 - Conclusion

This thesis has looked at how clustering methods and areal units used can impact the clusters (and thus hotspots) produced. In Chapter 1 I looked at the ways in which crime and place have been linked and how crime mapping developed and the social theories which developed alongside this. This was then expanded in Chapter 2 to provide an overview of how cluster analysis can identify hotspots i.e. clusters with a high crime count/rate. This also looked at the Modifiable Areal Unit Problem in greater detail and the ways in which MAUP can be mitigated. After identifying that it is the areal units chosen which can cause an issue with MAUP, the crime data areal units were identified for this thesis in Chapter 3. The data for this thesis covers all recorded crimes and offences in Strathclyde and these are then aggregated to the output area and data zone areal levels. The demographics of Strathclyde show that it is an area with higher deprivation and unemployment levels than Scotland as a whole suggesting it will be an interesting study area. The methodology chapter (Chapter 4) looked at the four methods that were chosen for this study and how they are constructed, k-means, finite mixture modelling, Local Moran's I and Getis Ord Gi*.

These methods were then applied in Chapter 5 to the Strathclyde dataset aggregated to both the output area and data zone areal levels. While visually the maps appear to highlight similar areas as hotspots (high crime clusters), when the cluster groupings are compared, there are not many similarities identified. At the output area level, each of the methods (bar Local Moran's I), produced clusterings which had over 90% of the output areas within the low crime cluster showing that most areas had low crime levels. The low crime cluster identified by Local Moran's I had less than 7% of the output areas within it and most of the output areas (81%) lay within the non-significant cluster. This suggests that most output areas did not have high or low crime neighbours and can be thought of as the medium crime cluster. Similarly, at the data zone areal level, most of the data zones for the clusterings identified by k-means and finite mixture models belonged in the low crime category (about 40%). However, for both Local Moran's I and Getis Ord Gi* clusterings, the low crime category had less than 8% of data zones within it with most (over 76%) data zones belonging to the non-significant (medium) cluster. This suggests that when spatial contiguity constraints are used, the majority of the output areas and data zones are not high or low crime areas with neighbours with high or low crime levels, and most would likely be medium crime areas.

The maps may also look similar as when we look at the maps we might look at both high and medium-high areas and see them as being similar clustering but when these are compared using ARI, these are two separate clusters. The results of ARI separately at both the output area and data zone areal levels (Table 5.10 and 5.15) show that clustering results were very different depending on the method chosen. The difference between the cluster groupings compared at output area and data zone levels (Table 5.16) highlights the MAUP as if MAUP did not cause an issue, the cluster groupings would be more similar and with ARI values perhaps closer to 1.

The ARI values for the clusterings at output area showed that there is some similarity between k-means and finite mixture models with an ARI of 0.676 and also between Local Moran's I and Getis Ord Gi* with an ARI of 0.477. When comparing the clusterings produced by k-means to the Local Moran's I and Getis Ord GI* clusterings the ARI value dropped to 0.134 and 0.287 respectively. When finite mixture modelling is compared to the Local Moran's I and Getis Ord GI* clusterings, these are less than 0.31 suggesting that these are quite different clustering solutions. The results followed a similar pattern at the data zone areal level. This suggests that the methods with no assumption of spatially contiguity were similar to each other as were the two methods with spatial contiguity constraints.

Limitations

One of the main limitations of any analysis on police crime data (recorded crime) is the 'dark figure of crime'. This refers to the crimes which are not reported and therefore, do not form part of the recorded crime dataset. This can lead to any analysis of recorded crime data 'missing' other crime information. It is usually assumed that the 'dark figure of crime' remains relative to the actual recorded crime counts so that it can be assumed that the overall crime patterns are identified. This can potentially cause issues at smaller scales as the 'missing' or 'hidden' crimes could be in one particular area, and this could cause it to appear to have a lower crime count/rate than it actually does which can cause any clustering results to be misleading. Other crime sources could be used to help mitigate this as discussed in Chapter 3 (self-report and victimisation reports) but each of these still can have crimes not reported. It is police crime data that is available to me and it is incredibly useful as it allows detailed analysis of crimes which are recorded accurately and have location of crime information available which is very useful for this type of analysis.

The use of spatial autocorrelation methods (in this case Local Moran's I) can be questionable when the data are skewed as identified in a study by (Fortin & Dale, 2005). This can bias the results if there are outliers, in the case of the crime data there are a few output areas and data zones which have particularly high crime counts in comparison with the other output areas and data zones. However, Local Moran's I provides results which are more intuitive to interpret the correlations and therefore, it is used regularly even if the data is not normally distributed. It would be of interest to look at the differences identified in this study, in particular between the clusters identified by Local Moran's I and the non-spatially contiguous methods to see if there is any cause for this method to be excluded from future hotspot analysis of non-symmetrical crime data.

A better index of comparison than the adjusted Rand index could be used. The adjusted Rand index has very small values when comparing across the output area and data zone level. Part of the reason for this is ARI is ignoring the ordinal nature of the clusterings i.e. low to high. A change of assignment for an area from low to medium between different clusterings is not as different as a change from low to high. We could instead look at this using a measure of association between ordinal variables. One such would be Kendall's tau (Kendall, 1938).

Future Work

Crime Type and Years

The longitude and latitude co-ordinates for the centre of all recorded crimes across 1999-2013 were provided by Police Scotland (formerly Strathclyde Police). The hotspots for different crimes could be investigated, in particular the below crimes are usually identified as crimes of place:

- Crimes of Dishonesty – Other Theft
- Fire-raising, Vandalism – Vandalism
- Breach of the Peace
- Assaults (both minor and serious).

The definition of "Breach of the Peace" included "Threatening or abusive behaviour", "Offence of Stalking", "Offensive behaviour at football" (2012 onwards due to new legislation introduced) and "Threatening Communication" (2012 onwards due to new legislation introduced) due to linking to the Scottish Government recorded statistics since these are definitions that were widely used. "Serious assault" definition was expanded to also include "Attempted murder", "Murder" and "Culpable homicide" as these definitions were included in the Scottish Government recorded statistics as being combined.

It could also be of interest to look at the Modifiable Temporal Unit Problem where the impact on clusters identified based on what timescales are used is investigated. Different years could be looked at to see whether the same clusters were identified for each year. Also, the same year split into months could be looked at to see if there is a seasonal monthly change in the hotspots identified.

Hot and Cold Spots

The results show that using different methods to cluster can produce different clusterings which show different areas to be hotspots. This could lead to different police strategies being employed for example at areas which show as crime hotspots across multiple methods. These areas could be targeted more closely with specific interventions with further work looking at a socio-economic analysis of these significant crime hotspot areas. More analysis could be carried out to look into the make-up of areas which are identified as hotspots and 'coldspots'. This could prove insightful for the police officers to identify why in two spatially contiguous areas, one has no recorded crime across the whole of 2011 while the other is seen to be a medium-high crime area. Local Moran's I and Gi* analysis in ArcGIS can be extremely useful for identifying low crime areas, particularly in this case as the cold spots can be used to identify areas for further investigation as to why there are less crimes occurring in these areas. For high-high and low-low clusters it could be useful to then further investigate the main socio-economic difference between these areas such as urban/rural settings, land usage. Analysis can be carried out to identify what exists in these areas which leads to low crime being identified here such as police interventions, community safety partnerships existing or other reasons which could lead to low crime counts. This can lead to more initiatives being developed to target the high crime areas and police resources can be targeted efficiently.

Another option for further work is to carry out a Monte Carlo simulation study. This technique involves generating random values for crime count/rate for each areal unit based on the range of (real data) estimates. Cluster analysis (e.g. k-means) is then carried out on the simulated data and this is repeated hundreds of times. Each time this is carried out, different simulated values for the crime count/rate are used. Examining the performance of the clustering in the context of MAUP in a situation where the truth is known (simulated from) would allow for further examination of the problem.

It could be of interest to identify if the main social theories relating to crime can be evidenced by looking at the hotspots or coldspots identified. Areas which are identified as high crime areas, may be lacking suitable guardians or could be areas which look unkempt but have plenty of targets. This could then provide further evidence towards the social theories discussed in Chapter 1.

Therefore, this thesis has laid the foundations for further study to look at crime hotspots and the modifiable areal unit. This could involve using more specific crime types and could look into linking to socio-economic neighbourhood factors which could influence other types of crime.

Appendix A: R Coding

k-means Output Area

(repeated the same using datazone data)

```
##Applying k-means to rates for all crimes for 2011 Output areas
AllCrimes<-  read.csv("R:/Becca/12Dec15/qryAllCrime_OA_2011_Rate.csv", header=TRUE,
sep=",")
attach(AllCrimes)
names(AllCrimes)
AllCrimes
Crimes<- cbind(AllCrimes)
Crimes
n<- length(Crimes[,1])
n
newdata <- cbind(Crimes)
newdata
summary(is.na(newdata))
names(newdata)
OAs<- newdata[,1]                   #outputareas
E<- data.frame(newdata[,7])      #crime rate

#calculate the within group sum of squares for k=1 to 10

wss1<- (n-1)*sum(apply(E,2,var))
wss<- numeric(0)
for (i in 2:10){
          W<- sum(kmeans(E, i, nstart=50)$withinss)
          wss<- c(wss,W)
}
wss<- c(wss1,wss)
wss
plot(1:10, wss, type="l", xlab="Number of Groups", ylab="Within groups sum of squares",
lwd=2)

### Use larger value for possible number of groups.  Try 100.
wss1<- (n-1)*sum(apply(E,2,var))
wss1
is.na(wss1)<-0
wss1
wss<- numeric(0)
for (i in 2:100){
          W<- sum(kmeans(E, i)$withinss)
          wss<- c(wss,W)
}
wss<- c(wss1,wss)
wss
plot(1:100, wss, type="l", xlab="Number of Groups", ylab="Within groups sum of squares",
lwd=2)
```

```
### 1 to 100 plot very small due to large wss at start so try smaller parts of the plot

wss_sub<- wss[1:20]
wss_sub
plot(1:20, wss_sub, type="l", xlab="Number of Groups", ylab="Within groups sum of
squares", lwd=2)

wss_sub2<- wss[10:20]
wss_sub2
plot(10:20, wss_sub2, type="l", xlab="Number of Groups", ylab="Within groups sum of
squares", lwd=2)

wss_sub3<- wss[10:60]
wss_sub3
plot(10:60, wss_sub3, type="l", xlab="Number of Groups", ylab="Within groups sum of
squares", lwd=2)

##6 groups
acrime.kmean<- kmeans(E, 6)
```

Finite Mixture Modelling

(repeated the same using datazone data)

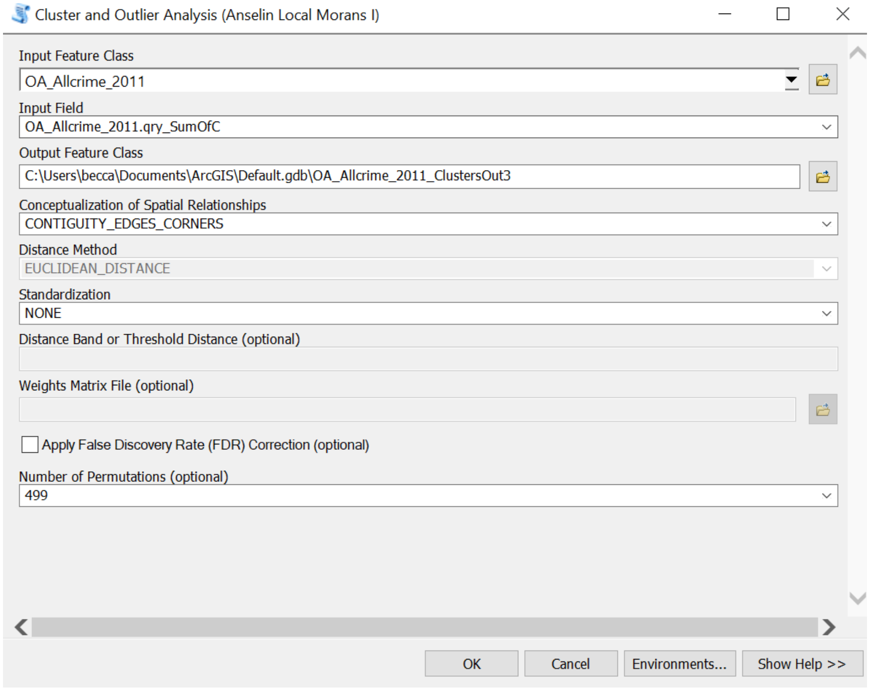
```
##applying to all crime 2011 output areas

AllC_2011<- read.csv("R:/Becca/12Dec15/qryAllCrime_OA_2011.csv", header=TRUE,
sep=",")
plot(AllC_2011)
AllC_2011
summary(is.na(AllC_2011))
newdata <- (AllC_2011)
newdata
summary(is.na(newdata))
library(flexmix)
names(newdata)
data<-newdata[,2]         #this is the crime counts
summary(data)
data<-as.data.frame(data)
res<-flexmix(data~1,data=data,k=20, model=FLXMRglm(family="poisson"))

table(clusters(res))
summary(res)
exp(parameters(res))
stepFlexmix(data~1,k=c(1:10),data=data, model=FLXMRglm(family="poisson"))
```

Appendix B: ArcGIS

Local Moran's I



Cluster and Outlier Analysis (Anselin Local Morans I)

Input Feature Class
OA_Allcrime_2011

Input Field
OA_Allcrime_2011.qry_SumOfC

Output Feature Class
C:\Users\becca\Documents\ArcGIS\Default.gdb\OA_Allcrime_2011_ClustersOut3

Conceptualization of Spatial Relationships
CONTIGUITY_EDGES_CORNERS

Distance Method
EUCLIDEAN_DISTANCE

Standardization
NONE

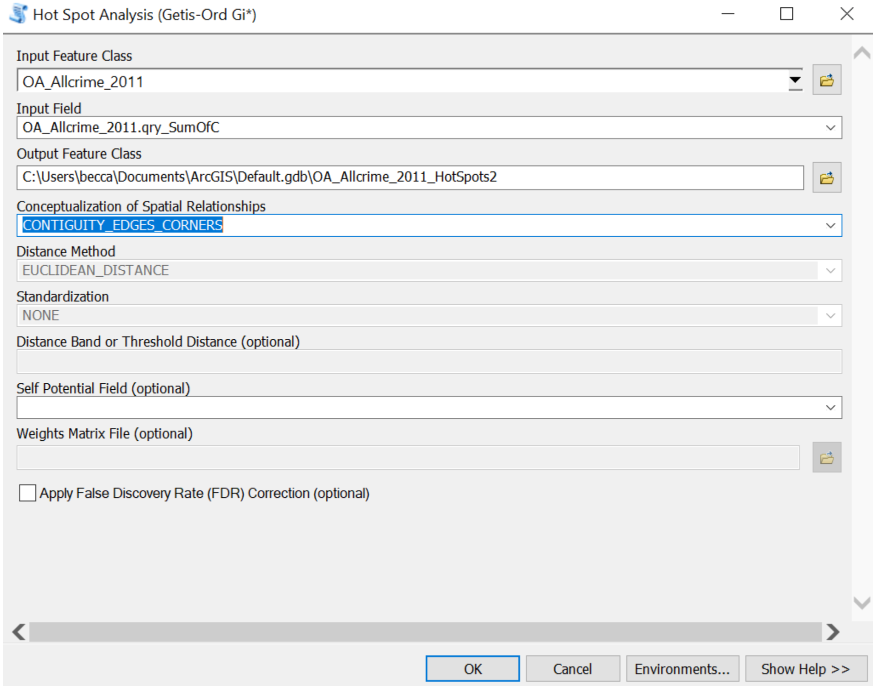
Distance Band or Threshold Distance (optional)

Weights Matrix File (optional)

□ Apply False Discovery Rate (FDR) Correction (optional)

Number of Permutations (optional)
499

Getis Ord Gi*



Hot Spot Analysis (Getis-Ord Gi*)

Input Feature Class
OA_Allcrime_2011

Input Field
OA_Allcrime_2011.qry_SumOfC

Output Feature Class
C:\Users\becca\Documents\ArcGIS\Default.gdb\OA_Allcrime_2011_HotSpots2

Conceptualization of Spatial Relationships
CONTIGUITY_EDGES_CORNERS

Distance Method
EUCLIDEAN_DISTANCE

Standardization
NONE

Distance Band or Threshold Distance (optional)

Self Potential Field (optional)

Weights Matrix File (optional)

□ Apply False Discovery Rate (FDR) Correction (optional)

Bibliography

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest, Hungary.

Anderson, C., Lee, D., & Dean, N. (2014). Identifying Clusters in Bayesian Disease Mapping. *Biostatistics*, *15*(3), 457–469. Applications. Retrieved from http://arxiv.org/abs/1311.0660

Andresen, M. A. (2015). Indentifying changes in spatial patterns from police interventions: the importance of multiple methods of analysis. *Police Practice and Research*, *16*(2), 148–160.

Andresen, M. A., Brantingham, P. J., & Kinney, J. B. (2010). *Classics in Environmental Criminology*. (M. A. Andresen, P. J. Brantingham, & J. B. Kinney, Eds.). Canada: Simon Fraser University Publications.

Andresen, M. A., Curman, A. S., & Linning, S. J. (2017). The Trajectories of Crime at Places: Understanding the Patterns of Disaggregated Crime Types. *Journal of Quantitative Criminology*, *33*(3). http://doi.org/10.1007/s10940-016-9301-1

Andresen, M. A., & Malleson, N. (2013). Spatial Heterogeneity in Crime Analysis. In M. Leitner (Ed.), *Crime Modeling and Mapping Using Geospatial Technologies*. London: Springer.

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

Audit Scotland. (2000). Safe and sound: A study of community partnerships in Scotland, (May).

Baker, S. P., Whitfield, R. A., & O' Neill, B. (1987). Geographic variations in mortality from motor vehicle crashes. *New England Journal of Medicine*, *316*, 1384–1387.

Barthe, E., & Stitt, B. G. (2009). Impact of casinos on criminogenic patterns. *Police Practice and Research*, *10*(3), 255–269. http://doi.org/10.1080/15614260802381067

Bates, E. (2014). *Vandalism : A Crime of Place? PhD Thesis*. University of Edinburgh.

Bernasco, W., & Elffers, H. (2010). Statistical Analysis of Spatial Crime Data. In A. R. Piquero & D. Weisburd (Eds.), *Handbook of Quantitative Criminology*. London: Springer.

Bichler, G., & Balchak, S. (2007). Address matching bias: ignorance is not bliss. *Policing: An International Journal of Police Strategies and Management*, *30*(1), 32–60.

Biderman, A. D., & Reiss, A. J. (1967). On exploring the "dark" figure of crime. *Annals of the American Academy of Political and Social Sciences*, *374*(1), 1–15.

Bleetman, A., Perry, C. H., Crawford, R., & Swann, I. J. (1997). Effect of Strathclyde police initiative "Operation Blade" on accident and emergency attendances due to assault. *Emergency Medicine Journal*, *14*(3), 153–156. http://doi.org/10.1136/emj.14.3.153

Bottoms, A. (2012). Developing socio-spatial criminology. In M. Maguire, R. Morgan, & R. Reiner (Eds.), *The Oxford Handbook of Criminology* (5th ed.). Oxford: Oxford University Press.

Bowers, K. J., & Johnson, S. D. (2004). Who commits near repeats? A test of the boost explanation. *Western Criminology Review*, *5*(3), 12–24.

Brantingham, P. L., & Brantingham, P. J. (1981). Notes on the geometry of crime. In P. L. Brantingham & P. J. Brantingham (Eds.), *Environmental Criminology*. Beverly Hills, CAL: Sage Publications.

Brantingham, P. L., Brantingham, P. J., Vajihollahi, M., & Wushke, K. (2009). Crime Analysis at Multiple Scales of Aggregation: A Topological Approach. In D. Weisburd, W. Bernasco, & G. J. N. Bruinsma (Eds.), *Putting Crime in its Place*. New York, NY: Springer New York.

Buerger, M. E., Conn, E. G., & Petrosino, A. J. (1995). Defining the "Hot Spots of Crime": Operationalizing Theoretical Concepts for Field Research. *Crime and Place*, *4*(2), 237–

257. Retrieved from http://www.popcenter.org/Library/CrimePrevention/Volume_04/11-Buerger.pdf

Burgess, E. W. (1916). Juvenile delinquency in a small city. *Journal of the American Institute of Criminal Law and Criminology*, *6*(5), 724–728.

Burns, R. (2008). Chapter 23 Cluster Analysis. Retrieved March 14, 2015, from http://www.uk.sagepub.com/burns/website material/Chapter 23 - Cluster Analysis.pdf

Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: a routine activities approach. *American Sociological Review*, *44*(4), 588–608.

Coleman, C., & Moynihan, J. (1999). *Understanding Crime Data: haunted by the dark figure of crime*. Buckingham: Open University Press.

Courtright, K. E., & Mutchnick, R. J. (2002). Cartographic School of Criminology. In D. Levinson (Ed.), *Encyclopedia of Crime and Punishment* (pp. 176–179). Thousand Oaks, CA: Sage.

Dark, S. J., & Bram, D. (2007). The modifiable areal unit problem (MAUP) in physical geography. *Progress in Physical Geography*, *31*(5), 471–479.

Davis, M. (2012). *The Modifiable Areas Unit Problem (MAUP) via Cluster Analysis Methodologies: a Look at Scale, Zoning, and instances of foreclosure in Los Angeles County*. University of South California.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete-data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, *B39*, 1–38.

Driver, H. E., & Kroeber, A. L. (1932). Quantitative expression of cultural relationships. *University of California Publications in American Archaeology and Ethnology*, *31*, 211–256.

Eck, J. E., & Weisburd, D. (1995). Crime places in crime theory. In J. E. Eck & D. Weisburd (Eds.), *Crime and Place. Crime Prevention Studies*. Monsey, NY and Washington, DC: Criminal Justice Press and The Police Executive Forum.

Everitt, B., & Hothorn, T. (2011). *An Introduction to Applied Multivariate Analysis with R*. (R. Gentleman, K. Hornik, & G. Parmigiani, Eds.). London: Springer.

Felson, M. (1987). Routine activities and crime prevention in the developing metropolis. *Criminology*, *25*(4), 911–932.

Flowerdew, R., Graham, E., & Feng, Z. (2004). *The Production of an Updated Set of Data Zones to Incorporate 2001 Census Geography and Data*. St Andrews.

Fortin, M-J., & Dale, M. R. T. (2005). *Spatial Analysis: a guide for ecologists*. Cambridge University Press. Retrieved from https://www-cambridge-org.ezproxy.lib.gla.ac.uk/core/services/aop-cambridge-core/content/view/1E6B08E5D90FB67137DDA12D45881FC1/9780511542039c3_p111-173_CBO.pdf/spatial_analysis_of_sample_data.pdf

Fotheringham, A. S. (1989). Scale-independent spatial analysis. In M. F. Goodchild & S. Gobal (Eds.), *Accuracy of Spatial Databases*. London: Taylor and Francis.

Fotheringham, A. S., & Wong, D. W. (1991). The modifiable areal unit problem in mulivariate statistical analysis. *Environment and Planning A*, *27*(7), 1025–1044.

Getis, A., & Ord, J. K. (1992). The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis*, *24*, 189–206.

Glyde, J. (1856). Localities of crime in Suffolk. *Journal of the Statistical Society London*, *19*(2), 102–106.

Gordon, A. D. (1996). How many clusters? An investigation of five procedures for detecting nested cluster structure. In P. Forer, A. Yeh, & J. He (Eds.), *Proceedings of 9th International Symposium on Spatial Data Handling*. Beijing International Geographic Union.

Grubesic, T. H. (2006). On the Application of Fuzzy Clustering for Crime Hot Spot Detection. *Journal of Quantitative Criminology*, *22*(1), 77–105.

Grubesic, T. H., & Murray, A. T. (2001). Detecting Hot Spots Using Cluster Analysis and GIS. In *Fifth Annual International Crime Mapping Research Conference*. Dallas, Texas.

Gruen, B., & Leisch, F. (2007). Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics & Data Analysis*, *51*(11), 5247–5252.

Gruen, B., & Leisch, F. (2008). Flexmix Version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, *28*(4), 1–35.

Guerry, A.-M. (1831). *Essai Sur la Statistique Morale de la France*. Paris: Chez Crochard.

Guerry, A.-M. (1833). *Essai sur la statistique morale de la France: precede d'un rapport a l'Academie de sciences*. Paris: Chez Crochard.

Harries, K. D. (1999). *Mapping crime: principles and practice*. Washington, DC: US Department of Justice.

Harris, C., Jones, P., Hillier, D., & Turner, D. (1998). CCTV surveillance systems in town and city centre management. *Property Management*, *16*(3), 160–165.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means Clustering Algorithm. *Journal of the Royal Statistical Society*, *28*(1).

Hipp, J. R. (2007). Block, tract and levels of aggregation: neighbourhood structure and crime and disorder as a case in point. *American Sociological Review*, *72*(5), 659–680.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 193–218.

Hunter, A. (1978). Symbols of Incivility. In *Annual Meeting of the American Society of Criminology*. Dallas, Texas.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 651–666.

Jean, S. (2007). *Pockets of Crime: Broken Windows, Collective Efficacy and the Criminal Point of View*. Chicago: University of Chicago Press.

Johnson, S. D., & Bowers, K. J. (2004a). The burglary as clue to the future: the beginnings of prospective hot-spotting. *European Journal of Criminology*, *1*(2), 237–255.

Johnson, S. D., & Bowers, K. J. (2004b). The stability of space-time clusters. *British Journal of Criminology*, *44*(1), 55–65.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, C. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 881–892.

Kaufman, L., & Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley.

Kendall, M. (1938). A New Measure of Rank Correlation. *Biometrika*, *30*(1–2), 81–89.

Kumar, M. V., & Chandrasekar, C. (2010). Crime Hotspot detection using spatial Clustering Clustering : A literature review of related work. *International Journal of Advanced Research in Computer Science*, *1*(3), 415–417.

Lawson, A. B. (2010). Hotspot detection and clustering: ways and means. *Environmental and Ecological Statistics*, *17*(2), 231–245. http://doi.org/10.1007/s10651-010-0142-z

Leisch, F. (2004). Flexmix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *Journal of Statistical Software*, *11*(8), 1–18.

Lembo Jr, A. J., Lew, M. Y., Laba, M., & Baveye, P. (2005). Use of spatial SQL to assess the practical significance of the Modifiable Areal Unit Problem. *Computers & Geosciences*, *32*(2), 270–274.

Lloyd, S. P. (1957). Least square quantization in PCM. *Bell Telephone Laboratories Paper*.

MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In *5th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281–297).

Maltz, M. D., Gordon, A. C., & Friedman, W. (1991). *Mapping crime in its community* *setting: event geography analysis*. New York: Springer.

Manley, D., Flowerdew, R., & Steel, D. (2005). Scales, levels and processes: Studying spatial patterns of British census variables. *Computers, Environment and Urban Systems*, *30*(2), 143–160.

Mason, T. J., McKay, F. W., Hoover, R., Blot, W. J., & Farumeni, J. F. J. (1985). Atlas of cancer mortality for U.S. Counties: 1950-69. *DHEW Publication (NJH). Bethesda, Md: National Cancer Institute*, 75–780.

MCE Insurance. (n.d.). Motorcycle Accidents: An Interactive Guide to Motorcycle Accidents in the UK. Retrieved December 3, 2014, from http://www.mceinsurance.com/resources/uk-motorcycle-accident-hotspots/

McLachlan, G., & Peel, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics.

Morissette, L., & Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology*, *9*(1), 15–24. http://doi.org/10.20982/tqmp.09.1.p015

National Records Scotland. (2013a). Census Results 2011 Release 1C. Retrieved September 25, 2018, from http://www.scotlandscensus.gov.uk/documents/censusresults/release1c/rel1ctablesA1toA5.pdf

National Records Scotland. (2013b). Census Results 2011 Release 2a. Retrieved September 25, 2018, from http://www.scotlandscensus.gov.uk/documents/censusresults/release2a/rel2asbfigure19.pdf

Office for National Statistics. (n.d.). Output Area (OA). Retrieved September 28, 2015, from http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/census/output-area--oas-/index.html

Openshaw, S. (1984). The modifiable areal unit problem. *Concepts and Techniques in Modern Geography (CATMOG)*, *38*.

Openshaw, S., Cross, A., Charlton, M., Brunsdon, C., & Lillie, J. (1990). Lessons learnt from a Post Mortem of a failed GIS. In *2nd National Conference and Exhibition of the AGI*. Brighton.

Ouimet, M. (2000). Aggregation bias in ecological research: how social disorganization and criminal opportunities shape the spatial distribution of juvenile delinquency in Montreal. *Canadian Journal of Criminology*, *42*(2), 135–156.

Pierce, G. L., Spaar, S. A., & Briggs, L. R. (1984). *The character of police work: Implications for the delivery of services*. Boston: Center for Applied Social Research, Northestern University.

Quetelet, L. A. J. (1842). *A treatise on man and the development of his faculties*. Edinburgh: W. and R. Chambers.

R Core Team. (2016). R: A language and environment for statistical computing. Retrieved from https://www.r-project.org/

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*, 846–850.

Ratcliffe, J. (2010). Crime Mapping: Spatial and Temporal Challenges. In A. R. Piquero & D. Weisburd (Eds.), *Handbook of Quantitative Criminology*. London: Springer.

Ratcliffe, J. H. (2001). On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units. *International Journal of Geographic Information Science*, *15*(5), 473–485.

Ratcliffe, J. H. (2004a). Geocoding crime and a first estimate of a minimum acceptable hit rate. *International Journal of Geographic Information Science*, *18*(1), 61–72.

Ratcliffe, J. H. (2004b). The hotspot matrix: a framework for the spatio-temporal targeting of crime reduction. *Police Practice and Research*, *5*(1), 5–23.

Ratcliffe, J. H., & McCullagh, M. J. (1998). Identifying repeat victimisation with GIS. *British Journal of Criminology*, *38*(4), 651–662.

Ratcliffe, J. H., & Rengert, G. F. (2008). Near repeat patterns in Philadelphia shootings. *Security Journal*, *21*(1–2), 58–76.

Reynald, D. M., & Elffers, H. (2009). The Future of Newman's Defensible Space Theory. *European Journal of Criminology*, *6*(1), 25–46. http://doi.org/10.1177/1477370808098103

Robinson, W. S. (1950). Ecological correlations and the behaviour of individuals. *American Sociological Review*, *15*, 351–357.

Schafer, S. (1969). *Theories in Criminology: Past and Present Philosophies of the Crime Problem*. New York: Random House.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

Scottish Index of Multiple Deprivation. (2015). Area Profile Report for Local Authority Glasgow City. Retrieved July 5, 2015, from http://www.sns.gov.uk/Reports/Report.aspx?ReportId=2&AreaTypeId=LA:Local Authority&AreaId=S12000046

Scottish Neighbourhood Statistics. (2004). *Data Zones Background Information*.

Scottish Neighbourhood Statistics. (2007). Geography Data Guide. Retrieved September 24, 2015, from http://www.sns.gov.uk/Guide/GeographyGuide.aspx?GeographyType=PU#Meta

Shaw, C., & McKay, H. (1942). *Juvenile Delinquency and Urban Areas*. Chicago: University of Chicago Press.

Shaw, C. R., & McKay, H. D. (1931). *Social Factors in Juvenile Delinquency*. Washington, DC: U.S. Government Printing Office.

Sherman, L. W., Gartin, P. R., & Buerger, M. E. (1989). Hot Spots of Predatory Crime: Routine Activities and the Criminology of Place. *Criminology*, *27*(1).

Sherman, L. W., & Weisburd, D. L. (1995). General Deterrent Effects of Police Patrol in Crime "Hot Spots": A Randomized Study. *Justice Quarterly*, *12*(4), 625–648.

Shjarback, J. (2014). Defensible Space Theory. *The Encyclopedia of Theoretical Criminology*. http://doi.org/10.1002/9781118517390/wbetc084

Skeggs, B., Moran, L., Tyrer, P., & Binnie, J. (2004). Queer as Folk : producing the real of urban space. *Urban Studies*, *41*(9), 1839–1856. http://doi.org/10.1080/0042098042000243183

Tanner, W. (2014). Preventative criminal justice in Glasgow, Scotland: Violence Reduction Unit, Scotland. Retrieved May 29, 2015, from http://www.reform.uk/wp-content/uploads/2014/11/Preventative_criminal_justice_in_Glasgow_Scotland.pdf

Taylor, R. B., & Harrell, A. V. (1996). *Physical environment and crime*. Washington, DC: National Institute of Justice, US Department of Justice.

The Scottish Government. (2006). Scottish Neighbourhood Statistics Guide. Retrieved August 15, 2013, from http://www.scotland.gov.uk/Publications/2005/02/20697/52626

The Scottish Government. (2009a). Statistical Bulletin Crime and Justice Series: Recorded Crime in Scotland, 2008-09. Retrieved November 25, 2014, from http://www.scotland.gov.uk/Resource/Doc/286378/0087196.pdf

The Scottish Government. (2009b). Tackling Knife Crime. Retrieved March 25, 2013, from http://www.scotland.gov.uk/News/Releases/2009/03/04092419

The Scottish Government. (2010). Knife Crime Initiative Rolled Out. Retrieved March 25, 2013, from http://www.scotland.gov.uk/News/Releases/2010/07/05095706

The Scottish Government. (2011). Scotland's Census 2011: Census Geographies Guide, *2013*(November 2012). Retrieved from https://www.scotlandscensus.gov.uk/documents/supporting_information/2011_Census_Geographies.pdf

The Scottish Government. (2018). Scotland's Census: Bulletin Figures and Tables. Retrieved October 12, 2018, from https://www.scotlandscensus.gov.uk/bulletin-figures-and-tables

Tobler, W. (1979). Cellular geography. In S. Gale & G. Olsson (Eds.), *Philosophy in Geography*. Dordrecht: Reidel.

Townsley, M., Homel, R., & Chaseling, J. (2003). Infectious burglaries: a test of the near repeat hypothesis. *British Journal of Criminology*, *43*(3).

Townsley, M., Homel, R., & Ratcliffe, J. H. (2008). Space time dyamics of insurgent activity in Iraq. *Security Journal*, *21*(3), 139–149.

Tyron, R. A. (1939). *Cluster analysis: Correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality*. Ann Arbor, Michigan: Edwards Brothers.

Vold, G. B., & Bernard, T. J. (1986). *Theoretical Criminology* (3rd ed.). New York: Oxford University Press.

Walklate, S. (1989). *Victimology: The Victim and the Criminal Justice Process*. London: Unwin Hymen.

Weisburd, D. (2008). Place-based policing. *Ideas in American Policing*.

Weisburd, D., Bernasco, W., & Bruinsma, G. J. N. (2009). *Putting Crime in its Place Units of Geographic Criminology*. (D. Weisburd, W. Bernasco, & G. J. N. Bruinsma, Eds.). New York: Springer.

Weisburd, D., Bushway, S., Lum, C., & Yang, S.-M. (2004). Trajectories of crime at places: a longitudinal study of street segments in the city of Seattle. *Criminology*, *42*(2), 283–321.

Weisburd, D., & McEwan, T. (1997). *Crime mapping and crime prevention, Vol 8*. New York: Criminal Justice Press.

Weisburd, D., & Telep, C. W. (2014). Hot Spots Policing: What We Know and What We Need to Know. *Journal of Contemporary Criminal Justice*, *30*(2). http://doi.org/10.1177/1043986214525083

Which. (2013). The UK's Worst Accident Hotspots. Retrieved December 3, 2014, from http://www.which.co.uk/news/2013/03/the-uks-worst-accident-hotspots-315038/

Wilson, J. W., & Kelling, G. (1982). Broken Windows. *The Atlantic Monthly*, *249*(3), 29–38.

Wolfgang, M. E., & Ferracuti, F. (1967). *The Subculture of Violence: Towards an Integrated Theory in Criminology*. New York: Tavistock.

Wortley, R., & Mazerolle, L. (Eds.). (2008). *Environmental criminology and crime analysis*. Collumpton, Devon: Willan Publishing.

Zhang, C., & Fang, Z. (2013). An Improved K-means Clustering Algorithm Traditional K-mean Algorithm, *1*, 193–199.

Zubin, J. A. (1938). A technique for measuring likemindedness. *Journal of Abnormal and Social Psychology*, *33*, 508–516.