ACTA UNIVERSITATIS LODZIENSIS
FOLIA IURIDICA 94, 2021

https://doi.org/10.18778/0208-6069.94.06



*Andrzej Porębski**
https://orcid.org/0000-0003-0856-5500

# APPLICATION OF CLUSTER ANALYSIS IN RESEARCH ON THE SPATIAL DIMENSION OF PENALISED BEHAVIOUR

**Abstract.** This paper is focused on some of the possibilities of the use of cluster analysis (clustering) in criminology and the sociology of law. Cluster analysis makes it possible to divide even a large dataset into a specified number of subsets in such a way that the resulting subsets are as homogenous as possible, and at the same time differ from each other substantially. When analysing geographical data, e.g. describing the location of crimes, the result of cluster analysis is a division of a territory into a certain number of coherent areas based on an objective criterion. The division of the territory under study into smaller parts is more insightful when the clustering method is applied compared to an arbitrary division into official administrative units.

The paper provides a detailed description of hierarchical cluster analysis methods and an example of using the Ward's hierarchical method and the *k*-means combinational method to divide data on crime reports in the city of Baltimore between 2014 and 2019. The analysis demonstrates that the resulting division differs considerably from the administrative division of Baltimore, and that increasing the number of groups emerging as a result of cluster analysis leads to an increase of variance of variables describing the structure of crime in individual parts of the city. The divisions obtained using clustering are used to verify the hypothesis on differences in crime structure in different areas of Baltimore.

The main aim of the paper is to encourage the use of modern methods of data analysis in social sciences and to present the usefulness of cluster analysis in criminology and the sociology of law research.

**Keywords:** cluster analysis, environmental criminology, geography of crime, crime in Baltimore, computational social science.

# ANALIZA SKUPIEŃ W BADANIACH NAD PRZESTRZENNYM WYMIAREM ZACHOWAŃ SPENALIZOWANYCH

**Streszczenie.** Prezentowany artykuł poświęcony jest wykorzystaniu w socjologii prawa oraz kryminologii środowiskowej jednej z nowoczesnych metod obliczeniowych przydatnych do badania dużych zbiorów danych – analizy skupień (grupowania; klasteryzacji). Metoda ta pozwala na podzielenie zbioru obserwacji na ustaloną liczbę podzbiorów takich, że elementy tego samego podzbioru są do siebie możliwie podobne, a elementy różnych podzbiorów – możliwie odmienne. Jeśli dane dotyczą położenia geograficznego, na przykład umiejscowienia przestępstw, rezultatem wykorzystania analizy skupień będzie podział obszaru na ustaloną liczbę wewnętrznie spójnych rejonów według zobiektywizowanego kryterium. Podzielenie badanego terytorium na mniejsze części

* Jagiellonian University in Krakow, Poland, Faculty of Law and Administration; AGH University of Science and Technology, in Krakow, Poland, Faculty of Management, poreand@gmail.com

z zastosowaniem klasteryzacji wydaje się być lepszym rozwiązaniem od kryteriów stosowanych tradycyjnie, w dużym stopniu arbitralnych, takich jak na przykład podział administracyjny.

W pracy przedstawiono szczegółową charakterystykę hierarchicznych metod analizy skupień, a następnie wykorzystano metody kombinatoryczną *k*-średnich oraz hierarchiczną Warda do podziału zbioru danych o zgłoszeniach przestępstw w mieście Baltimore w latach 2014–2019. Wykazano, że powstały w ten sposób podział różni się w sposób znaczący od podziału administracyjnego Baltimore, a także że zwiększanie liczby grup powstających jako wynik analizy skupień prowadzi do pożądanego w pewnych przypadkach wzrostu wariancji zmiennych opisujących strukturę przestępczości w poszczególnych częściach miasta. Utworzone przy użyciu klasteryzacji podziały wykorzystano także do zweryfikowania hipotezy o odmienności struktury przestępczości w różnych obszarach Baltimore.

Głównym celem pracy jest zachęcenie do stosowania w badaniach społecznych nowoczesnych metod analizy danych oraz pokazanie, że analiza skupień może być cennym narzędziem w kryminologicznych i socjologicznoprawnych analizach poświęconych relacji między prawem a przestrzenią.

**Słowa kluczowe:** analiza skupień, kryminologia środowiskowa, geografia przestępczości, przestępczość w Baltimore, obliczeniowe nauki społeczne.

## 1. INTRODUCTION

The rapid development of information technologies observed over the last years has significantly influenced not only sciences, but also fields of non-exact research. The potential of employing calculation technologies and big data in social sciences was first stressed within the computational social science movement, which has been growing rapidly (Lazer et al. 2009; Conte et al. 2012). Modern data analysis methods combined with a remarkable amount of data being continuously gathered by electronic devices and public information systems create unprecedented opportunities for social sciences research.

Sociologists and criminologists studied the influence of environmental factors on crime long before the era of computers. Already in the 19th century, works on the topic provided quantitative analyses of crime with respect to various geographic areas (e.g. Glyde 1856). At present, the entire branch within crime research can be distinguished, where emphasis is placed on how space affects criminal behaviour. This branch should be understood as containing both the field of environmental criminology, where the issue is usually analysed from an individual's point of view (Wortley and Townsley 2016), and the geography of crime, which tends to use a macroscopic description. A detailed overview of Polish literature dealing with the geography of crime can be found in (Mordwa 2016), in spite of the scarcity of publications on the topic (Mordwa 2016, 197). Recently, more attention has been drawn to the relation between space and – not only penal – law, which has resulted in the first Polish monography concerning the topic (Dudek, Eckhardt and Wróbel 2018).

Spatial research on crime constitutes a great example of a field where modern methods of data analysis and big data can be used. Unfortunately, modern

methods of quantitative analysis are still used rather infrequently in the fields of sociology of law and criminology. Contributions by, for instance, Kinga Kądziołka (Kądziołka 2016a; 2016b), where modern statistical models are employed in crime research, remain exceptions rather than a rule in the Polish subject literature, which is highly unfavourable.

Since modern data analysis methods can facilitate a better understanding of the spatial dimension of crime, they require a closer investigation. In this paper, I present the possibility of applying one of such methods, namely cluster analysis, which seems to be a tool perfectly fit for work with geographical data.

## 2. METHODOLOGICAL ISSUES

### 2.1. Objectives and Research Hypotheses

The primary aim of this work is to argue that the method of cluster analysis can be successfully applied to criminological and sociological analyses focused on links between space and crime, in which geographical data are used. Furthermore, it is the author's goal to encourage such analyses. In order to fulfil these goals, the cluster analysis method will be described in detail and an example of its usage will be presented. The example will be based on crime data from the city of Baltimore. The attempt of verificationying of the following hypotheses will be made:

1. The crime clusters resulting from performing a cluster analysis are substantially different from the administrative division of the city of Baltimore into districts.

2. The structure of crime is different across crime clusters, in particular with respect to the rate of a) murder, b) theft, c) car theft.

3. Variance of crime grows with an increasing number of areas defined by means of a cluster analysis. For a sufficiently large number of such areas, the variance for the areas exceeds the variance for administrative districts – in particular when it comes to a) murder, b) theft, c) car theft.

### 2.2. Cluster Analysis Method Description

Cluster analysis (or clustering – these two terms will be used interchangeably) is one of the statistical methods, which makes it possible to divide a dataset into a particular number of subsets grouping similar objects (observations) based on a defined similarity measure. The aim of cluster analysis – fulfilled in an ideal situation, which is of course not always the case – is to detect such groups (clusters) in the dataset that are naturally present in the set's structure as a result of data character. Distinguishing such groups should furthermore be interpretable

in a sensible way. In other words, the goal of the analysis is to split the initial set of objects into such subsets that elements belonging to different subsets should vary between one another more than elements of the same subset (see: Wierzchoń and Kłopotek 2015, 19–20; Krzyśko et al. 2008, 345–346).

A slightly more formal definition of cluster analysis is worth mentioning at this point: "a data analysis tool used to group *m* objects described with a vector of *p* features into *K* nonvoid, disjoint and, as much as possible, 'uniform' groups-clusters" ([translation mine – A.P.] Krzyśko et al. 2008, 345). Two aspects of this definition must be stressed. Firstly, the number of groups into which the set of objects will be split is not defined by the algorithm itself and the decision lies with the analyst. Secondly, objects are grouped based on the data known about them, that is the data the objects are described with. It is of utmost importance to be aware of the fact that the clustering of objects will always refer to their chosen representation. Another representation of the same objects, which means a different choice of features relevant to the research, is likely to result in obtaining radically different clusters. Naturally, the grouping algorithm has no knowledge about the real nature of objects, only about their features which have been intentionally formalised.

The most widespread clustering methods can be divided into hierarchical – based on linking and dividing observations – and combinational – in which a particular clustering performance function is optimised. Hierarchical methods can be further split into agglomerative (initially each object is a separate cluster, then the clusters are merged), and divisive (at the beginning there is a single group comprising all objects that are gradually divided into smaller and smaller clusters). Other methods are used less often and are more sophisticated. They include, among others, relational and graph methods. The most popular combinational method is called the *k*-means or centroid method. It must be borne in mind that for reasons of space, the subsequent sections of this paper will refer to the hierarchical agglomerative methods only. However, all the remarks regarding object formalisation and distance measures remain valid not only for the chosen group of clustering methods.

### *2.2.1. Formalisation of Real Objects*

The problem of real objects formalisation requires a more detailed description. The proper choice of examined objects features, i.e. aim-oriented objects formalisation, will be the factor distinguishing a well-designed analysis from a "blind" (haphazard) one. Formally, following the basic definition that will be considered in this paper, the problem is fairly simple.[1] A set of *n* objects is given by the following matrix:

[1] Such a presentation of the problem meets the requirements of this paper. For a more elaborate description of object formalisation, and a more complete formalisation of the entire problem, see Wierzchoń and Kłopotek (2015, 20–23).

$$X = (x_1, \ldots, x_n)^T,$$

whose each element is described by a $p$-dimensional vector ($p$ – the number of features):

$$x_i = (c_{i1}, \ldots, c_{ip}) \text{ where } i \in \{1, \ldots, n\}.$$

Then, $c_{ij}$ is the $j$-th feature of the $i$-th object, and the vector $x_i$ is that object's image (i.e. the vector of object's features).

In order to partially deformalise the above conclusions, it is useful to present the matrix in a tabular form. Let the rows be objects, and columns their features:

**Table 1.** A generalised matrix of $p$ features for $n$ objects

| Feature / Object | **Feature 1** | **…** | **Feature *p*** |
|---|---|---|---|
| Object 1 | Value of feature 1 for object 1 | … | Value of feature $p$ for object 1 |
| … | … | … | … |
| Object $n$ | Value of feature 1 for object $n$ | … | Value of feature $p$ for object $n$ |

Source: own elaboration.

As a result of assuming non-void and disjoint clusters, $n$ must be not less than the number of groups $K$. In practice, only the cases where $n$ is visibly higher than $K$ will be interesting.

Object formalisation allows displaying them as points in a $p$-dimensional space with each dimension representing a feature – that will be used further in this work. Better understanding of clustering's key point, the mathematical modelling of real objects, can be facilitated by noticing that each such object (for instance a man, an animal, an organisation, a usable item) can be to some extent described by enumerating its characteristics of interest. Even though the list of all features of any real object is enormous (if not infinite), the number of features relevant for a particular problem will certainly be finite and encompassed by a small set (a set of low cardinality).

From the perspective of crime research viewed from the spatial context, the features relevant for an object will naturally refer to the incident location, and disregard, for instance, offender's characteristics or incident's type. The formalisation of an incident will be based on its geographical coordinates. Cluster analysis will then refer to points of a plane, which can be intuitively understood as dots on a map.

### *2.2.2. Similarity (Dissimilarity) Measure*

It is not enough just to formalise the objects. By definition, clustering includes searching for groups of similar objects, which differ from the objects in other groups. Such a search would not be possible were it not for defined mathematical

criteria of similarity. Thanks to them, it is possible to calculate similarity of vectors on the basis of the data these vectors consist of. Such a criterion will be called the similarity measure or dissimilarity measure. It may seem counterintuitive that these two expressions can be used interchangeably but, indeed, the maximal dissimilarity is equivalent to the minimal similarity, and *vice versa*, therefore the maximisation of similarity is the minimisation of dissimilarity.

Mathematically speaking, a similarity measure is a function satisfying some specific (yet irrelevant for the purpose of the present paper) conditions in form $s: X \times X \rightarrow \mathbb{R}^+ \cup \{0\}$, i.e. associating each ordered pair of the considered set's Cartesian product with exactly one nonnegative real number.

As it has already been mentioned, having formalised objects, one can position them as points in a $p$-dimensional Euclidean space. Such display is a very convenient one since it allows for a very intuitive presentation of similarity (dissimilarity) as distance in such space. It is worth mentioning that for the present study, the distance between points is exactly the distance between the acts of crime. Thus, the interpretation of the mathematical distance is exceptionally simple in this case.

Even though numerous similarity measures can be considered and utilised (among which various variants of correlative measures using the Pearson, Spearman or Kendall correlation coefficient are worth mentioning), the most widely employed are distance functions, also called metrics in the subject literature. For a measure to be considered distance (metric), function $d: X \times X \rightarrow \mathbb{R}^+ \cup \{0\}$ (already here it can be noticed that its set of destination is nonnegative numbers) must fulfill the following conditions:

1) $d(x, y) = 0 \Leftrightarrow x = y$,
2) $d(x, y) = d(y, x)$,
3) $d(x, y) \leq d(x, z) + d(z, y)$.

Defining the similarity measure as a distance is convenient in this way as it lets the created system of formal terms to be highly intuitive. Objects "closer" to each other, i.e. less distant from each other, will also be less different, and thus more similar.

The most popular as well as the simplest distance used in cluster analysis is the Euclidean distance:

$$d_E(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2},$$

where $\boldsymbol{x_i}$ – value of the $i$-th feature of the object $x$,
$\boldsymbol{y_i}$ – value of the $i$-th feature of the object $y$,
$\boldsymbol{p}$ – number of features for the considered objects.

The Euclidean distance will be used in the analysis pursued in this paper. This being said, it is worth mentioning that when considering points that cannot be connected as the crow flies, the Manhattan distance (city block distance, taxicab distance) can be used instead:

$$d_M(x, y) = \sum_{i=1}^{p} |x_i - y_i|.$$

For the purpose of this paper, no further considerations of similarity measures are necessary. However, one should bear in mind that numerous measures exist (Walesiak 2002; Wierzchoń and Kłopotek 2015, 136–139; Krzyśko et al. 2008, 23–34). An adequate choice of either measure based on the characteristics of the available data as well as on the purpose and assumptions of the research is a vital element of cluster analysis.

Having defined the measure to be used for a given analysis, one can obtain a matrix with aggregated information about the similarity of any two considered objects. If the similarity measure is a distance, such a matrix is called a distance matrix. It is the base for hierarchical clustering.

*2.2.3. Clustering Algorithm*

As mentioned above, various types of object clustering algorithms can be distinguished (Wierzchoń and Kłopotek 2015; Krzyśko et al. 2008, 345–361). The presentation of both hierarchical and combinational methods lies outside of the scope of the present contribution. More thorough accounts are available for free on the internet, including both simpler (Gareth et al. 2017, 385–401, 404–407, 410–413) and more formalised ones (Hastie, Tibshirani and Friedman 2009, 501–528).

In this section, the agglomerative method used in this study will be presented. Several reasons justify its choice. First and foremost, this method is easy to present using informal terms, which matches the main goal of this paper – presenting cluster analysis. Secondly, unlike the combinational method of *k*-means, the agglomerative method does not require defining the number of clusters in advance. This facilitates a researcher's task by allowing for the modification of this number at later stages of data analysis. In this and similar research, this means dividing a city into a number of areas after the first iteration of the algorithm. Thirdly, the agglomerative methods, contrary to the combinational *k*-means method, are 1) deterministic (non-random): their result will always be the same for same data and measure; and 2) "monotonic": changing the number of clusters by *m* will make objects of exactly *m* groups move to another cluster, while the other groups will remain unchanged. Moreover, the agglomerative method makes it possible to display the result as a dendrogram, thus enabling the visualisation of higher

level relations between clusters. At the same time, hierarchical methods have a significant disadvantage: they require the distance matrix, which makes them unsuitable for very large datasets.[2] Even then, it is sufficient to limit the number of observations to make the agglomerative methods useful.

The algorithm of the agglomerative variant of hierarchical clustering begins with creating $n$ clusters containing one object each. The next step is finding two most similar objects. This step is equivalent to finding the two most similar clusters as at this stage each cluster contains exactly one object. The most similar objects are defined on the basis of the distance matrix (which can be called the similarity matrix as well, further on, the two terms will be used interchangeably – in spite of some formal nuances, that should not pose risk of misunderstanding). Subsequently, the two clusters closest to each other are merged into the first two-element cluster. In the following steps, the closest clusters continue to be merged until, depending on the algorithm version, 1) only one cluster of all $n$ objects is left or 2) there are $K$ clusters, where $K$ denotes the initially intended number of clusters. It may be unclear how to determine the distance (similarity) between clusters as the distance matrix is equivalent to the matrix of cluster distances only at the very first step. Such uncertainty is well justified – the intercluster distance measure must be defined for the grouping process to complete. The clustering algorithm as well as its result will depend on that definition. There is a variety of ways of determining the distance between clusters (Wierzchoń and Kłopotek 2015, 35; Murtagh and Contreras 2012, 88–89). For the purpose of this research, it is sufficient to mention the two most basic ones, and the third – more complex – used further in the paper.

a) Single linkage (minimum method) defines the distance between clusters as the closest distance between objects belonging to different clusters. Its informal, yet detailed, description can be found in Marek and Noworol (1983, 35–37).

b) Complete linkage (maximum method) defines the distance between clusters as the largest distance between two objects belonging to different clusters.

c) Ward's method defines the cluster closest to a group as the one whose incorporation will result in the lowest increase of variance within that group. In other words, the closest cluster is the one, whose incorporation will optimise an objective function (Ward 1963; more: Murtagh and Legendre 2014).

Having defined the measure of cluster similarity, one can present an informal and simplified algorithm of agglomerative hierarchical clustering:

1. Treat each object as a separate group.
2. Find two groups closest to each other and merge them to obtain one group.
3. Repeat step 2. until all objects belong to the same group.[3]

[2] For example, the distance matrix for the entire data set considered further in the paper (254442 observations) would require 241.2 Gb drive memory.

[3] Alternatively: Repeat step 2 until $K$ clusters are formed.

## 3. EMPIRICAL RESEARCH APPLYING CLUSTER ANALYSIS

In this section, the hypotheses outlined in 2.1. will be verified, and advantages of cluster analysis against using arbitrary divisions (like the division into districts) will be discussed. It must be emphasized that a division into districts is itself a result of some clustering. However, firstly, the criterion of that clustering is unknown and, secondly, it is not objective. Thirdly, the number of clusters, in this case districts, is set in advance.

The cluster analysis will consider objects formalised in such a way that they are described by just two variables: *Longitude* and *Latitude*. That is, clusters will be based on geographical data only. Thus, in this case, a cluster will be such a set of crimes – objects, for which their location diversity is minimised. Clustering will take into account the non-uniform distribution of crime acts across neighbourhoods. Therefore, the clusters formed will depend on the spatial distribution of crime, and must not be understood as based on geographical features of city terrain.

Statistical analyses were performed using the R v. 3.6.3 language (RCore Team 2020), RStudio IDE (RStudio Team 2020) and R packages: *ggplot2* (Wickham 2016), *dplyr* (Wickham et al. 2020), *VIM* (Kowarik and Templ 2016), *readr* (Wickham, Hester and François 2018).

### 3.1. Data

The data was extracted from an open-source (licence CC BY 3.0) database containing 260200 observations (records) of crime in the city of Baltimore.[4] Information stored in the database comes from testimonies of victims (reporting persons). Thus, whenever the word "crime" is used, it actually refers to "a reported event, in which an offence is likely to have been committed."

6 out of 16 variables included in the database were chosen for further analysis:

- **Longitude** – approximated longitude of the incident;
- **Latitude** – approximated latitude of the incident;
- **CrimeDate** – the date of crime according to its report; the newest offence found in the base was committed 6th June 2019, the oldest 1st January 2014;
- **District** – the city district where the incident took place; districts of Baltimore include: "Southwestern" (henceforth SW), "Northwestern" (NW), "Central" (C), "Northern" (N), "Eastern" (E), "Southeastern" (SE), "Southern" (S), "Western" (W) and "Northeastern" (NE);
- **Weapon** – indication of the weapon used: gun, knife, fist, fire or other;

[4] Open Baltimore, Part 1 Crime data (formerly: BPD Part 1 Victim Based Crime Data), https://data.baltimorecity.gov/datasets/part-1-crime-data-3/, accessed 13th February 2021. The study used the June 2019 version of the data.

• **Description** – reported crime type; fourteen kinds of offences are distinguished in the database (following its naming convention): aggravated assault (AN), arson (P), auto theft (KS), burglary (W), common assault (ZN), homicide (M), larceny (K), larceny from auto (WA), rape (G), robbery – carjacking (RS), robbery – commercial (RH), robbery – residence (RM), robbery – street (RU), shooting (S).

### 3.2. Data Pre-Processing; Descriptive Statistics

Some of the records did not contain any information of incident location. Even though methods of clustering for incomplete data are known (Matyja and Simiński 2014), in this case using them would be of no value since this particular cluster analysis considers objects defined only by geographical data. There were also observations referring to offences supposedly committed far beyond the borders of the city of Baltimore.

The size of the corpus used made it possible to disregard portions of the data. Therefore, all 5758 observations with incomplete geographical data as well as those of or (not in Baltimore) were removed. As a result, 254442 observations left in the database.

For 1824 further observations, there was no value of *District*, even though these incidents had complete location data. In such a case, when the location of a given incident was known and so was the District of the closest incident, the given crime's *District* could be correctly identified with nearly 100% accuracy. Thus, the classifier of *k*-nearest neighbours method (Zhang 2016; Jonge and Loo 2013, 48–49) was used to fill the missing values of the *District* variable with $k = 5$.

The data was neither normalised nor rescaled. Although normalisation is critical for most analyses and there are means of making variables comparable (Walesiak 2014; Jarocka 2015), the data considered in this paper do not require this procedure as longitude and latitude can be compared without any normalisation for such a small area.

The resulting dataset makes it possible to determine how often particular types of offences were reported, and if there was any information regarding the weapon used. This information is presented in Table 2 and Table 3.[5] The value of $\text{Percentage}_1$ shows the share of offences committed using a particular weapon among all offences, while $\text{Percentage}_2$ is the share of offences where a particular weapon was used among offences involving the use of any weapon.

Table 2 shows that the crimes reported most often in Baltimore are theft (about 22%), less aggressive assaults (about 16%) as well as burglaries (15%), and larceny from auto. The least commonly reported crimes involve arson, homicide, robbery on a street, carjacking and rape – each of them less often than in 1% of cases. Of course, the fact that some crimes are rarely reported does not necessarily

[5] Unless specified otherwise, precision of the results is three decimal digits, i.e. 0.1%.

mean they are also rarely committed. This is true particularly for rape, which is often hard to report due to victim's psychological distress or a privileged position of the offender. Table 3 reveals that in most cases no use of weapon is reported. Again, this observation does not have to mean that the offender had no weapon, however it is likely the weapon was not used, at least not in the way noticeable for the witnesses. In 19% of cases information about a weapon was provided. If it is present, it is most often a pistol, which is a pattern very much different from what is known in Poland. Using a knife was reported roughly three times less often than a pistol.

**Table 2.** Crimes of various types in Baltimore between 1.01.2014–1.06.2019

| **Type** | **AN** | **P** | **KS** | **W** | **ZN** | **M** | **K** |
|---|---|---|---|---|---|---|---|
| Number | 27170 | 1183 | 22578 | 38056 | 41604 | 1641 | 56637 |
| Percentage [%] | 10.7 | 0.5 | 8.9 | 15 | 16.4 | 0.6 | 22.3 |
| **Type** | **WA** | **G** | **RS** | **RH** | **RM** | **RU** | **S** |
| Number | 33620 | 1592 | 1985 | 4645 | 2610 | 17820 | 3301 |
| Percentage [%] | 13.2 | 0.6 | 0.8 | 1.8 | 1 | 0.7 | 1.3 |

Source: own elaboration.

**Table 3.** Using of weapon in crimes in Baltimore between 1.01.2014–1.06.2019

| **Type** | **Gun** | **Knife** | **Fist** | **Fire** | **Other** | **None/no data** |
|---|---|---|---|---|---|---|
| Number | 24730 | 8751 | 3284 | 1183 | 14871 | 201623 |
| Percentage$_1$ [%] | 9.7 | 3.4 | 1.3 | 0.5 | 5.8 | 79.2 |
| Percentage$_2$ [%] | 46.8 | 16.6 | 6.2 | 2.2 | 28.2 | – |

Source: own elaboration.

### 3.3. Comparison of *k*-means Clustering Results by Districts

As mentioned above, it is impossible to divide such a large set of data by means of hierarchical algorithms using ordinary computer devices. Before just a fragment of the observations is selected for further analyses, it is reasonable to use the *k*-means method, which has not been described in detail yet but which is capable of dividing a data set of virtually any size. This procedure will make it possible to compare an arbitrary division into city districts with an automated design based on mathematical criteria (distance between points).

Figure 1 displays all points representing approximated crime sites in the city of Baltimore. In the upper chart, points are grouped by the value of District variable (i.e. the set division into Baltimore city districts). The lower figure

shows a grouping resulting from clustering by the *k*-means method with nine clusters ($k = 9$). The groups presented in the upper chart will be called "districts," while the groups presented in the lower one will be referred to as "clusters." The clustering was performed on objects characterised only by their geographical coordinates, i.e. *Latitude* and *Longitude*.

It is visible that the resulting clusters are much more coherent than the districts. It is especially clear for the Southern (S) district which has been replaced by the cluster no. 8 which however does not reach that far north and west. The Northeastern district has remarkably decreased its area too – the border of cluster no. 1 crosses the area around (–76.58; 39.34) where there are no offences.

The first hypothesis will shortly be verified using the reduced set of data but already now it can be assessed that there is a solid support for deeming it true. The result of clustering is remarkably different from the division into city districts. This is not surprising as it was certainly not among the objectives of the city of Baltimore while creating the district division to ensure internal coherence of the criminal incident locations set. Detailed information about the distribution of points from the districts across the clusters is presented in Table 4.

All figures presented here constitute an own elaboration.

**Table 4.** Distribution of points in districts within each cluster

| **Cluster** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
|---|---|---|---|---|---|---|---|---|---|
| **C** | 0 | 8178 | 0 | 369 | 0 | 0 | 0 | 0 | 20191 |
| **E** | 0 | 0 | 0 | 0 | 0 | 0 | 7275 | 0 | 14725 |
| **N** | 0 | 6255 | 14549 | 0 | 0 | 4317 | 0 | 0 | 2709 |
| **NE** | 23789 | 0 | 10624 | 0 | 0 | 0 | 3708 | 0 | 212 |
| **NW** | 0 | 1661 | 0 | 0 | 1659 | 21168 | 0 | 0 | 0 |
| **S** | 0 | 0 | 0 | 13148 | 0 | 0 | 66 | 11135 | 5159 |
| **SE** | 0 | 0 | 0 | 0 | 0 | 0 | 27878 | 0 | 9296 |
| **SW** | 0 | 138 | 0 | 7375 | 18404 | 4 | 0 | 0 | 0 |
| **W** | 0 | 12790 | 0 | 6580 | 1080 | 0 | 0 | 0 | 0 |

Source: own elaboration.

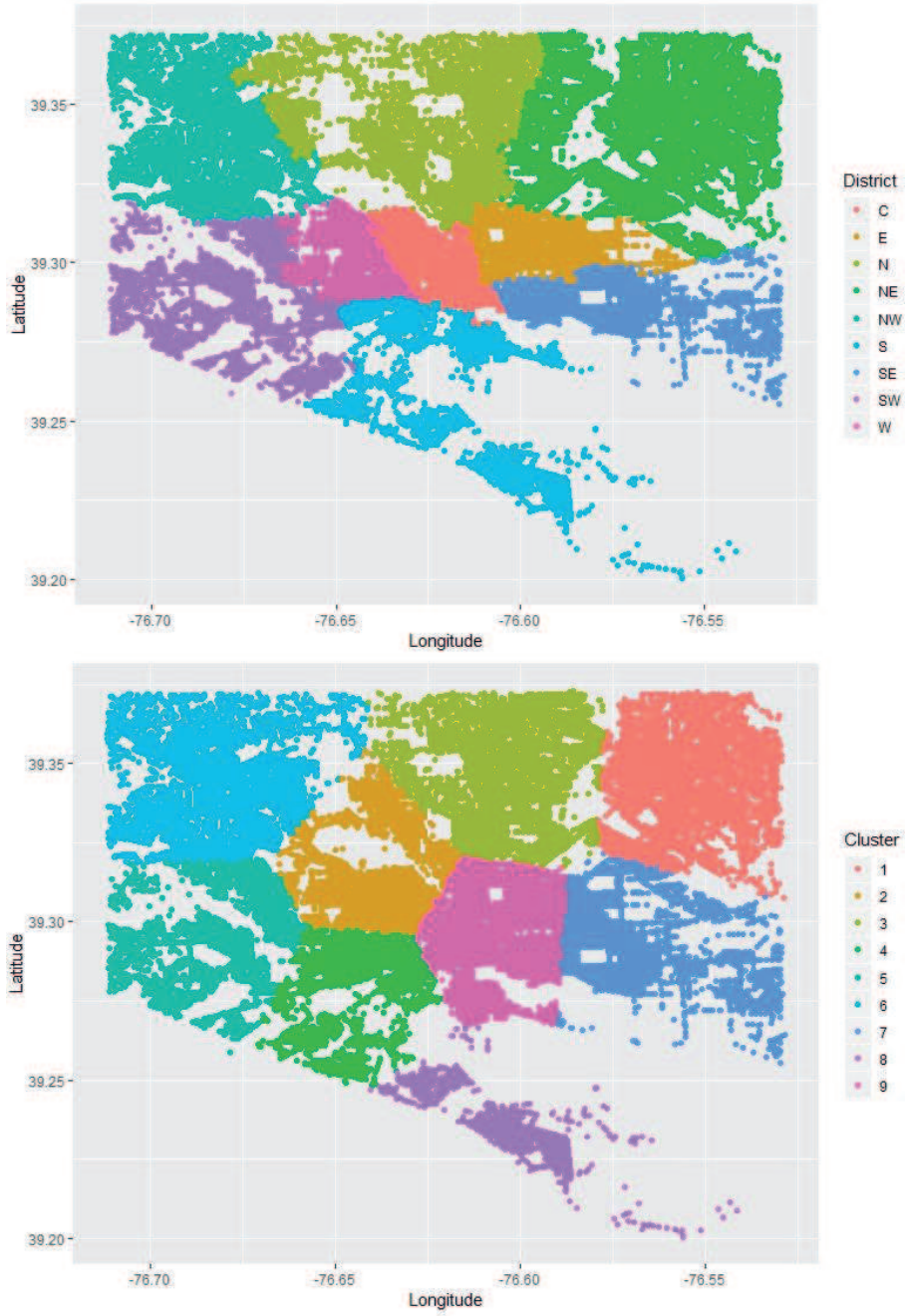


**Figure 1.** Location of crimes displaying data from the period between 1.04.2014–1.06.2019 in Baltimore by districts (upper chart) and clusters obtained with the *k*-means clustering, $k = 9$ (lower chart)

### 3.4. Hierarchical Clustering and Its Result

In order to group the data hierarchically, a subset of 10000 newest observations was extracted. They represent crimes committed between 3rd March 2019 and 6th June 2019. Same as previously, the clustering was performed on variables referring to geographical coordinates only (*Latitude* and *Longitude*). The corresponding Euclidean distance matrix was created, and Ward's clustering algorithm was executed. As a result, a dendrogram presented in Figure 2 was generated. It represents the process of merging smaller clusters into larger ones. The vertical axis presents the distances between clusters being merged.

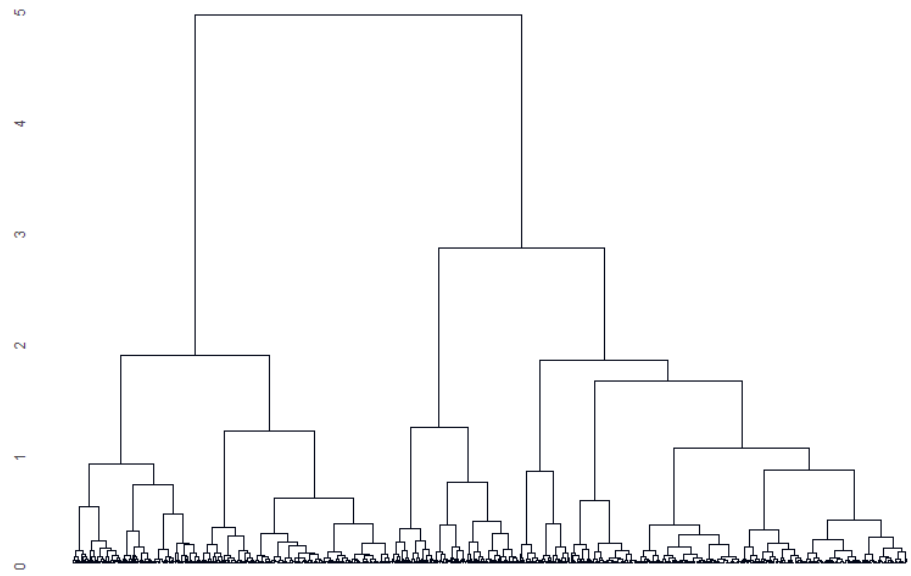


**Figure 2.** Dendrogram of Ward's method clusterization

As can be seen, the observations get merged into larger and larger clusters until a cluster grouping all of them is obtained. It seems reasonable to divide crime incidents in Baltimore into less than nine clusters, on the one hand, and more than that, on the other. This decision will facilitate assessing if the variance of crime patterns can be observed for large-scale city parts as well as for small neighbourhoods, provided that such variance can be observed. To begin with, the division into nine clusters was chosen again so as to verify the first hypothesis in conditions differing from those presented in section 3.2. The result is shown in Figure 3.

What seems especially striking at the first glance, is the decrease in the "density" of the chart when compared to Figure 1. It results from limiting

the number of visualised points, which is about 25 times less than it used to be. Still, the distribution of points in clusters is clearly different from their distribution in the districts. Hence, visual inspection again lends support to the first hypothesis: cluster analysis divides Baltimore into areas different from its official districts.

Moreover, at least a few areas visible in Figure 3 seem naturally separated (for example, the group of points in the southern or easternmost part). Therefore, the set was divided into 14 clusters. To contrast an additional division into three clusters only was created. Its result can be seen in Figure 4, while Figure 5 presents the division with 14 clusters. It can be observed that almost all of 14 clusters are clearly separated from each other and intrinsically homogenous. It would be still reasonable to split some of the clusters, however it is clear that increasing the number of clusters has already increased internal coherence and intuitiveness of the division. On the contrary, a reasonable division cannot be provided with only three clusters, which are clearly too few for such a varied set.

In order to verify the second hypothesis, which postulates a different crime structure in various city areas, the share of particular types of incidents (homicide, auto theft and larceny) in the general number of incidents was examined for the obtained clusters. Subsequently, for the division with $k = 14$ the mean and standard deviation was calculated for the set of the fourteen shares. The Table 5 contains information about the calculations made.[6] The value of the coefficient of variation[7] calculated for the three types of offence exceeds 15%, which suggests that at least when it comes to these kinds of crime, the crime structure in Baltimore is not uniform. In order to check more reliably if this assumption is justified, the test was performed separately for each of three offences considered. The null hypothesis here was a uniform distribution of the percentage of the given crime in clusters. Another test was performed for the three offence types combined. The null hypothesis of the test is that crime structure is independent of a district. Test results are presented in Table 6. Each of them provides very strong evidence for rejecting the null hypothesis ($p < 0.01$), which strongly suggests confirming crime structure variance for the offences considered. This leads to confirming the second hypothesis that the crime structure is not the same in various areas of Baltimore.

[6] Because of small order of magnitude of some shares, the results are presented as percentages with 0.01% precision, i.e with four decimal digits.

[7] Defined as the standard deviation divided by the mean: $\frac{\sigma}{\bar{x}}$.

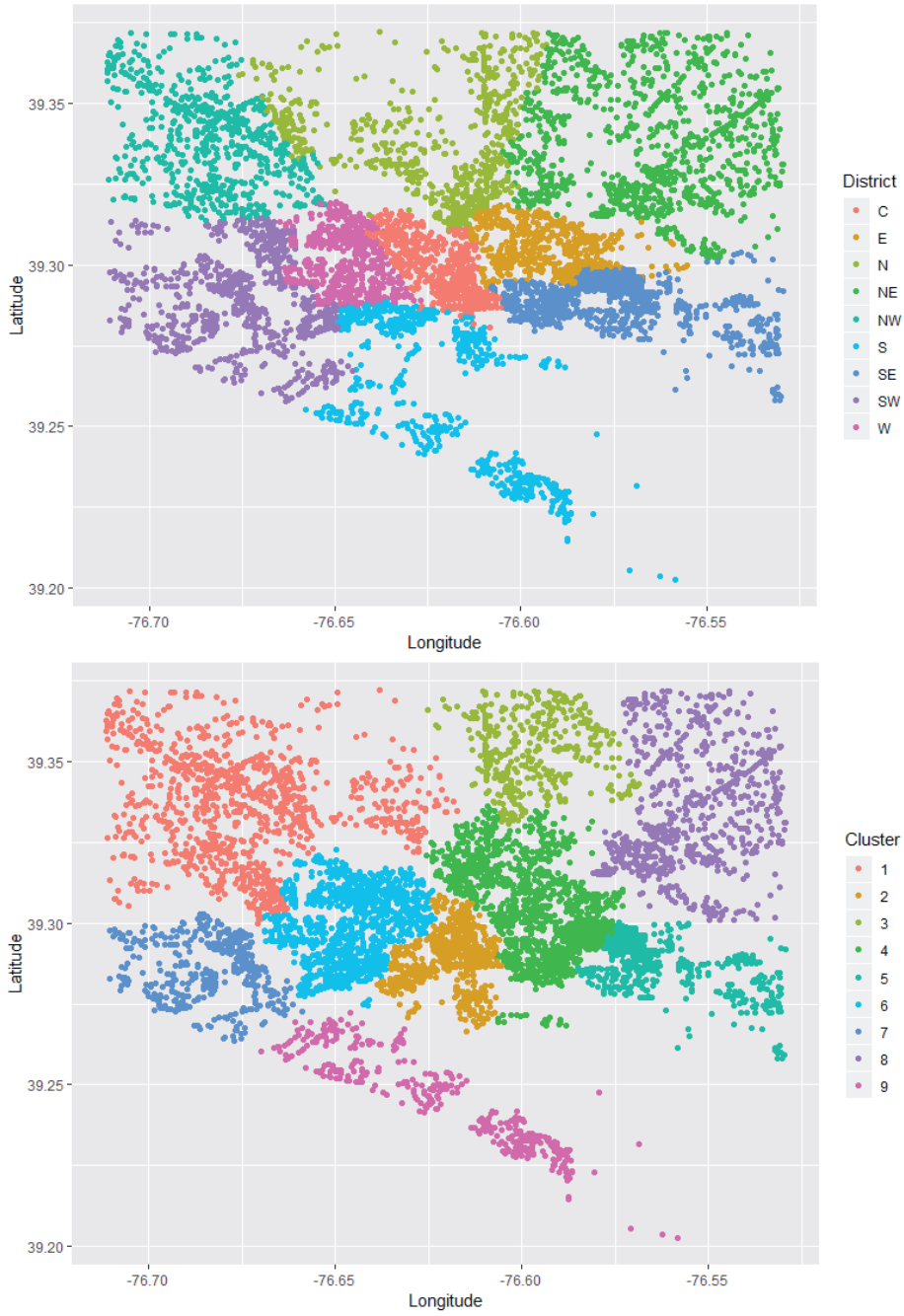


**Figure 3.** Location of crimes displaying data from the period between 3.03.2019–1.06.2019 in Baltimore by districts (upper chart) and clusters obtained with the Ward's method, $k = 9$ (lower chart)

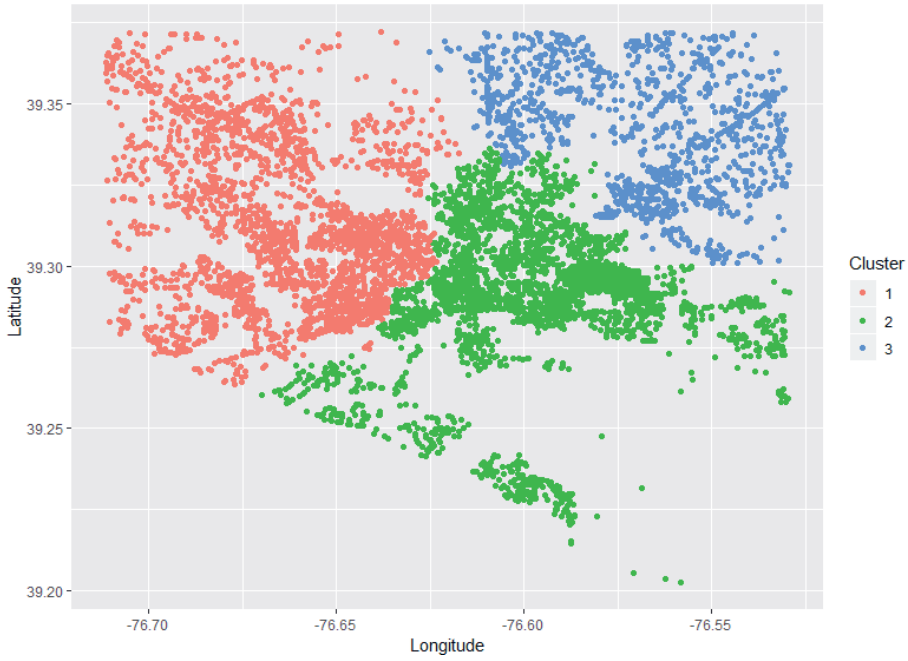


**Figure 4.** Plot for clustering obtained with the Ward's method, $k = 3$

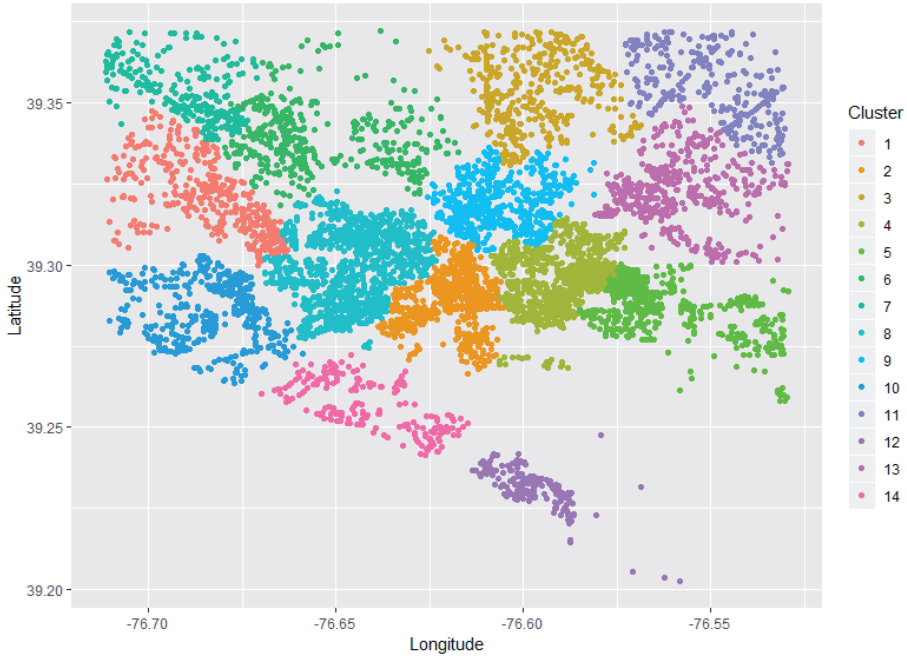


**Figure 5.** Plot for clustering obtained with the Ward's method, $k = 14$

**Table 5.** The percentage of selected crimes in particular clusters, $k = 3$, $k = 9$ and $k = 14$

| Cluster | Percentage M [%] | Percentage KS [%] | Percentage K [%] |
|---|---|---|---|
| Number of clusters to divide into: *k = 3* | | | |
| 1 | 1.2 | 9.21 | 22.2 |
| 2 | 0.52 | 7.34 | 25.6 |
| 3 | 0.45 | 9.74 | 21.6 |
| *k = 9* | | | |
| 1 | 0.78 | 9 | 24.4 |
| 2 | 0.16 | 4.47 | 33.2 |
| 3 | 0.38 | 9.28 | 20.5 |
| 4 | 0.86 | 7.94 | 21.4 |
| 5 | 0.12 | 8.12 | 25.8 |
| 6 | 1.65 | 9.68 | 21 |
| 7 | 1.09 | 8.54 | 19.9 |
| 8 | 0.89 | 9.98 | 22.2 |
| 9 | 0.67 | 10.2 | 23.2 |
| *k = 14* | | | |
| 1 | 0.86 | 7.2 | 20.6 |
| 2 | 0.16 | 4.47 | 33.2 |
| 3 | 0.38 | 9.28 | 20.5 |
| 4 | 0.82 | 8.08 | 18.6 |
| 5 | 0.12 | 8.12 | 25.8 |
| 6 | 0.51 | 9.48 | 29.4 |
| 7 | 1.08 | 11.1 | 22.2 |
| 8 | 1.65 | 9.68 | 21 |
| 9 | 0.92 | 7.72 | 25.8 |
| 10 | 1.09 | 8.54 | 19.9 |
| 11 | 0.28 | 9.89 | 22.3 |
| 12 | 1.2 | 8 | 22.8 |
| 13 | 0.61 | 10 | 22.2 |
| 14 | 0.29 | 11.8 | 23.6 |
| **Mean** | **0.71** | **8.81** | **23.41** |
| **Standard deviation** | **0,44** | **1,74** | **3,83** |

Source: own elaboration.

**Table 6.** Results of uniformity tests of crime structure in different clusters of Baltimore

| Considered variables | Number of degrees of freedom | $\chi^2$ value | *P*-value |
|---|---|---|---|
| M | 13 | 33.447 | 0.001 |
| KS | 13 | 112.82 | < 0.001 |
| K | 13 | 43.227 | < 0.001 |
| M; KS; K | 39 | 176.62 | < 0.001 |

Source: own elaboration.

### 3.5. Variance of Crime Structure by Number of Clusters

The third hypothesis to be examined referred to changes in variance of crime among clusters depending on the number of clusters. For a few reasons, it is worthwhile to consider dividing the area into subareas according to objective criteria.

Firstly, designing social research and testing the effectiveness of social campaigns or suggested environment changes is typically constrained by limited funding. Thus, such initiatives can rarely be performed or introduced on a large area. Even when disregarding criteria of administrative divisions, the created areas may be too large, making the implementation of projects impossible due to the high costs involved. Cluster analysis suggests an objective way of delimiting smaller and more homogenous areas based on the spatial distribution of crime incidents.

Secondly, sometimes it is especially vital to monitor some variables with high resolution. For instance, research examining how the presence of police affects the crime rate in a particular area would definitely benefit from using sub-district level data.

Thirdly, operating with data referring to entities of administrative division may effectively disable the creation of high quality econometric models unless the number of such entities is high enough. Moreover, high variance of variable values is sometimes wanted. For example, when modelling the relations between a certain type of the crime rate and some local factors, it is beneficial if the variables have high level of variance. Cluster analysis makes it possible to handle the first of the above mentioned problems by splitting a dataset into smaller parts, and thus obtaining several values of the variable. But can this method increase the variance of the variable value? Verifying the third hypothesis will bring an answer to this question.

In order to verify this issue, a function dividing the previously selected set into $k$ clusters was created using the Ward's method, with $k$ ranging from

3 to 40. Next, standard deviation (square root of variance)[8] was calculated for these samples, accounting for the following crimes: homicide, larceny and auto theft. The standard deviation for the sample representing the Baltimore division into districts was used as a reference. The results obtained as well as the regression line (estimated with the ordinary least squares method) are presented in Figures 6–8. The black horizontal line represents the standard deviation obtained for the Baltimore district division.

Even though the standard deviation does not increase monotonically with the increase of the number of clusters, the correlation is evidently positive. For each clustering, such $k$ can be determined that the variance of a set of values for this $k$ is higher than for Baltimore districts. In spite of a local variance decrease, the coefficient in each linear regression model with k as an independent variable and standard deviation as a dependent variable was significant ($p < 0.001$) and positive. There was no rationale to reject the null hypothesis of normality of the distribution of model residuals ($p$-value of Shapiro-Wilk tests > 0.1). Moreover, the coefficient of determination in the models ranges from about 0.77 to 0.92, suggesting that the number of clusters obtained constitutes the main variable responsible for the increase in crime structure variance.

The third hypotheses should then be considered true. It means that this clustering method has a greater potential of explaining the influence of various factors on crime in cities than arbitrary administrative divisions into districts. Even though this relation has been proved neither for all kinds of offences, nor for cities other than Baltimore, there is no reason to assume that it cannot be at least partly generalised.

Of course, the findings of this section do not imply that it is always reasonable to maximise the number of clusters. Increasing their number yields positive results as long as it can be justified by the aim of the research undertaken or the structure of the dataset. If an adequate number of clusters cannot be determined based on the above mentioned criteria, analytical criteria can be helpful (Milligan 1985; Jung et al. 2003).

[8] An ordinary descriptive statistics of standard deviation (a biased estimator of standard deviation) was calculated, given by: $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}}$.

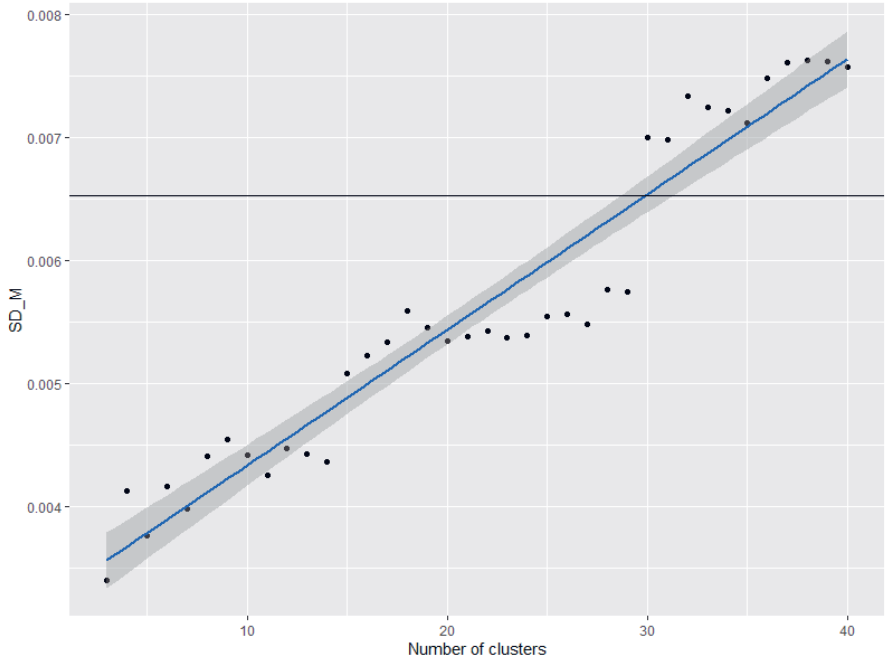


**Figure 6.** Standard deviation of homicide percentage by number of clusters

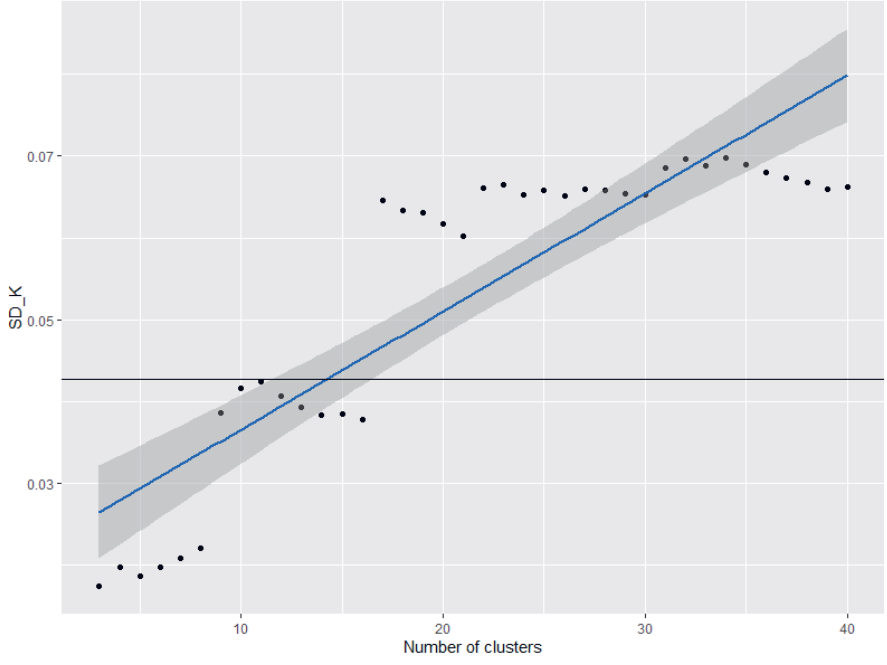


**Figure 7.** Standard deviation of larceny percentage by number of clusters

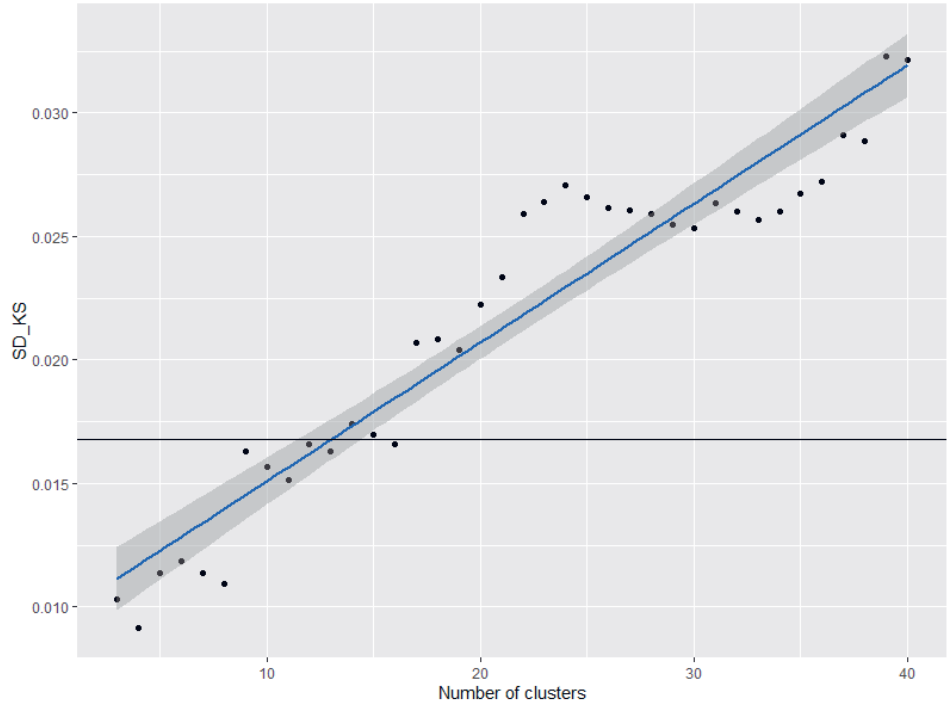


**Figure 8.** Standard deviation of auto theft percentage by number of clusters

## 4. CONCLUSIONS

The goal of the paper was to present hierarchical cluster analysis methods and to apply them to the clustering set of 10000 observations concerning crime in Baltimore. Moreover, the entire dataset was divided into groups using the *k*-means method. The results obtained have revealed that cluster analysis leads to the division of the city different from the administrative boundaries. It was also demonstrated that spatial factors correlate with the crime structure in Baltimore, which is different in various areas. Finally, it was shown that the variance of percentage of crime types rises with an increase in the number of clusters.

It seems that the cluster analysis method can be extremely useful in criminological and sociological research concerning the relation between crime and space. Not only does it ensure an objective criterion for division, but also allows for adjusting the number of areas and takes into account the spatial distribution of crime. Broader use of clustering in the research concerning the spatial dimension of law, including breaking it, is recommended. In fact, only a small part of the applications of the method were discussed in the present paper.

## BIBLIOGRAPHY

Conte, Rosaria, Nigel Gilbert, Giulia Bonelli, Claudio Cioffi-Revilla, Guillaume Deffuant, Janos Kertesz, Vittorio Loreto, Suzy Moat, Jean-Pierre Nadal, Anxo Sanchez, Andrzej Nowak, Andreas Flache, Maxi San Miguel and Dirk Helbing. 2012. "Manifesto of Computational Social Science." *The European Physical Journal Special Topics* 214(1): 325–346.

Dudek, Michał, Piotr Eckhardt and Marcin Wróbel, Eds. 2018. *Przestrzenny wymiar prawa* [*Spatial Dimension of Law*]. Kraków: NOMOS.

Gareth, James, Daniela Witten, Trevor Hastie and Robert Tibshirani. 2017. *An Introduction to Statistical Learning with Applications in R.* New York: Springer.

Glyde, John. 1856. "Localities of Crime in Suffolk." *Journal of the Statistical Society of London* 19(2): 102–106.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning. Data Mining, Inference and Prediction Second Edition.* New York: Springer.

Jarocka, Marta. 2015. "Wybór formuły normalizacyjnej w analizie porównawczej obiektów wielocechowych" ["The Choice of a Formula of the Data Normalization in the Comparative Analysis of Multivariate Objects"]. *Ekonomia i Zarządzanie* 1: 113–126.

Jonge, Erwin de and Mark van der Loo. 2013. *An Introduction to Data Cleaning with R*. The Hague: Statistics Netherlands.

Jung, Yunjae, Haesun Park, Ding-Zhu Du and Barry Drake. 2003. "A Decision Criterion for the Optimal Number of Clusters in Hierarchical Clustering." *Journal of Global Optimization* 25(1): 91–111.

Kądziołka, Kinga. 2016a. "Determinanty przestępczości w Polsce. Analiza zależności z wykorzystaniem drzew regresyjnych" ["Determinants of Crime Rate in Poland. Analysis using Regression Trees"]. *Ekonomia. Rynek, Gospodarka, Społeczeństwo* 45: 53–81.

Kądziołka, Kinga. 2016b. "Przestrzenne zróżnicowanie zagrożenia przestępczością w Polsce" ["Spatial Diversity of Crime Rate in Poland"]. *De Securitate et Defensione. O Bezpieczeństwie i Obronności* 2: 31–43.

Krzyśko, Mirosław, Waldemar Wołyński, Tomasz Górecki and Michał Skorzybut. 2008. *Systemy uczące się. Rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości* [*Machine Learning Systems. Pattern Recognition, Cluster Analysis and Dimensionality Reduction*]. Warszawa: Wydawnictwo Naukowo-Techniczne.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy and Marshall Van Alstyne. 2009. "Computational Social Science." *Science* 323(5915): 721–723.

Marek, Tadeusz and Czesław Noworol. 1983. *Wprowadzenie do analizy skupień* [*Introduction to Cluster Analysis*]. Kraków: Uniwersytet Jagielloński.

Matyja, Artur and Krzysztof Simiński. 2014. "Comparison of Algorithms for Clustering Incomplete Data." *Foundations of Computing and Decisions Sciences* 39(2): 107–127.

Milligan, Glenn and Martha Cooper. 1985. "An Examination of Procedures for Determining the Number of Clusters in a Data Set." *Psychometrika* 50(2): 159–179.

Mordwa, Stanisław. 2016. "The Geography of Crime in Poland and Its Interrelationship with Other Fields of Study." *Geographia Polonica* 89(2): 187–202.

Murtagh, Fionn and Pedro Contreras. 2012. "Algorithms for Hierarchical Clustering: An Overview." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2: 86–97.

Murtagh, Fionn and Pierre Legendre. 2014. "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?" *Journal of Classification* 31(3): 274–295.

Walesiak, Marek. 2002. “Pomiar podobieństwa obiektów w świetle skal pomiaru i wag zmiennych” [“Similarity Measures from the Point of View Scales of Measurement and Variables Weights”]. *Prace Naukowe Akademii Ekonomicznej we Wrocławiu* 950: 71–85.

Walesiak, Marek. 2014. “Przegląd formuł normalizacji wartości zmiennych oraz ich własności w statystycznej analizie wielowymiarowej.” [“Data Normalization in Multivariate Data Analysis. An Overview and Properties”] *Przegląd Statystyczny* 61(4): 363–372.

Ward, Joe. 1963. “Hierarchical Grouping to Optimize an Objective Function.” *Journal of the American Statistical Association* 58(301): 236–244.

Wierzchoń, Sławomir and Mieczysław Kłopotek. 2015. *Algorytmy analizy skupień* [*Algorithms for Cluster Analysis*]. Warszawa: Wydawnictwo WNT.

Wortley, Richard and Michael Townsley. 2016. “Environmental Criminology and Crime Analysis: Situating the Theory, Analytic Approach and Application.” In *Environmental Criminology and Crime Analysis*, 2nd ed. 1–26. Edited by Richard Wortley and Michael Townsley. New York: Routledge.

Zhang, Zhongheng. 2016. “Introduction to Machine Learning: K-nearest Neighbors.” *Annals of Translational Medicine* 4(11): 218: 1–7.

**Software used in analysis**

Kowarik, Alexander and Matthias Templ. 2016. “Imputation with the R Package VIM.” *Journal of Statistical Software* 74(7): 1–16.

RCore Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. Accessed 13th February 2021: https://www.R-project.org/

RStudio Team. 2020. *RStudio: Integrated Development for R*. RStudio. PBC. Boston, MA. Accessed 13th February 2021: http://www.rstudio.com/

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. New York: SpringerVerlag. Accessed 13th February 2021: https://ggplot2.tidyverse.org/

Wickham, Hadley, Jim Hester and Romain François. 2018. *readr: Read Rectangular Text Data. R package version 1.3.1*. Accessed 13th February 2021: https://CRAN.R-project.org/package=readr/

Wickham, Hadley, Romain François, Lionel Henry and Kirill Müller. 2020. *dplyr: A Grammar of Data Manipulation. R package version 0.8.5*. Accessed 13th February 2021: https://CRAN.R-project.org/package=dplyr/