Global Journal of human-social science: C
Sociology & Culture
Volume 16 Issue 4 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 2249-460X & Print ISSN: 0975-587X

Quantitative Criminology: An Evaluation of Sources of Crime Data

By Refat Aljumily

Abstract- Crime data is at the heart of quantitative criminology research in particular and social science research in general. In the past years, many sources of crime data have been proposed to understand, describe and explain crime and criminality, but never before have the majority of these sources been tested using a huge number of crimes and applying different multivariate methods. A large-scale analysis and comparison of various sources of crime data is crucial if current analytical methods are to be used effectively and if new and more powerful methods are to be developed. This article presents the results of a comparison of the four main sources of crime data commonly used in quantitative criminology, in order to determine the best data source that can tell the whole truth about the extent or the true level of crime occurring in a society. Based on the results of these tests, a more comprehensive approach to measure crime is proposed, which represents all categories of crime and covers the offences committed. The result of the analysis is empirically-based, objective, and replicable evidence which can be used in conjunction with existing literature on the quantitative methods in criminology.

Keywords: quantitative; multivariate; hierarchical; vector; space; matrix; SOM; Euclidean distance; Prison statistics; Court records; PCR; CSEW.

GJHSS-C Classification : FOR Code: 160299

Strictly as per the compliance and regulations of:

Quantitative Criminology: An Evaluation of Sources of Crime Data

Refat Aljumily

Abstract- Crime data is at the heart of quantitative criminology research in particular and social science research in general. In the past years, many sources of crime data have been proposed to understand, describe and explain crime and criminality, but never before have the majority of these sources been tested using a huge number of crimes and applying different multivariate methods. A large-scale analysis and comparison of various sources of crime data is crucial if current analytical methods are to be used effectively and if new and more powerful methods are to be developed. This article presents the results of a comparison of the four main sources of crime data commonly used in quantitative criminology, in order to determine the best data source that can tell the whole truth about the extent or the true level of crime occurring in a society. Based on the results of these tests, a more comprehensive approach to measure crime is proposed, which represents all categories of crime and covers the offences committed. The result of the analysis is empirically-based, objective, and replicable evidence which can be used in conjunction with existing literature on the quantitative methods in criminology.

Keywords: quantitative; multivariate; hierarchical; vector; space; matrix; SOM; Euclidean distance; Prison statistics; Court records; PCR; CSEW.

I. Introduction

Sources of crime data grew out of the work of the sociologist Émile Durkheim in the 1897s when suicide rates across different populations were considered as a quantitative data. Sources of crime data changed massively during the 20th century. In the 1915s, the recorded convictions, environment and social experiences were used as statistics to generate a hypothesis for a study or to test hypotheses related to the proneness to criminal behaviour. Since the 1950s, criminology saw the raise of many attempts to measure crime, also in a quantitative context, mainly by British criminology due to the large number of social scientists that developed criminology theories (Dantzker and Hunter, 2000). With the development of data collection methods and analytical methods, many of the old sources and measures have been modified or have continued to be used in one form or another up to the present day. While examiners of quantitative criminology have proposed many sources of measuring crimes over the past years, never before has a large-scale analysis and evaluation of these data sources been conducted to determine which are most useful for measuring crime. Such an evaluation should have done a long time ago: if we are to know, for example, crime levels as to whether crime is increasing or decreasing, then we must use accurate crime data source to adequately draw firm conclusions. The aim of this study is thus to analyse and evaluate the four most commonly available sources of crime data, in order to determine the best source that can tell the whole truth about the extent of crime in a society. In addition, based on the results of these tests, a more comprehensive approach to measure crime is proposed, which represents all categories of crime and covers the offences committed.

This paper is organized as follows: the next section discusses the various sources of crime data typically used in quantitative criminology. Section three presents and describes the data, as well as the three analytical methods that will be used. Analytical test results and their interpretation are included in section four while the conclusions drawn by this study are discussed in section five.

a) Major Sources of Crime Data

A variety of data sources to measure crime have evolved over the years. Each source has different strengths and limitations. The most frequently cited data sources are those collected from official/national crime statistics: official documentation by government and quasi-government agencies. What follows is a variety of these data sources, and it is useful to define each one of these sources and consider briefly the respective advantages and disadvantages of each source.

i. Police Crimes Records (PCR)

It is also known as Crime Related Statistics (CRS) or Police Crime Statistics (PCS). However, whatever name it is given, this source records all the crimes (felonies, misdemeanors, infractions) detected by the police or reported to them. More specifically, police records often include any person(s) of the society who committed a crime or crimes cleared by arrest. The main advantage of this data source is that it provides a government with a summarized account of the crime information obtained regionally and nationally by identifying trends in illegal behavior and patterns of crime, and also particular crimes, recorded over time and across different areas. It is often believed that this data source can also be used to suggest areas for improvement and help in the constant effort to prevent crime from taking place. However, the most important

Author: Somerton Road, Immingham, NE Lincs, United Kingdom.
e-mail: refat.a.70@outlook.com

disadvantage of PCR is that unless a crime has been reported to the police and classified as a criminal act or an offence it will not be recorded. For example, sexual assaults or sexual offences are not always (immediately) reported to the police or unrecorded (i.e. reported to the police but not recorded as an offence), or, as in some cases, are reported long after the incident has committed. Also, there are times when the victims are more willing to report an incident or a crime to the police and, conversely, when the victims are less willing to do it. Another disadvantage of PCR is that victimless crimes (e.g. prostitution, public orders, etc) and all minor crimes are also excluded from being recorded, not to mention that most offending activities do not always result in an arrest. For example, incidents of assault between people who know each other are less likely to be reported to the police or recorded by the police (considered private matter) than incidents of assault between two strangers or incidents of assault with a weapon or a sharp instrument or injury.

ii. *Victim Surveys (VS)*

This source of data aims to record crimes that have not been recorded by the police or have not been reported to the authorities and this way to show the so called 'dark figure' or 'grey figure' of crimes occurring in a society. However, this source is usually done through surveys and interviews with various members of the public. Victim surveys can be conducted at home, by visiting door to door or over the phone. Asking peoples (individuals, households, members of neighborhood, etc)what crimes they have been a victim of or if they have been victims of crimes is a good way to measure crimes and let peoples speak about their attitudes toward police and concerns about crime. The primary advantage of this data source is that it can help in the analysis of reporting behaviour and also can identify the factors that affect reporting decisions. It is often suggested that this data source gives an indication about patterns of crime within society and in particular crimes committed against different sociological and minority groups (e.g. in cases where a range of varied people is involved). An additional advantage is that this data gives an indication of crimes that may not be otherwise reported or considered as a criminal act. One of the main weaknesses of this data is that it records incidents and actions that the police might consider as not criminal since this increases the tendency to make some types of crime over-reported or exaggerated. Being dependent on an individual's honesty and personal understanding of how he/she has been affected or the effect of crime, the reliability of victim surveys is questionable: individuals may provide exaggerated responses or false information. Another disadvantage is that victim surveys account only for crimes that are committed by individuals, i.e. commercial or corporate crimes are not recorded.

iii. *Offender Surveys (OS)*

Surveys of offenders are used just like victimization surveys, but these are for the offenders. The surveys often ask what crime or how many crimes the offender has committed. The main advantage of this data source is that it detects some victimless crimes that have escaped from the police attention such as illegal drug use, prostitution, public order and delinquency crimes, as well as rarely reported crimes such as shoplifting, offender surveys. However, offender surveys have potential for bias. It is often recognized that these surveys reflect the biases and personal career objectives of those involved in reporting crimes. For example, there is a tendency sometimes to under-report more serious crimes (e.g. sexual offences) or to remove the suspects (who are likely to have been detected and convicted) for some serious offences from the sampling frame.

iv. *Self-Report Studies (SRS)*

Like surveys of victims and offenders, this data source asks particular groups or a sample of people as to whether they have themselves committed a crime in a particular period of time. This measure is helpful especially in revealing much about crimes that are victimless and those less observed, and also in identifying hidden offenders who are not caught or detected by the police. In particular this data source makes it possible to find out about the social characteristics of offenders such as ages, gender, social class, and even their location. Besides these advantages, this data source has also a lot of disadvantages. This data source doesn't make good use of a representative sample of a society. Many or most self-report studies are often on simple crimes and young people and students, asking them about their involvement in criminality and law breaking. There are no such studies on professional criminals or drug traffickers for example. Another disadvantage is that this data depends on the honesty of those being surveyed. That is, respondents may lie or exaggerate about their criminal behaviour and, even if they do not deliberately seek to mislead, they may simply be mistaken about their criminal history.

v. *Court Records (CR)*

This data source records all the convictions for criminal offences. It provide accurate information about how many offenders are heard by a court and tried or imprisoned for reported crimes or offences, and what crimes they were convicted of. This data source also provides statistics on type and volume of cases that are received and processed through the criminal court system of a country. However, some believe that one disadvantage of court records is that it underestimate the true extent of crime. That is, after the police identify and arrest a suspect, a relevant court may decide that there is insufficient evidence to mount a prosecution.

Another disadvantage is that a jury may not be convinced by the prosecution's case. A further disadvantage is that in cases where a single incident has multiple offences (e.g. burglary and rape) the offenders are tried and convicted of only one offence they have actually committed (i.e. the most serious crime), and in cases where one or more offences committed by the same person the offenders are tried and convicted of a few of many offences they have actually committed.

vi. *Prison Records(PR)*

Prison records or statistics provides accurate information about the total number of offenders or how many offenders are actually entered prisons to serve ordered sentences and the types of crimes they have committed. The major advantage of this data source is that it shows the relationship between prison numbers and levels and types of crimes, and thereby reveals scope for community solutions to prevent or reduce crime. Another major advantage of prison statistics is that it provides important information relating to prisoners' general categorization, such as ethnicity, gender, religion, sexuality or disability, and prisoners' group types or categories, such as imprisoned juveniles, elderly prisoners, foreign prisoners, minority ethnic prisoners, with statistics for the main types of crimes they have committed. In addition to these advantages, prison statistics provides statistics and information on the criminal justice system such as prisoner re-offending and ex-offenders, prison rehabilitation and education, budgets and costs, staffing, violence, mental health, drugs and alcohol.

Like most things, prison statistics suffers from specific disadvantages related to sentencing policies that may be politically determined. If a government decides on a series of sever measures to restrict, for example, burglaries, theft or drug crimes, then this might translate into sever sentencing policies, which result in more people being imprisoned for those offences, even if the actual rate of offending has not really changed.

vii. *Observation and Reports (OR)*

Crimes are usually detected in two ways: observation and reports by other people. Observation is used to measure crimes when some crimes such as traffic offences and victimless crimes are observed directly by the police. Reports by other people (e.g. households, individuals, neighbourhoods, etc) are also used to measure crimes when someone goes to the police and informs of crime that either he/she observed it or someone else told him/her about it. If we rely on the observation or reports by other people as methods or ways to detect or inform the police of crime, we would find that many crimes will not be well measured. This source of data is far from being the most efficient way to provide information about the actual crime rate in a society. For example, shoplifting or drug use. There are many cases where shoplifting, theft, or drug use will neither be observed by the police nor reported by other people. Therefore, crimes like shoplifting, drug possession and sales, etc. will not be accurately measured.

(Criminal Justice-Quantitative Criminology, 2016; Strengths and Weakness of Crime Statistics and Victimization Surveys, 2015; Siegel, 2012; and Maxfield, 1995).

In summary, the forgoing discussion shows that there is a wide range of available data sources used to measure different categories of crimes and provide statistics on each type, which may be useful for different purposes. It also shows that no single source has a complete advantage over the others; rather it shows that these data sources might be complementary and could be used alongside each other. Each data source has strengths and weaknesses and each provides different information on the nature and extent of crime in a society. Thus a study attempts to address (particular) questions or solve (particular) problems through the analysis of data sources of crime statistics should use one or two or as many data sources as are relevant to a particular research aim.

II. Data and Methods

a) The data

i. *Dataset gathering*

In this study, the researcher used four data sources of crime statistics such as Police Crime Records (PCR), Crime Surveys for England and Wales (CSEW), Prison Population Figures, and Court Figures with statistics and figures related to thirty-six types of crimes committed by adults for the years from 2011 to 2012, 2012 to 2013, and 2013 to 2014.

Figures for crimes that are uncautioned, untried or unsentenced were excluded. These data sources are used by central and local government and police service for planning and monitoring service delivery and for resources allocation. They are also used to inform public debate about crime and the public policy response to it. These crimes are shown in Table/1 below.

Table 1 : Categories of crimes selected from the four data sources

	Type of crime		Type of crime		Type of crime
1	Violence against the person	13	Miscellaneous crimes against society	25	Vehicle related theft
2	Sexual offences	14	Robbery	26	Theft from a vehicle
3	Theft offences	15	Shoplifting	27	Vehicle offences
4	Criminal damage and arson	16	Violence without injury	28	Theft from person
5	Drug offences	17	Domestic burglary in a dwelling	29	Bicycle theft
6	Possession of weapons offences	18	Domestic burglary in a in a non-connected building to a dwelling	30	Other theft of personal property
7	Public order offences	19	Domestic burglary	31	Fraud and forgery
8	Murder	20	Rape of a male	32	Other theft of household
9	Manslaughter	21	Harassment	33	Health and safety offences
10	Infanticide	22	Handling stolon goods	34	Immigration offences
11	Attempted murder	23	Vehicle/driver fraud	35	Assault with a knife or sharp instrument
12	Rape of a female	24	Dangerous driving	36	Blackmail

ii. *Dataset sources used*

The researcher has used multiple sources on crime statistics to address parts of the analysis undertaken in the study. Data sources used are the following:

- Statistical bulletin: Crime in England and Wales 2014 to 2015;
- Statistical bulletin: Crime in England and Wales 2013 to 2014;
- National Statistics: Crimes Detected in England and Wales 2011 to 2012;
- Statistical Bulletin: Crime Outcomes in England and Wales 2013 to 2014;
- National Statistics: Crime Outcomes in England and Wales, year ending June 2014;
- National Statistics: Crime in England and Wales, year ending September 2013;
- Criminal Justice Statistics Quarterly-March, 2013;
- Criminal Justice System Statistics Quarterly-December, 2014;
- Criminal Justice System Statistics March, 2012-March, 2013;
- Home Office Statistical Bulletin: Crime in England and Wales 2010/11;
- National Statistics: Crimes Outcomes in England and Wales statistics and +others 2011 to 2012;
- Prison Population Figures: 2012;
- Prison Population Figures: 2013;
- (XLS) Prison Population Figures: 2014.

Dataset used in this study therefore derives from figures and statistics available in the online Bulletins and collections, published by the home office/ Office for National Statistics (ONS) and ministry of justice.

iii. *Data representation*

In the current application, in order to conduct a fair analysis and comparison of the most commonly used data sources of crime statistics it is necessary that each type of crime be inserted into the same analytical methods and tested using the thirty-six types of crimes listed in Table/1 above. To do so, vector space model (VSM) was used to represent each data source mathematically, that is, each data source was a statistical vector profile with the same (types of crime) information. After each data source was mathematically represented in a vector profile, the associated set of vectors stored together as a matrix row vector, in which the rows are the data sources and the columns the types of crimes. That is, the current data is represented as a 12 x 36 data matrix D in which D_i (for i=1..m) is the *i'th* crime measure, D_j (for j=1..n) is the *j'th* crime, and D_{ij} the value of crime j for measure i.

b) The Methods

The field of quantitative criminology is fundamentally a 20th century movement with the appearance and major advances in computing technology occurring during and immediately after World War II. What began with an emphasis on suicide rates across different populations gradually became focused on the methodological and statistical tools that have led to rapid increase of methodological and statistical tools, and as a result quantitative criminology has developed rapidly. In brief, the field of quantitative criminology now regularly employs statistical univariate methods and statistical bivariate methods (e.s. Boba, 2012). The statistical univariate methods measure only a single variable, for example, frequency distributions or graphical representation of murder. Common univariate methods are measures of central tendency, measures of dispersion, measures of the form of a distribution, mean, median, mode, etc. These univariate methods are used

to examine crimes in terms of a single variable and the results derived from them are therefore described as a simple form of statistical analysis. The statistical bivariate methods measure relationships between two variables, for example, murder rate and burglary rate, or violent crime and total average income. Common bivariate methods are linear regression, measure of association, T-test, Pearson's correlation. This study does not, however, use statistical methods because the analysis of the relevant data is not statistical. The reasoning which led to the decision not to take a statistical approach is as follows. The position adopted here is that each data source of crime statistic consists of various types of crime that have values and these sources can't be described by a single or even two descriptive crimes, and that simultaneous analysis of numerous crimes is required to create a more accurate analysis to evaluate or explain the different measures of crime. Each measure of crime is a combination or more or less numerous crimes, but univariate analysis permits investigation of only one characteristic of a crime at a time, bivariate analysis permits only two, and results for different characteristics are not always or even usually compatible, and the consequence is unclear overall results. This means that univariate and bivariate statistical methods are insufficient for present purposes, and that, if statistical methods are to be used, a multivariate methodology is required. The main class of multivariate statistical methods is multivariate regression, which investigates the relationship between more or less numerous independent variables and one or more dependent ones. At an early stage of the research reported here, however, it became clear that selection of sets of independent and dependent variables was problematic: which variables should be independent, which dependent, and why should the sets, once selected, have an independent-dependent relationship? There may well be answers to these questions, but the decision was taken to abandon multivariate regression and to use an entirely different class of methods. In principle, after all, to decide on the best measures that can give a clear picture about the extent of crime requires only an evidence to be identified; that evidence does not have to be statistical in the sense of having been derived from regression analysis.

For this study, cluster analysis was used. Cluster analysis divides data into clusters based on information found in them that describes the data and its relationship. The data items within cluster are similar or related to one another (since they share common characteristics) and different from or unrelated to the data items in other clusters (since they do not share common characteristics). There is a large number of cluster analysis methods and a large literature associated with each. An extensive range of these methods is discussed and covered in (e.g. Moisl, 2015; Everitt et al. 2001). The methods used here were Agglomerative Hierarchical Clustering (AHC), Principal Components Analysis (PCA), and U-matrix Self-Organizing Map (SOM). The rationale for using these methods is that it is often recognized that that a single class of methods cannot safely be relied on, and that at least one additional method or class of methods must be used to corroborate the results from hierarchical analysis: (i) AHC is based on preservation of distance relations in data space, ii) PCA is a non-hierarchical method based on preservation of data variance, and iii) U-matrix SOM is a nonlinear method based on preservation of data topology.

i. *Agglomerative Hierarchical Cluster Analysis (AHCA)*

Hierarchical clustering is characterized by atree-like structure called a cluster hierarchy ordendrogram. Most hierarchical methods fall into acategory called agglomerative clustering. In this category, clusters are consecutively formed from vectors on the basis of the smallest distance measure ofall the pairwise distance between the vectors. LetX={x1, x2, x3,…,xn} be the set of vectors. We begin with each vector representing an individual cluster. We then sequentially merge these clusters according to their similarity. First, we search for the two most similar clusters, that is, those with the nearest distance between them and merge them to form a new cluster in the dendrogram or hierarchy. In the next step, we merge another pair of clusters and link it to a higher level of the hierarchy, and so on until all the vectors are in one cluster. This allows a hierarchy of clusters to be constructed from the left to right or the bottom to top. The proximity between two vector profiles is calculated as the Euclidean distance between the two profiles taken on by the two vectors. Euclidean distance is the actual geometric distance between vectors in the space and Euclidean distance is the square root of the sum of the squared differences in the variables' values. This is expressed by the function:

$$d_{Euclid\ (BC)} = \sqrt{\left(X_{B-X_C}\right)^2 + \left(Y_{B-Y_C}\right)^2}$$

AHCA is not one specific method but a family of related methods, often minor variants of each other, and it can seem difficult to select an appropriate method for a particular study since all of them operate in a similar way but their calculation (i.e. how distance between clusters is measured) is different. Four AHCA methods based on Sq. Euclidean distance were selected for the analyses that follow: single linkage, complete linkage, average linkage, and Ward method, the aim of which was to examine and differentiate the four data sources at an individual rather than group level with the aid of 21 types of crimes.

ii. *Principal Components Analysis (PCA)*

PCA is a non-hierarchical linear method based on preservation of data variance. Specifically, given D

matrix of 12 data sources, where D described by 36 crimes, principal component analysis re-described the 12 data sources in terms of a number of crimes, such that most of the variability in the original variables was retained. This allowed us to plot the 12 data sources in two-dimensional space and to directly perceive the resulting clusters. The principal components analysis was in a four-stage procedure. The first step was the construction of a symmetric proximity matrix for distances among vectors. The second was the construction of an orthogonal basis for the covariance matrix in such a way that each axis was the least-squares best fit to one of the *n* directions of maximum of variation in D. The third was the selection of dimensions in which we removed the axes that had relatively little variation and kept an *m*-dimensional basis for D, where $m < n$. The fourth step was the projection into *m*-dimensional space, which yielded data matrix D', that is dimensionality-reduced but still had the property of maximum variation in D, that is the total combined variance of all vectors (Jain and Dubes, 1988).

iii. *Self-Organizing Map (SOM) U-Matrix*

The unified distance matrix or U-matrix is a representation of SOM that calculates the nonlinear distances between data vectors and is presented with different colorings. It is based on preservation of data topology. SOM U-matrix generates graphical representations in two-dimensional space such that, given a suitable measure of proximity, vectors which are spatially or topologically relatively close to one another in high-dimensional space are spatially or topologically close to one another in their two dimensional representation, and vectors which are relatively far from one another in high-dimensional space are clearly separated, either by relative spatial distance or by some other graphical means, resulting—in the case of nonrandom data—in a configuration of well-defined clusters (Kohonen, 2001). The analysis was a two-stage process. The first was the training of SOM by loading all the vectors comprising D into the input space. The second was the generation of the two-dimensional representation of the D on the map. For each vector, the values in the input space were propagated through all the connections to the units in the lattice. Because of the variation in connection strength, a given vector activated one unit more strongly than any of the others, thereby associating each vector with a specific unit in the lattice. When all the vectors had been projected in this way, the result was a pattern of activation across the lattice. The U-matrix representation of SOM output used the relative distance between connection vectors to find cluster boundaries. Specifically, given 12×36 output map D, the Euclidean distances between the connection vector associated with each map unit and the connection vectors of the immediately adjacent units were calculated and summed, and the result for each was stored in a new matrix UD, having the same dimensions as D. U was plotted using a color coding scheme to represent the relative magnitudes of the values in U Din which a dark coloring between the vectors corresponds to a large distance and, thus, represents a gap between the values in the input space. A light coloring is the boundaries between clusters or the vectors, indicating that the vectors are close to each other in the input space. Light areas represent clusters and dark areas cluster separators. Any significant cluster boundaries will be visible. The colour scale is displayed near (to the left or right of the map), which contains numbers denoting to the values of U-matrix data vectors and that of the distances between neighboring data vectors.

c) *Analysis and Results*

The position adopted here is that if a more comprehensive picture of crime is the goal, then a source of crime data must represent the total number of crimes that take place and cover all types of crime that people can experience. To put it in quantitative terms, if the resulting structure being tested is valid, then the data sources within a cluster are similar or related to one another and different from or unrelated to the data sources in other clusters. The more consistent the data source is in every clustering analysis, the better and more robust the data source model is likely to be. A source of crime data that doesn't feature consistent clustering would be a data source that lacks information on certain crime categories that could help criminologists or social scientists to draw firm conclusions about the levels and trends of crime and criminality. In this section, the analytical methods described above were applied on PCR (11/12; 12-13; 13-14), CSEW (11/12; 12-13; 13-14), Prison Statistics (11/12; 12-13; 13-14), and Court Statistics (11/12; 12-13; 13-14), and the main determinants for the resulting structures were identified.

d) *AHC methods*

The hierarchical analyses are first presented without comment, and subsequently discussed.

The four AHC methods assign five clusters to the similarity relations among the data sources of crime statistics, as shown in Figure/1 below.

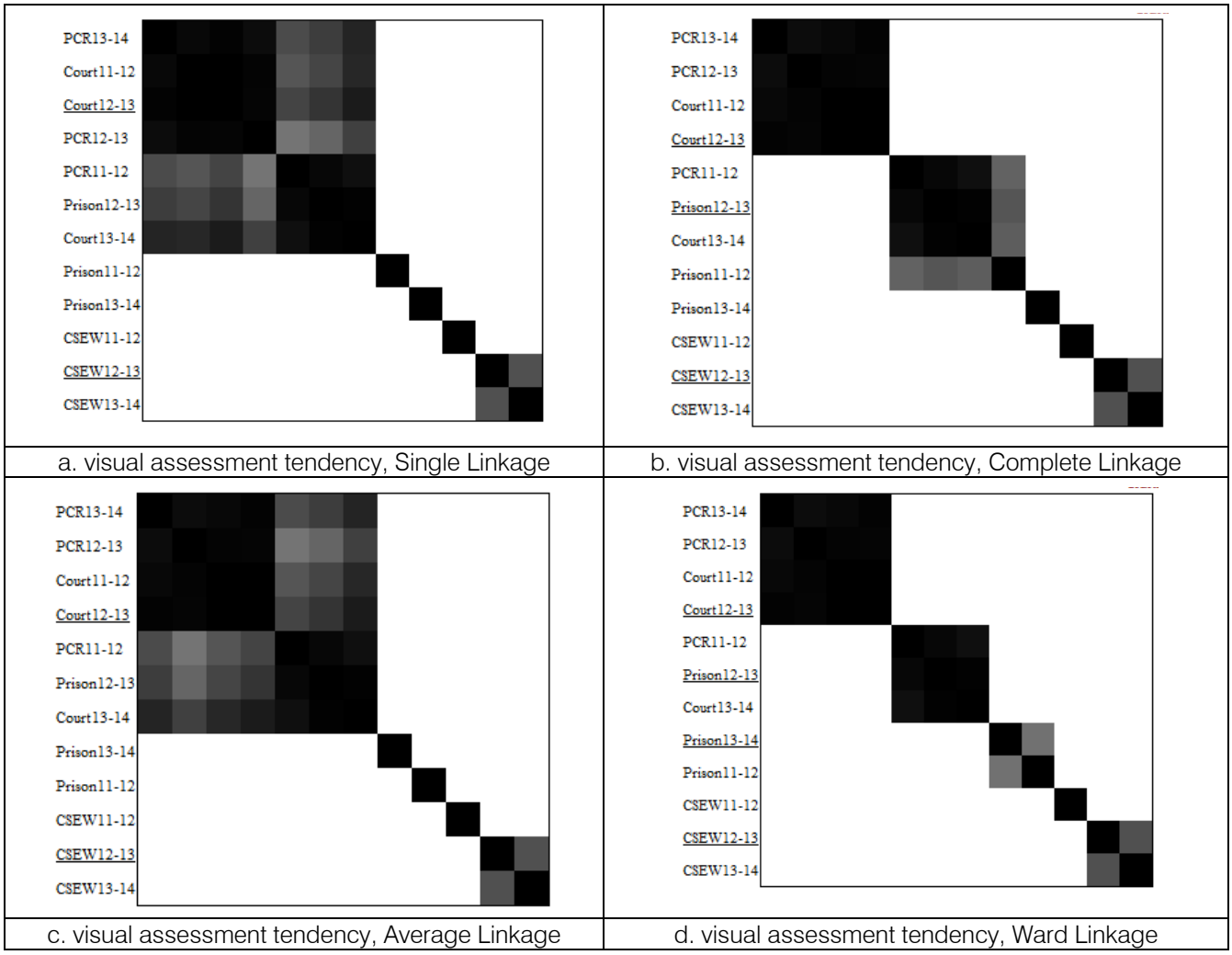

Figure 1 : Visual Assessment Tendency of D

The analytical results given in Figure (1) show a strong match in the way that the data sources of crime statistics are clustered in the hierarchical trees in terms of their relative distance from one another, as shown in Figure (2):

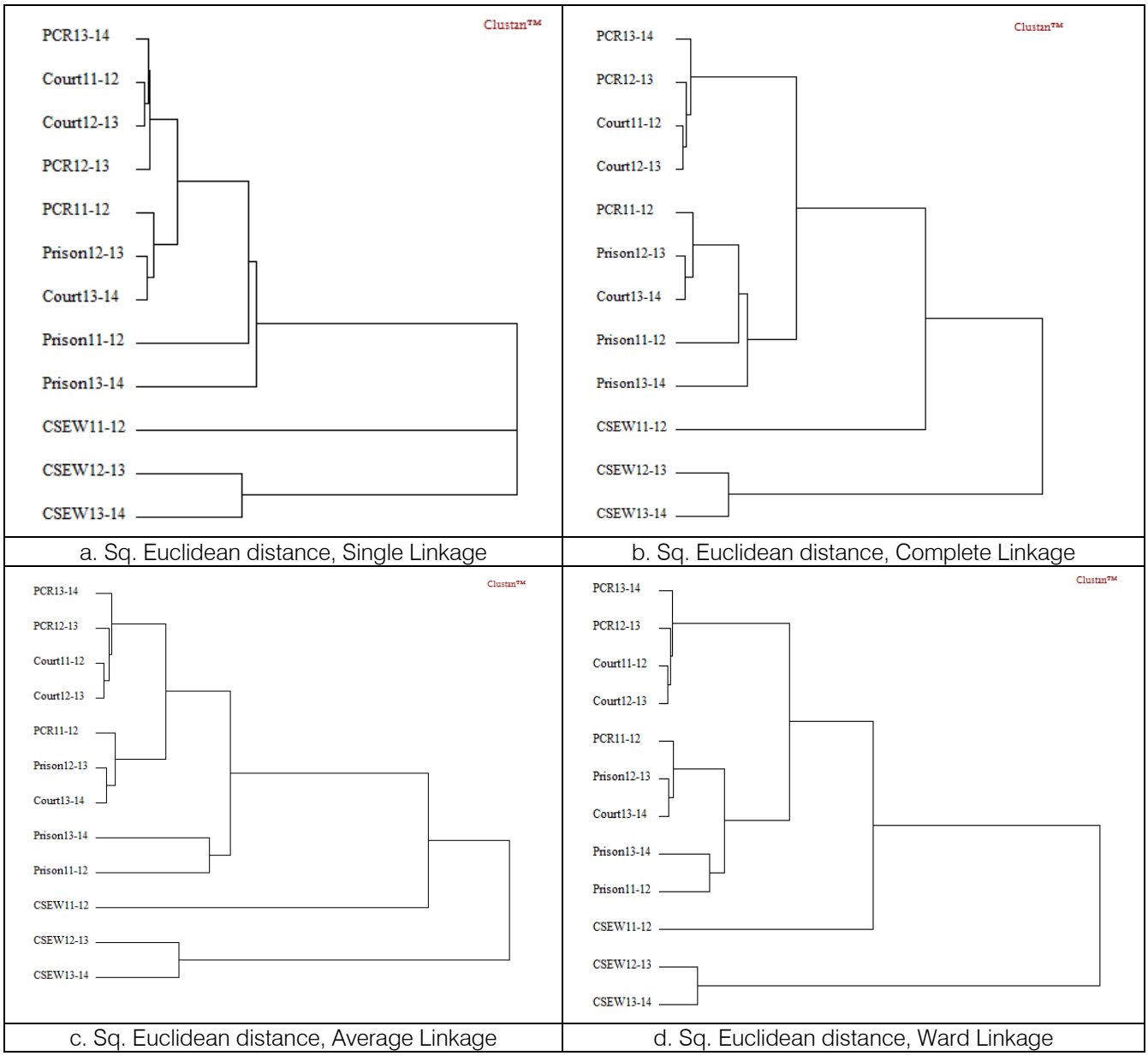

Figure 2 : Four AHC Analyses of D

Nevertheless, to establish the validity of cluster results and select the analysis that best captures the similarity relations among the data source clusters, the cophentic coefficient correlation was used, and the result is shown in Table/2.

Table 2 : Validation of the four AHC in Figure/1 using Cophenetic Coefficient Correlation

AHC method	Cophenetic Coefficient Correlation
Single	0.8075
Complete	0.6123
Average	0.9119
Ward	0.5443

The AHC tree generated by Average linkage seemed to fit the data matrix D more well than the clusterings produced by Single, Complete, and Ward method. Average linkage defines the degree of closeness between any pair of subtrees (X,Y) as the mean of the distances between all ordered pairs of objects in X and Y: If X contains *x* objects and Y contains y objects, the distance is the mean of the sum of (X_i, Y_j), for $i = 1...x$, $j = 1...y$.

$$D_{avg}(A,B) = \frac{\sum_{1=1..m,\, j=1..n} d(a_i \in A, b_j \in B)}{m \times n}$$

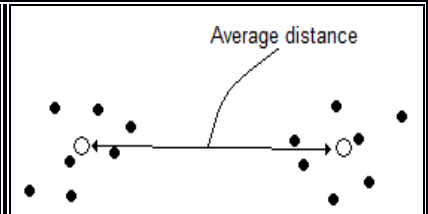

Figure 3 : Average AHC

In the light of the validation results, the analysis generated by average linkage method was used in the present analysis.

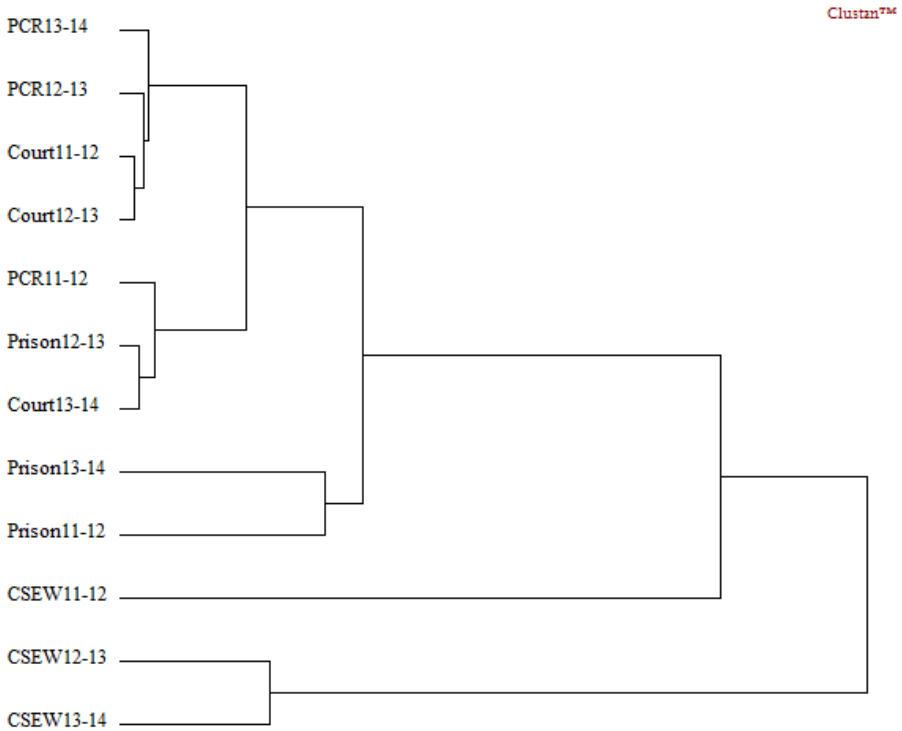

Figure 4 : Average AHC analysis of D

In this figure there are five main clusters. The first cluster consists of four data sources grouped into three sub-clusters: the first sub-cluster consists of (PCR13-14), the second consists of (PCR12-13) and the third sub-cluster consists of (Court11-12 and Court 12-13). The second cluster consists of three data sources grouped into two sub-clusters: the first sub-cluster consists of (PCR11-12) on its own and the second sub-cluster consists of (Prison12-13 and Court13-14). The third cluster consists of two data sources (Prison13-14 and Prison11-12). The fourth cluster consists of data source on its own (CSEW11-12). The fifth and last cluster consists of two data sources (CSEW12-13 and CSEW13-14).

i. *PCA*

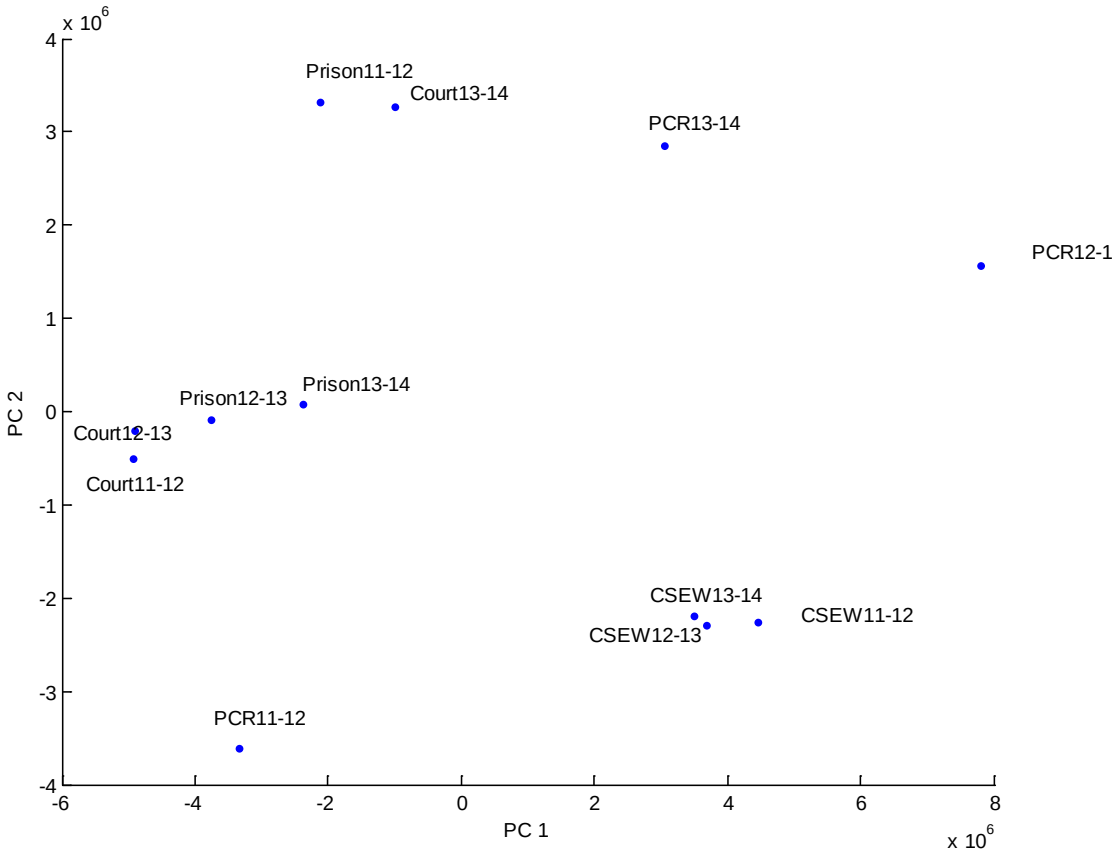

Figure 5 : PCA of D

PCA results in Figure (5) shows the followings:

- The first cluster consists of three data sources (Prison11-12, Court13-14, and PCR13-14). Prison11-12 and Court13-14 are plotted next to each other in the 2-D space, but PCR13-13 is far apart from them.
- The second cluster consists of only one data source (PCR12-13) located on its own in the space.
- The third cluster consists of three data sources (CSEW13-14, CSEW12-13, and CSEW11-12). These are plotted close one another in the 2-D space.
- The fourth cluster consists of only one data source (PCR11-12) located on its own in the space.
- The fifth cluster consists of four data sources (Prison12-13, Prison13-14, Court12-13, and Court11-12). These are positioned close to each other in the 2-D space.

ii. *U-matrix SOM*

As with the AHC and PCA, the SOM one is first presented without comment, and subsequently discussed. The analysis of the data sources using SOM represented by U- matrix is presented in Figure (6).

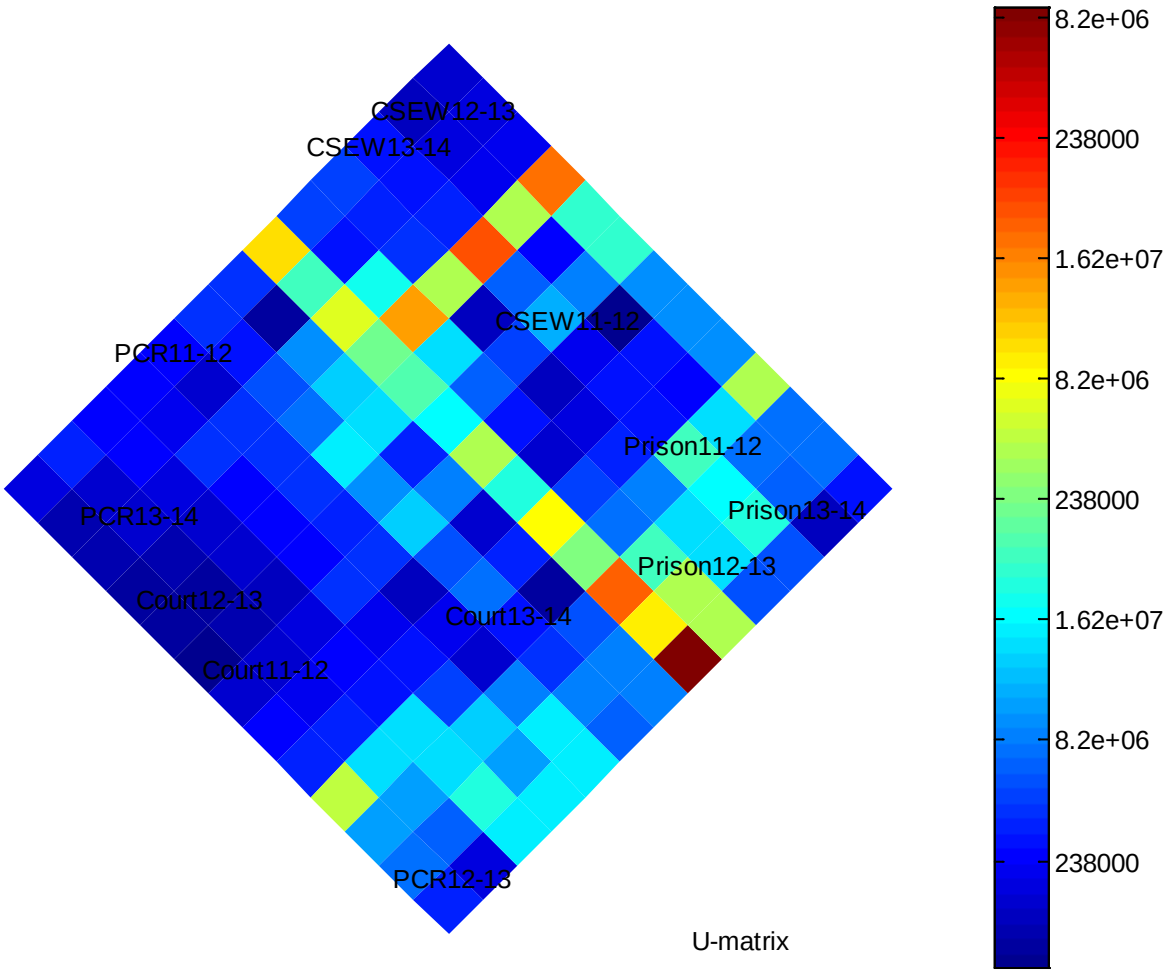

Figure 6 : U-matrix SOM of D

In presenting and understanding the results given in Figure (6), the above discussion of SOM U-matrix representation have to be kept in mind. Specifically, the yellow or green/ light areas of the maps are the regions where the data sources are topologically close, that is, where they cluster, and the blue-green-orange/dark areas are where they are topologically far apart. However, in this figure, we obtained five main clusters:

- The first cluster (top part of the map) consists of two data sources (CSEW12-13 and CSWE13-14). The data sources in this cluster are positioned next to each other in the map.
- The second cluster (right part of the map) consists of the data source (CSEW11-12) which is assigned to one cluster in the map.
- The third cluster (right part of the map) consists of three data sources (Prison11-12, Prison13-14, and Prison12-13). The data sources in this cluster are clustered close to one another in the map.
- The fourth cluster (bottom part of the map) consists of one data source (PCR12-13) which is assigned to one cluster in the map.
- The fifth and final cluster (left part of the map) consists of five data sources (PCR11-12, PCR13-14, Court12-13, Court11-12, and Court13-14). (PCR11-12, PCR13-14, Court12-13, and Court11-12 are grouped next to each other in the map, but (Court 13-14) is in the close periphery of the Prison cluster.

Although procedures vary from one method to another, comparison showed a close match between the results from the AHC, the PCA, and the U-matrix SOM analysis. Specifically, there is a good degree of correspondence between the data sources in the five main clusters generated by AHC, PCA, and U-matrix SOM clusters:

- CSEW/12-13, 13-14, 11-12: the three versions of CSEW are more or less closely adjacent in the space/map.
- PCR/11-12/13-14/12-13: two versions of PCR are relatively close to each other in the map or space. The third version is in a distant cluster or region of the map.
- Prison/11-12, 12-13/13/14: the three versions of Prison are either immediately adjacent in the boundary region of the cluster, or nearby in an immediately adjacent cluster.

- Court/12-13, 11-12, 13-14: two of them are positioned close to another in a boundary region or space while the third is placed nearer Prison cluster.
- Among all the pairs of data sources, there are two pairs that consistently closest: Prison 12-13 and Court 13-14 and Prison 13-14 and Prison 11-12. There's some slight variation in degree of closeness to these, but the overall picture is clear.

On the basis of this comparison it is possible to define two core clusters, where a core cluster consists of those data sources that are assigned to it by the AHC, the PCA, and the SOM analyses:

Core cluster/1(AHC/PCA/SOM)	Core cluster/2 (AHC/PCA/SOM)
Prison12-13, Court13-14 Prison11-12, Court13/14	Prison13-14, prison11-12 Prison13-13, Prison12-13

These results show similarity in a way that is quite easy to interpret.

iii. *Interpretation of the Results*

How many crimes and what types of crime are committed are one of the most fundamental characteristics of arobust source of crime data. Which source of crime data indicates the most and which one the least? Although this may sound difficult to answer, it really is not. The answer is Prison Records and Court Records have the most value of all sources of crime statistics. The justification for this claim is very straightforward: each of these two sources of crime data clusters has the same types of crime that differentiate it from the others.

The difficulty with evaluating different sources of crime statistics is that the interpretation of the results would be highly subjective, and very often this may create a misleading conclusion. This means that one qualified quantitative criminologist may not interpret the same information in the same way as another qualified quantitative criminologist. It is, however, possible to objectify it to some degree using a quantitative criterion, which is now proposed. Cluster analysis clusters multidimensional data vectors on the basis of their relative similarity: data vectors in any given cluster are more similar to one another on some measurement criterion than they are to vectors in any other cluster. In the present application, the four sources of crime data were clustered on the basis of crime statistics vectors. The existence of distinct clusters therefore implies that each cluster has a characteristic crime statistics profile which distinguishes it from the others. By comparing the crime statistics profiles of the four data source clusters, therefore, it should be possible to determine the crime categories in which they differ most, and, on the basis of the figures of these categories, to identify the categories of crime of the respective data source clusters. What is a 'crime statistics profile' for a cluster? It is an average column vector constructed from the various data source statistics vectors that constitute the cluster by adding the corresponding crimes of each source column vector and taking the mean of the sum:

$$P_j = (\sum_{i=1\ldots n} C_{ij})/n$$

Where *j* is the index to the *j* thcrime of the profile vector *p*, *i* indexes the vectors of the source of crime data C that comprise the cluster, and *n* is the total number of vectors in the cluster. Such a profile vector is constructed for each of the two core clusters. For the data matrix D, the average column vector of crime values for each source of data was calculated and the results were bar plotted. The amount of variability was used as a criterion to select a relevant set of crimes. A crime type with a larger amount of variability in its average than the other types of crime was taken to be the most important discriminator between the (core) clusters because there was much change in the values of that crime throughout crime data source row vectors, i.e. if the difference is large, it is clearly significant. Various possibilities were tried, and it was found that 12 out of the 36 categories of crime were sufficient for the present purpose. These crimes are shown in Table/3, and their bar plots are shown in Figure/7.

Table 3 : A set of 12 crimes based on variance

	Crime type		Crime type
1	Violence against the person	7	Shoplifting
2	Sexual offences	8	All other theft offences
3	Theft offences	9	Violence without injury
4	Criminal damage and arson	10	Domestic burglary
5	Drug offences	11	Vehicle offences
6	Robbery	12	Theft from person

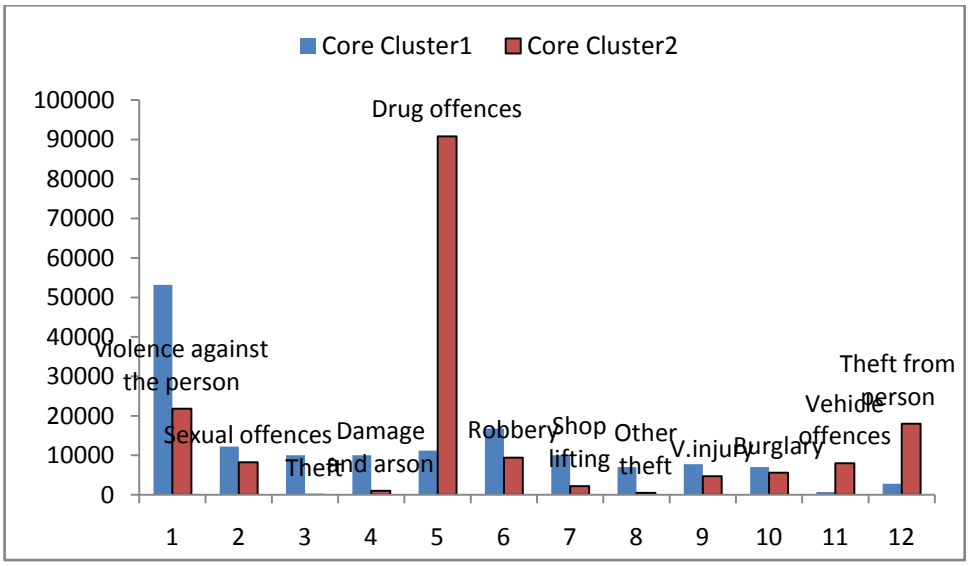

Figure 7 : Bar plot of 12 crimes based on variance

Now it is possible to determine which crimes are most and least characteristic of each cluster, and which differentiate them most. It is evident from the plot of the 12-average column crime vectors that the variation in the average bar representing the crimes of 'drug offences' and 'violence against the person' more than the average bars representing other crimes and that the crime of 'drug offences' is the most important crime in the consistent clustering of (Prison12-13) and (Court13-14) and/or (Prison13-14) and (Prison11-12) close to one another.

III. Conclusions

The study has focussed on the main seven sources of crime data, an area which has not garnered a great deal of widespread attention, and presented the results of testing four of these sources using a large number of crimes and applying three different multivariate analytical methods. For the first time in the history of quantitative criminology, criminologists and social scientists now have the opportunity to use the most useful or reliable source of crime statistics to adequately test theories of offending and victimization as well as to assess the effectiveness of public policies.

In this study, the generated data was assessed using Visual Assessment Tendency and the results were validated using Cophenetic Coefficient Correlation and different clustering methods in combination.

The analysis of the conducted test shows that Prison records and Court records are the most reliable measure to represent the true extent of crime or the total number of crimes that take place.

However, no indications were found supporting the two other sources of crime statistics, namely PCR and CSEW. This could possibly be ascribed to not including or covering all forms of crimes in these two types of data. The PCR and CSEW measure crime in different ways since each covers different views of crime.PCR records exclude crimes that are not reported to, or not recorded by the police. Also not involved are allless serious crimes (e.g. motoring offences), and much more. Due to quality recording concerns, PCR doesn't record crimes consistently (probably due to changes in police recording practice); therefore the true level of recorded crime is understated. CSEW excludes crimes that are difficult to estimate robustly (e.g. sexual offences, fraud and much more) or that have no victim who can be interviewed (e.g. homicides, and drug offences). Of course, this does not mean that CSEW and PCR are invaluable, but it does mean that, on the one hand, CSEW is useful for covering crimes not recorded to the police and providing information on the characteristics of people they interview and the relationship between victims and police. On the other hand, PCR is more useful and more valid in providing information about the nature of crimes in term of time and place, the characteristics of offenders, and the relationship between victims and offenders, etc.

Rather, the analysis of the test indicated two categories of crime that have the direct effect on clustering Prison Records and Court Records all together. These are 'drug offences' and 'violence against the person'. Prison figures and Court statistics contain information on the number and characteristics of people tried or convicted; information that the other data sources lack. Prison Records and Court Records can also provide information on the level of criminal activity for a particular type of crime, which other data sources can't provide (a separate database on a particular crime type is out of the question here). The bottom line is that Prison Records and Court Records are representative of the officially recorded crimes and are closest

procedurally to the actual amount of crime committed; together they provide a more comprehensive picture of crime.

All things considered, criminologists and social scientists are advised to take both Prison Records and Court Records into account when tracking trends and patterns in the crime rate or when formulating a conclusion for a study. Nevertheless, as with every measure of crime, Prison Records and Court Records do not provide information on the dark figure of crime or unknown or unrecorded offences.

In this study, cluster analysis methods and techniques are proven to be effective in analysing different crime data sources described by a large number of crimes and in identifying a particular crime type. We hope expansion in the use of cluster analysis in the future as multivariate tools in the resolution of different problems in criminology and criminal justice research.

The author explicitly document the approach to the data, ensuring that the results presented here are objective and replicable.

IV. Acknowledgments

The author wishes to thank the Editor of the journal and the reviewer(s) for devoting their time and effort towards this manuscript.

Conflicts of Interest

The author declares no conflict of interest.

References Références Referencias

1. Anil K. Jain, and Richard C. Dubes. *Algorithms for Clustering Data*. Englewood Cliffs: Prentice-Hall, 1988.
2. Criminal Justice-I Research Net. (2016). *Quantitative Criminology* .Retrieved from http://criminaljustice.-iresearchnet.com/criminology/reserachmethods/-quantitative-cri.
3. Dantzker, ML and Ronald D. Hunter. *Research Methods for Criminology and Criminal Justice*. Butterworth-Heinemann: A Primer, 2000.
4. Everitt, B. S.; Landau, S.; and Leese, M. *Cluster Analysis*. 4th ed. London: Arnold, 2001.
5. Hermann Moisl. *Cluster analysis for corpus linguistics*. Berlin: De Gruyter Mouton, 2015.
6. Larry J. Siegel. *Criminology*. 11th ed. Australia: Wadsworth Cengage Learning Publishing, 2012.
7. Maxfield, M.G. *Research methods For Criminal Justice and Criminology*. Ed. Earl Babbie. California: Wadsworth Publishing Company, 1995.
8. Rachel Boba. *Crime Analysis with Crime Mapping*. California: SAGE Publications, Inc, 2012.
9. Teuvo Kohonen. *Self-Organizing Maps*.3rd ed. Berlin: Springer, 2001.
10. UK Essays. (2015). *Strengths and Waeness of Crime Statistics and Victimization Surveys*. Retrieved from http://www.ukessays.com/essays/criminology.

References to data used

11. Chris Mckee. (2014). Home Office Statistical Bulletin: Crime Outcomes in England and Wales2013/2014 (first edition). 17th July, 2014.
12. Crimes in England and Wales, years ending June 2014.
13. Gavin Berman and Aliyah Dar. Prison Population Statistics. Social and general statistics, 2013.
14. Gov.UK. XLS Prison Population, 2014.
15. Gov. UK. Prison Population Statistics, 29 July, 2013.
16. John Flatley. (2016). Statistical Bulletin: Crimes in England and wales, year ending September 2015.
17. National Statistics: Criminal Justice System. Statistics Quarterly: December, 2014. Ministry of Justice.
18. National Statistics: Crimes Detected in England and Wales 2011 to 2012: Statistics, crime outcomes in England and wales statistics and–others. Home Office-July 2012.
19. Official Statistics: Criminal Justice Statistics Quarterly-March, 2013. Ministry of Justice.
20. Official Statistics: Prison Population Figures: 2012. Ministry of Justice 11/July/2013.
21. Official Statistics: Crimes outcomes in England and wales 2013 to 2014: Data Tables.
22. Statistical Bulletin: Crimes in England and wales, year ending September 2013.