Development of cluster analysis methodology for identification of model rainfall hyetographs and its application at an urban precipitation field scale

Karol Mikołajewski[1,2], Marek Ruman[2,*], Klaudia Kosek[3,*], Marcin Glixelli[4], Paulina Dzimińska[5], Piotr Ziętara[4], Paweł Licznar[5]

[1]RETENCJAPL Sp. z o.o., Gdańsk 80-868, Poland; karol.mikolajewski@retencja.pl

[2]Faculty of Natural Sciences, University of Silesia in Katowice, Sosnowiec 41-200, Poland; marek.ruman@us.edu.pl

[3]Faculty of Civil and Environmental Engineering, Gdansk University of Technology, Gdańsk 80-233, Poland; klaudia.kosek@pg.edu.pl

[4]Krakow Water, Kraków 30-106, Poland; marcin.glixelli@wodociagi.krakow.pl (M.G.), piotr.zietara@wodociagi.krakow.pl (P.Z.)

[5]Faculty of Environmental Engineering, Wroclaw University of Science and Technology, Wrocław 50-377, Poland; dziminska.paulina@gmail.com (P.D.), pawel.licznar@pwr.edu.pl (P.L)

[*]Corresponding authors: marek.ruman@us.edu.pl (M.R.); klaudia.kosek@pg.edu.pl (K.K.)

Abstract: Despite growing access to precipitation time series records at a high temporal scale, in hydrology, and particularly urban hydrology, engineers still design and model drainage systems using scenarios of rainfall temporal distributions predefined by means of model hyetographs. This creates the need for the availability of credible statistical methods for the development and verification of already locally applied model hyetographs. The methodology development for identification of similar rainfall models is also important from the point of view of systems controlling stormwater runoff structure in real time, particularly those based on artificial intelligence. This paper presents a complete methodology of division of storm rainfalls sets into rainfalls clusters with similar temporal distributions, allowing for the final identification of local model hyetographs clusters. The methodology is based on cluster analysis, including the hierarchical agglomeration method and k-means clustering. The innovativeness of the postulated methodology involves: the objectivization of clusters determination number based on the analysis of total within sum of squares (wss) and the Caliński and Harabasz Index (CHIndex), verification of the internal coherence and external isolation of clusters based on the bootmean parameter, and the designated clusters profiling. The methodology is demonstrated at a scale of a large urban precipitation field of Kraków city on a total set of 1806 storm rainfalls from 25 rain gauges. The obtained results confirm the usefulness and repeatability of the developed methodology regarding storm rainfall clusters division, and identification of model hyetographs in particular clusters, at a scale of an entire city. The applied methodology can be successfully transferred on a global scale and applied in large urban agglomerations around the world.

Keywords: precipitation modelling, storm rainfalls, cluster analysis, classification quality assessment indices

1. Introduction

Conducting calculations of non-stationary surface runoff of stormwaters and their further transformation in the stormwater system, including detention and retention in water bodies, requires the scenarios availability of rainfall temporal distribution. Precipitation scenarios employed by hydrological models usually determine the values of outflow volumes observed in the calculation results, the dynamics of their changes in time, and changes in the water volume retained in various elements of the drainage system. Due to historical conditions, and particularly calculation limitations, for decades, only the simplest models naturally employing the simplest rainfall scenarios could be implemented in practice in hydrology at its various scales, and particularly in urban hydrology. An extreme example can be the rational model, still commonly used today in designing urban and road drainage systems. It employs block rainfall, i.e., a scenario of rainfall with a specified duration and constant intensity determined based on the local IDF (Intensity-Duration-Frequency) model for the adopted frequency of its occurrence (Kundzewicz and Licznar, 2021(2022)). The block model is obviously a very specific rainfall scenario. Unlike in that model, as observed for natural rainfall, hydrological analyses usually assume the necessity of taking into consideration the temporal variability of point rainfall intensity, usually described in the graphic form by means of hyetographs. Adopting hyetographs in design aimed at reflecting rainfall temporal distributions analogical to the temporal courses of locally recorded rainfalls, and implicitly also potential rainfalls that may occur in the future. This has encountered numerous obstacles since the very beginning, resulting not only from the stochastic (random) nature of rainfall, but also from the multifractal nature of rainfall still unknown at the time. It is currently already evidenced that even simple time series of rainfall records from single rain gauges have a multifractal structure escaping the principles of simple Euclidean geometry (de Lima, 1998; Deidda et al., 1999; Licznar, 2009). Due to this, and due to the limited availability of research material (sets of precise rainfall records at a high temporal resolution), the historical methodology of hyetographs representative development for modelling, customarily called model hyetographs, had to be based on generalisation and simplification, and assumptions that should be currently rejected or subject to suitable validation. All this creates the need for undertaking new research in the scope of identification of model hyetographs.

Interest in model hyetographs certainly increased with first attempts of transition from stationary methods of calculation of urban stormwater systems towards non-stationary methods of simulation of their operation during stormwater runoff. It became possible due to the implementation of digital hydrodynamic models of sewage networks, such as e.g., programme SWMM (Storm Water Management Model) (Nix, 1994). Initially, however, due to the limited computing power of the available computers, focus was only placed on flows simulations in the underground canal system for single rainfalls. The model of stormwater runoff transformation in the sewage system was only combined with the hydrological surface runoff model, describing stormwaters inflow to network nodes (manholes and inlets*)* from the catchment area, and the stormwater system effect. The interaction of the stormwater system with the rainfall receiver was only possible to reflect by means of a suitable threshold condition on the network outlet. Rainfall introduced to the first generation of stormwater systems hydrodynamic models was a hyetograph of either a subjectively selected actual storm rainfall recorded by a pluviograph, or an artificial model hyetograph (often called model rainfall).

With rapid improvement of the computing power of computers, but without excessive complication of the algorithms of hydrodynamic models, it became possible to conduct simulations for entire series of local storm rainfalls. Their use enabled actual implementation of long-term simulations. After statistical processing, their final results could provide the basis for the probabilistic verification of the drainage system functioning in terms of stormwater system damming up frequency (Gires et al., 2012, 2013; Licznar et al., 2008), as well as e.g., the necessary volume of retention reservoirs of stormwaters (Licznar, 2013). This new reality seemingly appeared to bring an end to the application of model hyetographs, at least in the area of urban hydrology. Moreover, with the installation of urban precipitation monitoring networks equipped with new generation electronic rain gauges, high temporal resolution rainfall series became much more available. In simulations of large stormwater systems, particularly in terms of development of their Real Time Control, records of rainfalls from entire networks of rain gauges began to be applied, or even spatial data from weather radars (Jakubiak et al., 2014). Afterwards at operational phase of RTC systems above mentioned observational rainfall data sources are often coupled with numerical weather predictions. It allows proper short-term prediction of extreme rainfall events, indispensable for urban flood alert broadcasting and selection of best scenarios of next RTC strategies. Here, there is a vivid need for effective whether pattern recognition algorithms, based on machine learning technique, as for example support vector machine implemented by Nayak and Ghosh (2013) in case of Mumbai, India. Finally, in response to the stochastic nature of precipitation processes, it was determined that it is the most credible to supply hydrodynamic models with not so much historical precipitation data (i.e., implementations of local precipitation processes that already took place) as much richer synthetic data generated by means of local precipitation generators (Molnar et al., 2006). Owing to progress in the scope of multiplicative random cascades, it became possible to develop generators not only for generating rainfall series (1-D) (Güntner et al., 2001; Hingray and Ben Haha, 2005; Licznar et al., 2011a, 2011b), but also spatial data (2-D) (Over and Gupta, 1994; Rupp et al., 2012), or even spatiotemporal data on rainfall (3-D) (Deidda, 2000).

Despite all the aforementioned conditions, the end of application of model hyetographs in urban hydrology has not arrived yet. To a certain extent this probably results from the conservatism of the engineers themselves, preferring design and drainage systems modelling based on simple and long-familiar model hyetographs. This facilitates their work e.g., in terms of time and equipment requirements necessary for conducting hydrodynamic simulations, and in accordance with the conservative provisions of rarely modified technical standards (Schmitt, 2000). Paradoxically, however, the application of model hyetographs is currently not limited to simple engineering works. They prove useful in the most advanced consulting works. Such works employ a completely qualitatively new approach to modelling stormwater networks, called integrated modelling. The software used for the purpose is much more advanced, because it combines three models: the hydrological streams model (receivers of stormwaters), the hydrodynamic stormwater network model, and the hydrological surface runoff model. The first two models are 1-dimentional, whereas the surface runoff model is a 2-dimentional model with an additional fill parameter. Naturally, the latter must be coupled with the digital terrain model (DTM). Although the application of integrated models brings numerous benefits, e.g., in the form of the possibility of tracing floodings on the surface of the DTM, it requires the use of complicated software launched on equipment with high computing power. Also in this case, simulations of larger drainage systems prove time-consuming and prone to numerical instabilities. As a result, simulation of runoff from the drainage system ceases to be feasible for tens or even hundreds of scenarios of storm rainfalls. In practice, complicated simulations on an integrated model can be only conducted for certain characteristic precipitation. This explains the return to the concept of the model hyetograph application or a narrow group of model hyetographs in the case of time-consuming simulations. It therefore remains an important issue to improve the methodology that could efficiently and objectively determine what model hyetographs would reflect local rainfall distributions variable in time in a satisfactory way.

In the case of the methodology modernisation of the model hyetographs development, it should be remembered that it began forming in the situation of strongly limited access to high temporal resolution rainfall records and statistical tools supporting their processes. Rainfall was recorded by means of simple rain gauges, e.g., pluviographs, and the records in the analogue form were difficult to process. Their processing usually involved a review of the records with designation of maximum rainfall with different durations and structure, and determination of the resulting empirical DDF (Depth-Duration-Frequency) or IDF (Intensity-Duration-Frequency) dependencies. Due to this, it was proposed to use the information contained in DDF and IDF dependencies as the starting point for constructing model hyetographs. Next to the rectangular shape of block rainfall mentioned in the introduction, different authors recommended simple modifications of the shape of hyetographs. For example, Sifald (1973), analysing the hydraulic operation of the stormwater network, postulated a hyetograph with a trapeze shape. At the scale of small catchments, Yen and Chow (1980) proposed transformation of rainfall depth read for the predefined duration and frequency from the DDF model into a triangular hyetograph. Probably due to the common use of the unit hydrograph method in the contemporary hydrology aimed at providing an estimate of direct runoff hydrographs resulting from given excess rainfall hyetograph, hyetographs with possibly simple shapes were eagerly used, composed of regular geometric figures. For example, in urban hydrology, Desbordes (1978) implemented the application of a hyetograph composed of "three" triangles. Peyron et al. (2002) attempted to simplify the shape through replacing two triangles representing the start and end impulses with rectangular courses. At the same time, in the conditions of Taiwan, Lee and Ho (2002) postulated the application of a model hyetograph with a shape built from two triangles. Even earlier, independently from the aforementioned papers, a concept appeared for the structure of the model hyetograph to be completely based on the reading of the entire IDF curve. The most classic example of the approach can be the continuous hyetograph for the city of Chicago developed by Keifer and Chu (1957). Another representative of this concept can be Euler type II model hyetograph (Schmitt, 2000), frequently encountered in the hydrodynamic modelling practice in Poland and Germany. The underlying idea of this type of hyetographs structures assumes a single artificial rainfall scenario including maximum point intensities for the entire hierarchy of partial times. It appears to be at variance with the observation of nature, where all maximums of point intensities are usually not recorded in a single rainfall at a specified level of frequency for durations lower and equal to total duration. Nonetheless, relying on the Euler type II hyetograph theoretically offers engineers the possibility to test the operation of the drainage system in a single simulation for all maximum point intensities simultaneously at a specified level of frequency of occurrence. Engineers are accustomed to using IDF curves, and their transformation into a Euler type II model hyetograph requires only simple algebra operations. In the case of analyses of larger drainage systems, it is even simpler for engineers in Poland and Germany to reach for a hyetograph recommended in the guidelines of DVWK (1984). Its development is based on the distribution of rainfall depth read from the DDF model to three rectangles with different intensity levels.

With time, the methodology of model hyetographs determination could be improved owing to the availability of increasingly richer measurement data from rain gauge networks. It became feasible to determine model hyetographs based on the large rainfall datasets analysis. The precursor of such an approach was Huff (1967), who analysed data from 49 rain gauges and 12 years (1955-1966) from the state of Illinois in the USA. Because the rainfalls retrieved from the records had different total durations and depths, he proposed normalisation of hyetographs, and application of dimensionless hyetographs. This permitted comparison of a dataset of 261 rainfalls that Huff (1967) classified according to whether the greatest percentage of cumulative rainfall occurred in the first, second, third, or fourth quarter of the storm duration. The resulting model Huff mass curves found numerous applications in hydrology. The analogical methodology of temporal variability analysis of rainfall distributions was not only repeated in further research by Huff himself (1970, 1990), but also by many other scientists, e.g., Pani and Haragan (1981), Bonta and Rao (1987), Bonta (2004), Terranova and Iaquinta (2011), Elfeki et al. (2014), Pan et al. (2017). Hydrologists in the USA also commonly apply model hyetographs defined by means of dimensionless mass curves recommended by the SCS (Soil Conservation Service) (McCuen, 1986). The increasingly rapidly growing digital databases of rainfalls records from electronic rain gauges, and the progress in the scope of data mining techniques nowadays allows for the continuation of the trend determined by Huff (1967), and model hyetographs determination based on the actual local rainfall analysis.

Access to a large data base from 25 electronic weight rain gauges from the municipal precipitation monitoring network of Warsaw (Poland) has become the impulse for the verification whether the Euler type II model hyetograph applied in practice, developed based on the IDF curve, corresponds with temporal distributions of actual rainfalls. For this purpose, out of approximately 20-year precipitation series, Licznar and Szeląg (2014) desinated a total of 669 storm rainfalls. The set was then divided into subsets with increasing durations, expressed in minutes: [0-45], (45-60], (60-90], (90-120], (120-180], (180-240], (240-300], (300-360], (360-420]. Each of the subsets was moreover supplemented by an additional Euler precipitation model (type II) developed based on the local IDF curve. All subsets were then analysed with the application of the hierarchical agglomeration method. Based on the obtained dendrograms, Licznar and Szeląg (2014) observed that precipitation recorded in Warsaw, even those with approximate durations, have evidently differing hyetographs. Moreover, in each of the subsets, Euler model precipitation (type II) was an extreme outlier in the structure of the dendrogram with the highest bond distance towards all actual rainfalls. Results of the study have become an impulse for the search for a method of designation of more representative local model hyetographs. For this purpose, Licznar et al. (2017) proposed the application of cluster analysis tools in the form of not only the hierarchical agglomeration method, but also non-hierarchical k-means clustering. The methodology also found applications for example in research by Licznar (2018) and Wartalska et al. (2020). Nonetheless, it leaves evident gaps. Its greatest weakness is completely subjective a priori adoption of number *k* of clusters, i.e., the number of final rainfall models. Two further missing components of the methodology include the objective assessment of the classification results and cluster profiling. They are increasingly important, because in the context of new challenges in urban hydrology, striving for ‘smart city’ solutions involves not so much searching for model hyetographs themselves as the possibility of fast and efficient search of similar temporal models of rainfalls. Runoff control systems employing artificial intelligence aim at the implementation of the most effective strategy of the forecasted scenario control of stormwater runoff, adopted based on the already implemented in nature and recorded precipitation phenomenon and runoff caused by rainfall with possibly similar rainfall temporal distribution in reference to local rainfall forecast from the nowcasting system. The final question to be answered from the engineering practice point of view is: to what extent the application of the cluster analysis brings results repeatable at the natural spatial scale of extensive municipal drainage systems? Can for example model hyetographs developed based on rainfall series from nearby rain gauges be treated as credible for analyses of precipitation-runoff phenomena in the territory of the entire city?

Considering the overview of the methodology state of rainfall classification in terms of its temporal distribution and model hyetographs identification, the primary objective of this paper is to present the complete cluster analysis methodology supplemented by objective determination of the number of classes, credible assessment of the classification results, and cluster profiling. Another primary objective of the paper is to demonstrate the postulated complete cluster analysis methodology at the scale of a large urban precipitation field, and the resulting answer to the question whether model hyetographs retrieved from records of different rain gauges show mutual compatibility.

The pragmatic objective of the study is the complete methodology development of designation of local model hyetographs throughout Poland. In the years 2016-2020, the project of the Polish Atlas of Rains Intensities PANDa was implemented, resulting in a digital base of rainfalls at high temporal resolution (after 30 years of observation from 100 stations), and the national rainfall atlas composed of 12885 local IDF models ascribed to areas designated through the division of Poland with a grid with field dimensions of 5 km per 5 km (Burszta et al.; 2019; Licznar et al., 2020). The PANDa atlas provides the basis for the publicly accessible digital design platform www.waterfolder.com, where you can design and select among others: retention reservoirs of stormwaters, infiltration reservoirs for stormwaters, linear drainage units, gravitational canals of stormwater systems, stormwater pumps and pumping stations, and green roof surfaces. In the scope of the new WaterFolder Connect project, works are currently undertaken aimed at the integration of tools of selection and enabling hydrodynamic simulations of newly designed drainage systems. This will require the designation of local hyetographs for particular areas of Poland. The developed methodology presented in this article is planned to be applied at the scale of the entire country, and then globally.

2. Materials and methods

2.1. Study area

Research on objective classification of storm rainfall hyetographs by means of classification quality assessment indices was conducted at a large scale of an urban precipitation field in the territory of Kraków. Kraków is the second largest city in Poland in terms of population (~ 767 000) and surface area (327 km²), located in the south of the country on the Vistula River. The study employed part of the resources of the Polish precipitation data base of the Polish Atlas of Rains Intensities (PANDa) project, and records from the local rain gauge network of the Municipal Water Supply and Sewerage Company (MWSSC) in Kraków.

At the initial stage of the PANDa project in the years 2016-2017, a digital base of rainfall series was developed for a total number of 100 rain gauges in Poland. Then it was analysed in terms of occurrence of maximum rainfall intensity (Burszta-Adamiak et al., 2019). The base included among others records from a rain gauge installed in the area of the Kraków–Balice airport (50°04′40″, 19°47′42″) at a height of 237 m a.s.l., and rain gauge in station Kraków–Wola Justowska (50°03′50″, 19°53′25″) at a height of 204 m a.s.l. (Figure 1). For station Kraków–Balice, records from the multiannual period 1986-2006 were available, and for station Kraków–Wola Justowska records from the multiannual period 2007-2015. In that set, digital record series from the years 1986-1998 resulted from the digitalisation of pluviograph recording strips. The computer-aided method of conversion of pluviograph recording strips to digital format similar to that proposed by Licznar et al. (2011a) was adopted. Records from a standard unheated pluviograph (200 cm^2 orifice) covered warm year periods between spring (April or May) and autumn (October or November) when most of storm rainfall events occur in Poland. For later years, i.e., for the multiannual period 1999-2015, all-year records from electronic tipping bucket gauges were already available. On station Kraków–Balice, rainfall was recorded by an electronic rain gauge Aster TPG, and on station Kraków–Wola Justowska, by an electronic rain gauge Met One Instruments 60030. The resolution of recording rainfall depth in the case of the aforementioned devices was 0.1 mm and 0.2 mm, respectively, and their inlet surface was 200 cm^2. Records of local rainfall provided for the research by the MWSSC came from a network of a total of 23 rain gauges distributed throughout the city (Figure 1).

Figure 1. Location of rain gauges belonging to Waterworks Kraków and the Institute of Meteorology and Water Management

The entire network was composed of electronic tipping bucket gauges operating during both the summer and winter half-year. In all stations, rain gauges Hobo RG3M were installed, with rainfall record resolution of 0.2 mm and inlet surface of 200 cm^2. Unfortunately, records of rainfall series from these rain gauges were considerably shorter than in the case of the data base of the PANDa project, covering a period of 30 years (1986-2015). Detailed information regarding periods of rainfall records by particular rain gauges of the measurement network in Kraków is provided in Table 1 in the chapter discussing results obtained in the scope of designation of storm rainfalls. It should be emphasised that all rain gauges used in the analysis were located within the administrative boundaries of the city. The only exception was rain gauge Kraków–Balice located in the direct vicinity of the city boundaries, in the area of the nearby airport. All rainfall series from the entire period of monitoring of the precipitation field, recorded directly in digital form by tipping bucket gauges, as well as those resulting from digitalisation of records on pluviograph record strips, had a uniform temporal resolution of 1 minute.

2.2. Applied methodology

Based on the available time series of rainfall records (from the rain gauges of the PANDa project and from the network of MWSSC), storm rainfalls were designated by standard criteria proposed by Schmitt (2000) for identifying storms for urban drainage systems modelling. They are standard criteria applied in Germany and Poland, and used in the already cited paper by Licznar et al. (2011a). The adopted threshold minimum value of the total amount of storm rainfall was 10 mm, and the minimum time interval between single rainfall events was at least 4 hours. The designation of rainfalls in reference to dry periods also employed the minimum value of rainfall depth of 0.1 mm during 5 minutes as the threshold value for the interval to be considered as part of a rainfall event in terms of duration and precipitation amount. The result of designation of storm rainfalls separately from rain gauges of the PANDa project and the MWSSC network were data sets called set No. 1 and No. 2, respectively. Set No. 3 analysed in the final part of the research constituted a combination of sets No. 1 and No. 2.

Due to the differing durations and total depths of the designated storm rainfalls, further analysis of the sets, i.e., mutual comparison and identification of typical (quantifiable, model) distributions during rainfalls, involved double normalisation of cumulative hyetographs. This procedure was conducted in accordance with the methodology described in the publication (Licznar et al., 2017). It corresponded with the methodology of development of dimensionless hyetographs by Huff (1967). For this purpose, for each storm rainfall, total duration and total rainfall depth was determined. Next, each cumulative storm rainfall hyetograph with known total duration was divided into 100 even time intervals. For each of the subsequent intervals, a corresponding cumulative rainfall increase was determined. Subsequent cumulative precipitation increases were divided by total rainfall depth, obtaining unitary cumulative precipitation increases. As a result, the shape of each of the storm rainfalls was reflected by a hyetograph normalised to a range from 0 to 100% for duration and to a range from 0 to 1 (100%) for rainfall depth.

The mutual comparison of shapes of normalised (dimensionless) hyetographs within the analysed sets, and further designation of typical storm hyetographs that could be considered model hyetographs, useful e.g., for modelling of local drainage systems, employed tools for mining large data sets in the form of cluster analysis algorithms. Unlike in the case of earlier attempts to apply cluster analysis for the classification of storm rainfall hyetographs and for identification of model hyetographs (Licznar et al. 2017; Licznar, 2018; Wartalska et al., 2020), this study applied a complete cluster analysis methodology. The complete cluster analysis methodology involves seven stages (Milligan, 1996; Zhou et al., 2014):

1) Selection of objects and variables;
2) Selection of the formula of normalisation of variable values;
3) Selection of distance measure;
4) Selection of classification method;
5) Determination of the number of classes;
6) Assessment of classification results;
7) Class description (interpretation) and profiling.

The implementation of the first two points of the methodology, i.e., selection of objects and their variables, combined with normalisation of their values, was already characterised earlier, and aimed at the development of sets of normalised (dimensionless) hyetographs. Two further stages 3 and 4, covering the selection of classification methods and distance measures, were analogical to those in the already published papers by Licznar et al., 2017, Licznar, 2018, Wartalska et al., 2020. In comparison to these publications, three last missing stages of the cluster analysis were added in this paper, including: objective determination of the number of classes, and assessment of classification results, combined with simplified class profiling. This resulted in a coherent research methodology allowing for the classification of storm rainfall hyetographs, their division into the objectively determined number of clusters, determination of courses of model hyetographs, and their simplified profiling.

Cluster analysis tools find broad practical application in collating large data sets. Their implementation in the case of such sets permits the separation of their objects into a certain number of subsets, called clusters, covering mutually similar objects. The requirement of decouplability and sufficiency of the designated clusters is met, i.e., each of the elements belongs to a specific single cluster (Larose, 2005; Stanisz, 2007). Therefore, the sum of all clusters corresponds with the initial large set of objects, and particular clusters are separate and have no elements in common. In each cluster, objects are approximate, mutually similar, and simultaneously different from objects in other clusters. Depending on the adopted method, the division into clusters can be conducted to an *a priori* determined or undetermined number of clusters. In research in the Kraków polygon, cluster analysis was applied to the division of sets No. 1, No. 2, and No. 3 of dimensionless cumulative hyetographs of storm rainfalls into an undetermined, and then determined number of clusters. For this purpose, the hierarchical agglomeration method and non-hierarchical k-means clustering method were implemented.

Agglomeration methods have already found application in research on precipitation, not only in the context of search for similarities in temporal distribution of storm rainfalls for particular locations (Licznar et al., 2017; Wartalska et al., 2020), but also in the case of research on the variability of precipitation conditions within a large municipal precipitation monitoring network (Licznar et al., 2015). In their research conducted on the municipal monitoring network in Warsaw (including 25 rain gauges), comparable to that in Kraków in terms of size, Licznar et al. (2015) successfully implemented agglomeration methods to evidence similarities of empirical distributions of breakdown coefficients (BDCs) of rainfall from rain gauges located in different points in the city, for hierarchies of timescales, corresponding with time from 5 min to 1280 min. Agglomeration methods applied in this type of research aim at combining mutually similar objects through the application of appropriate measures of their mutual distance and the agglomeration method. The starting point is treating each object as a separate cluster. At the subsequent stages of agglomeration, objects most mutually approximate according to the defined measure (i.e., most similar by default) are combined into new clusters, covering objects and clusters resulting from earlier stages, until a single cluster is obtained. Based on the experience of other authors (Licznar et al., 2017; Wartalska et al., 2020), in research in the Kraków polygon, the already verified distance metrics were applied: Euclidean and Euclidean squared distance. The Euclidean distance metric $d(x,y)$ for two objects $x = (x_1, \ldots, x_r)$ and $y = (y_1, \ldots, y_r)$, characterised by r measurement values, is expressed in the following formula (Larose 2005, Stanisz 2007):

$$d(x,y) = \sqrt{\sum_{i=1}^{r}(x_i - y_i)^2}. \tag{1}$$

The Euclidean distance metric has simple and natural interpretation in the case of objects defined by only two or three measurement values ($r = 2$ or $r = 3$), because in that case its equivalent is distance on a plane and in space of two points x and y. Through analogy, in the case of the analysed sets of normalised hyetographs, the Euclidean distance between pairs of hyetographs was determined in space with a considerably higher number of dimensions (for $r = 100$).

Next to adopting a particular mutual distance metric for objects in the measurement space, the objective of combining mutually similar objects and clusters also requires the application of a particular method of their agglomeration. In this case, also based on the already cited papers (Licznar et al., 2017; Wartalska et al., 2020), the popular unweighted pair-group method was applied. In this method, differences in distances between all pairs of elements included to particular clusters are calculated. Values of the averaged differences between all pairs are adopted as the measure of distance between particular clusters. Due to this, it is known which elements of the sets are mutually similar and can be included to shared clusters, and moreover to what extent particular clusters are mutually similar and can be agglomerated into structures of larger clusters. A natural consequence of this is forming on the dendrograms (the resulting diagrams of the set structure in relation to the increasing bond distance, and therefore decreasing similarity between its elements) characteristic ‘chains’ made of similar objects developing increasingly extensive clusters.

Considering the primary study objective, i.e., the determination of model hyetographs, a more significant research tool was k-mean clustering. The application of this tool permitted separation of the analysed sets into *k* independent clusters differing to the greatest possible degree. In the case of earlier research on model hyetographs (Licznar et al., 2017, Wartalska et al., 2020), the application of the k-means method always involved questionable adopting of *k* number of clusters subjectively estimated based on the analysis of previously prepared dendrograms. It was assumed that at a certain level of bond distance, the chains of clusters can be cut to obtain several separate subsets of mutually similar dimensionless hyetographs. The obvious weakness of such an approach was lack of justification of the adopted level of bond distance at which the dendrogram was divided (cut). Moreover, as not observed in earlier papers (Licznar et al., 2017; Wartalska et al., 2020), simple cut-off of dendrograms at a given level of and distance and obtaining *k* clusters was not equivalent to the determination of the same number *k* of independent clusters by means of the k-means clustering method. Particular clusters could include differing subsets and objects as a result of differences in the classification methods between the hierarchical agglomeration method and non-hierarchical method, namely k-means clustering.

Unlike in the case of the hierarchical agglomeration method, aimed at combining objects and subclusters, agglomeration by means of the k-means clustering method aims at fragmenting the entire set into the *a priori* defined number *k* of clusters, whereas none of the *k* clusters is the subcluster of another cluster. The algorithm of agglomeration by means of the k-means clustering method therefore involves development of *k* subclusters, and then moving objects across them for distances between them within the subclusters to be as small as possible, and for distances between subclusters to be as large as possible. The moving procedure is repeated iteratively aiming at the most efficient separation of clusters (Larose, 2005). The final objective is arriving at a solution in which the designated clusters meet two criteria: that of internal coherence and external isolation (Gordon, 1999). This task was implemented in the computing environment of language R due to the resulting calculation difficulties. Their primary source was the number of classified objects. The number of all divisions of a set of *n* elements into *k* non-empty clusters is escribed by the following formula (Everitt et al., 2001; Gordon, 1999):

$$L(n,k) = \frac{1}{k!}\sum_{s=1}^{k}(-1)^{k-s}\binom{k}{s}s^{n}, \qquad (2)$$

where s is the number of class ($s = 1\ldots, k$). In accordance with formula (2), even in the case of a very small set of 10 objects with their division into 4 non-empty clusters, the number of all possible divisions is 34 105 (*L (10.4)* = 34105). For comparison, research has involved division of large sets of hyetographs including hundreds or even thousands of storm rainfalls.

These calculation challenges were further considerably multiplied during analyses of measurement sets No. 1, 2, and 3 due to multiple launching of the clustering algorithm with the application of the bootstrap method. In statistics, bootstrap methods are used to estimate the distribution of estimation errors by means of multiple sampling with replacement. Their implementation in the case of research on storm hyetographs from Kraków meant that the clustering algorithm was performed 150 times, each time for random samples from the analysed sets. Results obtained in subsequent iterations were compared, allowing for the designation of values of the bootmean parameter. The bootmean parameter was calculated as a mean value of the Jaccard index (Jaccard similarity coefficient) for each of the designated clusters. The Jaccard coefficient itself measures similarity between two sets, and is determined as the ratio of power set of the intersection of sets and power set of sum of these sets. High values of the Jaccard coefficient approximate to 1 strongly suggest perfect repeatability of the separation of objects into clusters. It is assumed that exceeding the threshold of 0.6 for Jaccard coefficients for each of the clusters suggests no occurrence among the designated clusters of clusters with random character, i.e., those including rainfall models deviating from the remaining clusters, but simultaneously not mutually similar. The verification of the above criterion allows drawing conclusions regarding the resulting clusters meeting the criteria of internal coherence and external isolation (Gordon, 1999). The bootmean parameter was used in the study for conducting assessment of results of the classification of particular sets No. 1, 2, and 3.

The compute-intensive process of k-means clustering was preceded by the stage of determination of the number of classes. For this purpose, a methodology was adopted analogical to that applied in research on temporal distributions of hourly water uptakes by Dzimińska et al. (2021). The analysis of previously prepared dendrograms provided the basis for the determination of a potential range of the number of clusters that should be considered in the case of division of storm rainfalls in Kraków. The optimal number of clusters was designated from that range based on the analysis of the total within sum of squares (wss) and values of the Caliński and Harabasz Index (CHIndex) for a variable number of clusters. Values of total within sum of squares (wss) and the Caliński and Harabasz Index (CHIndex) were calculated in accordance with the following formulas (Walesiak and Gatnar, 2009):

$$wss = \sum_i^k \sum_{x \in C_i} \|x - m_i\|^2, \tag{3}$$

$$CHIndex = \frac{SS_B}{SS_W} \cdot \frac{N-k}{k-1}, \tag{4}$$

where: k – number of clusters, x – element of a set, C_i – i-th data cluster, m_i – cluster centroid i, $\|x - m_i\|^2$- Euclidean distance between two vectors, N – total number of observations (elements of a set), SS_B – total variance between clusters (trace of interclass covariance matrix), SS_W – total internal cluster variance (trace of intraclass covariance matrix). The value of *wss* naturally decreases with an increase in the number of clusters k. Nonetheless, the gradient of the decrease evidently decreases after reaching a certain number of clusters considered optimal. In the case of the Caliński and Harabasz Index, the number of clusters is searched, considered optimal, for which its value is the highest. In accordance with formula (4), this means the occurrence of maximisation of the ratio of SS_B and SS_W, i.e., particular clusters differ from one another very significantly, and the elements of the set agglomerated in particular clusters are strongly mutually similar (relatively weakly variable).

The last seventh stage of the research covered the description (interpretation) and profiling of the obtained clusters. For all the clusters, their centroids were determined (arithmetic averages calculated from the original values of each variable based on objects developing a given cluster). This provided the basis for obtaining sets of normalised hyetographs for each of the analysed sets No. 1, 2, and 3. They were subject to mutual comparison. Finally, simplified profiling of clusters was conducted. The objective of cluster profiling is the identification of characteristic features of particular clusters allowing for the determination of differences between them. Cluster profiling is conducted based on variables that did not take part in the process of classification of the set of objects. In the case of research on hyetographs of storm rainfalls from Kraków, the available variables that did not participate in the process of classification of the set of objects were total rainfall depths and total rainfall durations. These general rainfall characteristics permitted the determination of distributions of mean precipitation intensities in particular clusters, i.e., simplified profiling of clusters in terms of explanation of differences in the obtained types of model hyetographs from the point of view of intensities of the represented storms.

3. Results and discussion

Based on the adopted criteria (Schmitt, 2000), from set No. 1, covering time series of precipitation records from 30 years for rain gauges of the PANDa project, a total of 313 storm rainfalls were designated (Table 1), corresponding to the frequency of their occurrence at a level of approximately 10 storms per year. Analogically, from the larger set No. 2 of precipitation series from 126 years, recorded on 23 rain gauges of the MWSSC network, a total of 1494 storm rainfalls were designated (Table 1), corresponding to the frequency of their occurrence at a level of approximately 12 storms annually. The obtained quantities of data sets were in accordance with the expectations and results of previously published research from the territory of Poland. For example, from a set of digitalised pluviography storm records from 38 years from Wrocław (south-western Poland), Licznar et al. (2011a) designated 250 storm rainfalls based on the same criteria, determining their frequency of occurrence at a level of 6.6 times per year. The lower precipitation frequency results from the fact that at each stage, precipitation records covered only periods of 5-6 months with positive air temperatures that allowed for the exposure of non-heated pluviographs. For comparison, in the polygon of the rain gauge network of the city of Warsaw composed of 25 electronic weighing rain gauges recording precipitation throughout the year, the criteria proposed by Schmitt (2000) permitted the designation of a total of 669 storm rainfalls (Licznar and Szeląg, 2014). The latter figure, in the context of records somewhat longer than two years (114 weeks) for each of the rain gauges, translates into the frequency of occurrence of storms equal to 12.3 events per year per single observation point.

Table 1. List of rain gauges belonging to sets No. 1 and 2 with characteristics of designated storms

Particular rainfalls differed in their courses in time. It is evident based on the example of cumulative normalised (dimensionless) hyetographs of set No. 1 in figure 2. A large majority of rainfalls, both in the case of set No. 1 and No. 2, had hyetographs considerably differing from theoretical hyetographs recommended for application as rainfall scenarios for hydrodynamic modelling of municipal drainage systems, such as: block rainfall, model rainfall according to DVWK, or model rainfall according to Euler type II. This observation also remains in accordance with research of other authors who already previously questioned the justification of drainage systems modelling in Poland commonly being based on model rainfall according to Euler type II (Licznar and Szeląg, 2014; Licznar et al., 2017).

Figure 2. Cumulative dimensionless hyetographs of 313 storm rainfalls from Kraków (set No. 1)

The designated storm rainfalls in particular sets No. 1 and 2 also differed in terms of total durations and total rainfall depths. In set No. 1, total depths of storm rainfalls were in a range from 10.0 to 105.0 mm, and their durations varied from 24 to 2329 minutes. These parameters in set No. 2 varied within similar ranges from 10.0 to 145.4 mm and from 14 to 4576 minutes in the case of total depths and total durations, respectively. These values also showed similar distributions in both analysed sets, as confirmed by histograms included in figure 3. A vast majority of storm rainfalls was characterised by total rainfall depth in a range from 10 mm to 20 mm, and their total durations varied from 2 h to 12 h. The similarity of total distributions of rainfall depths and durations in sets No. 1 and No. 2 presented in figure 3 at least partially justifies the possibility of use of sets of storm rainfalls recorded in stations in city outskirts and belonging to the national precipitation monitoring network in the context of issues regarding urban hydrology at a scale of a city as large as Kraków.

Figure 3. Histograms of rainfall depths and durations for the designated sets of rainfalls from Kraków, diagrams in the top row for set No. 1 (313 rainfalls), diagrams in the bottom row for set No. 2 (1493 rainfalls)

The analysed sets No. 1 and No. 2 of normalised cumulative precipitation were used for preparing dendrograms presented in figures 4a and 4b, respectively. As demonstrated based on the example of the least abundant set No. 1 in figure 4a, even in the case of number of objects only slightly exceeding 300, their abundance makes the prepared detailed dendrogram largely illegible. Due to this, the same dendrogram and all remaining dendrograms for more abundant sets were prepared for a number of leaf nodes reduced to 30 (Figures 4a, 4b, 5a, and 5b), considerably improving their legibility. In practice, this measure corresponded with cut-off of detailed dendrograms at a level of bond distance of approximately 1.0. On the prepared dendrograms in figures 4a and 4b, it is easy to recognise the characteristic structures in the form of chains connecting similar objects or subclusters of objects occurring for lower bond distances. The obtained image of dendrograms therefore corresponded in terms of quality to dendrograms published in earlier papers regarding the application of agglomeration methods in the analysis of precipitation hyetographs (Licznar and Szeląg, 2014; Licznar et al., 2017, Licznar, 2018; Wartalska et al., 2020). In the case of figure 4b, notice that its structure included a cluster with number 29 with extremely high bond distance (approximately 5.5). Detailed analysis of cluster No. 29 in figure 4b showed that it is composed of only one object.

Figure 4 a. Dendrograms obtained for set No. 1 composed of 313 dimensionless cumulative rainfall hyetographs from Kraków (top panel). Below find the same dendrogram prepared for the reduced number of 30 leaf nodes (bottom panel). In the diagrams, vertical axes show bond distances for particular rainfalls and rainfall clusters. The horizontal axis of the dendrogram on the bottom panel shows numbers of rainfalls in particular clusters; **b.** Dendrogram obtained for set No. 2 composed of 1494 dimensionless cumulative rainfall hyetographs from Kraków. The dendrogram was prepared for the reduced number of 30 leaf nodes. The vertical axis of the diagram shows bond distances for particular rainfall clusters, and the horizontal axis shows their numbers

Figure 5 a. Dendrogram obtained for adjusted set No. 2 composed of 1493 dimensionless cumulative rainfall hyetographs from Kraków. The dendrogram was prepared for the reduced number of 30 leaf nodes. The vertical axis of the diagram shows bond distances for particular rainfall clusters, and the horizontal axis shows their numbers; **b.** Dendrogram obtained for adjusted set No. 3 composed of 1806 dimensionless cumulative rainfall hyetographs from Kraków. The dendrogram was prepared for the reduced number of 30 leaf nodes. The vertical axis of the diagram shows bond distances for particular rainfall clusters, and the horizontal axis shows their numbers

The object was a hyetograph of a storm rainfall designated from records of rain gauge No. 12 in Płaszów in 2014 (Figure 6). The rainfall had a total depth of 13.2 mm and total duration of 507 min. Although the rainfall event formally meets the adopted criteria of storm rainfall (Schmitt, 2000), it showed very specific course in time (Figure 6). For the first 460 min. of the rainfall, only 0.8 mm of rain was recorded. In the second, considerably shorter, final part of the rainfall lasting 47 min, 12.4 mm of rain was recorded. Considering this very specific course of the rainfall in time, it was excluded from set No. 2, and the analysis was continued for a set of 1493 normalised hyetographs of the remaining rainfall events.

Figure 6. Hyetograph of a rainfall event recorded on station Kraków-Płaszów in 2014

For the adjusted set No. 2 including 1493 objects, the dendrogram presented in figure 5a was obtained. As a result of elimination of outliers in the form of a single rainfall event from rain gauge No. 12, a dendrogram was obtained with a maximum bond distance at a level somewhat higher than 4.0, comparable like in the case of set No. 1 (Figure 3). Importantly, after the adjustment, the most deviating cluster with number 24 was not composed of a single object, but covered normalised hyetographs for 9 different storm rainfalls (combined into a single cluster for bond distances lower than 1.0). At the final stage of application of hierarchical methods for analysing structures of sets of normalised hyetographs, a dendrogram was also prepared for set No. 3, constituting a combination of sets No. 1 and No. 2, presented in figure 5b. The dendrogram obtained in the case of set No. 3, prepared for 1806 dimensionless storm rainfall hyetographs, had a structure analogical to that of the already discussed dendrograms in figures 4a and 5a. Its maximum bond distance did not exceed 4.0, and it showed characteristic chain connections of similar subclusters of objects, mutually connecting for bond distances of less than 1.0. Irrespective of the similarities indicated herein, the comparison of three dendrograms obtained for sets No. 1, 2 and 3 (Figures 4a, 5a, and 5b) provided no basis for answering the question regarding the appropriate and unquestionable number of classes in the division of normalised hyetographs from Kraków into clusters. To illustrate the problem in a simple way, even if a certain subjective level of bond distance is adopted a priori, e.g., 1.75, then the structure of the dendrogram of sets No. 1, 2 and 3 allows for designating the following mutually divergent numbers of clusters: 11, 9 and 6. Considering the substantial discrepancy in terms of number of clusters in further research, a potential range of the number of clusters *k* was subject to analysis, covering values from 2 to 20.

The results of calculations of the Caliński and Harabasz Index values (*CHIndex*) as well as total within sum of squares (*wss*) for the number of clusters k within a range from 2 to 20 for set No. 1 are presented in figure 7.

Figure 7. Value of the CHIndex and total within sum of squares (wss), and for a set 1 of 313 rainfalls from Kraków, depending on the adopted number of clusters k

Diagrams of both parameters directly suggest that the optimum number of clusters for set No. 1 should be adopted as equal to 4 ($k = 4$). By theory, *wss* values naturally decrease with an increase in the number of clusters *k*, although the decrease gradient evidently decreases after reaching number of clusters $k = 4$. For the same number of four clusters, maximum CHIndex value is also observed (*CHIndex*=170). In accordance with formula (4), a high CHIndex value is correlated with maximisation of the ratio of SS_B and SS_W. This means that particular clusters very significantly differ from one another, and elements of the set grouped in particular clusters are strongly similar to one another (relatively weakly variable).

The determination of the optimum number of clusters $k = 4$ was followed by the k-means clustering process for set No. 1 of normalised cumulative hyetographs from Kraków. Through the application of the bootstrap method, the following values of the bootmean parameter were obtained for subsequent clusters from 1 to 4: 0.72; 0.65; 0.84, and 0.74. Bootmean parameter values were higher than 0.6 for all four clusters. This evidences that the designated four clusters included no cluster with random character, i.e., one that includes rainfall models deviating from the remaining three clusters, but at the same time not mutually similar. As a result of the k-mean clustering process for set No. 1 including 313 cumulative normalised hyetographs from Kraków, clusters No. 1, 2, 3, and 4 were ascribed 102, 93, 68, and 50 rainfalls, respectively. This corresponded with the share of 35%, 32%, 23%, and 17%, respectively, throughout the set of analysed rainfalls.

For the designated clusters, averaged dimensionless cumulative storm rainfall hyetographs were determined, presented in figure 8. The cumulative dimensionless hyetographs were also transformed to the form of storm rainfall hyetographs presented in figure 9.

Figure 8. Diagrams of averaged dimensionless cumulative rainfall hyetographs for four clusters designated by means of k-means clustering for Kraków based on set No. 1 for 313 storm rainfalls

Figure 9. Model dimensionless rainfall hyetographs developed by means of the k-means clustering method for set No. 1 for 313 rainfalls from Kraków. The horizontal axis shows percent increase in rainfall duration, and the vertical axis shows percent shares in total precipitation amount

The analysis of the obtained model hyetographs shows that the most frequently occurring hyetographs of type 1 and 2 (35% and 32%, respectively) have relatively even values of point rainfall intensity (point rainfall depths for unitary duration intervals of 1/100 of total duration do not exceed 2% total rainfall depth). The rainfalls, however, do not correspond with the simplified block rainfall model, commonly used in designing drainage systems, and even sporadically applied in their hydrodynamic modelling. In the case of numerous unitary duration intervals, point rainfall depths differ from 1%, they are considerably lower, and frequently approximate to 0.5%, or considerably higher, reaching approximately 2% of total rainfall depth.

The substantially more seldom occurring model hyetograph for cluster 3 has a shape very generally approximate to model rainfall according to Euler type II recommended by Schmitt (2000) for modelling stormwater systems. Unlike in the case of rainfall model according to Euler type II, the greatest rainfall accumulation occurs not in 1/3 of rainfall duration, but already in its initial part, during the first 10% of the duration of the entire rainfall. Moreover, this accumulation has no character of a very sharp peak (point rainfall depths for unitary duration intervals of 1/100 of total duration do not exceed 4% of total rainfall depth). The most seldom occurring hyetograph shape is that determined in the case of cluster 4. With a high degree of generalisation, the hyetograph can be treated as a mirror reflection of the model hyetograph of cluster 3. This is not strictly accurate, because the rainfall accumulation occurring in the final part of the hyetograph is very obscure. It is observed in the final interval covering approximately 30% of the entire rainfall, and point rainfall depths for unitary duration intervals of 1/100 of total duration do not exceed 2.5% of total rainfall depth in that case. Moreover, at the very beginning of the model hyetograph for cluster 4, a small rainfall peak occurs, with no equivalent in the final part of the model hyetograph for cluster 3. Referring to he discussed results obtained in the analysis of set No. 1 from Kraków, it is worth emphasising that the obtained courses of cumulative model hyetographs in figure 8 were very approximate to the cumulative model hyetographs obtained by Licznar (2018) for another rain gauge from Poland, as a result of application of clustering of a set of 213 storm rainfalls by means of the k-means clustering method for a subjectively adopted number of 4 clusters. The obtained results also remain in accordance with earlier studies from the territory of Poland that question the justification of common application of the synthetic rainfall model according to Euler type II due to its deviation from the vast majority of scenarios of temporal course of actual storm rainfalls (Licznar and Szeląg, 2014; Licznar et al., 2017).

Diagrams of the dependencies of the Caliński and Harabasz (*CHIndex*) index values and total within sum of squares (*wss*) on the number of clusters *k* developed for set No. 2 are presented in figure 10.

Figure 10. Value of the CHIndex and total within sum of squares (wss), and for set No. 2 of 1493 rainfalls from Kraków, depending on the adopted number of clusters k

Analogically to set No. 1, they provide the basis for the determination that the optimum number of clusters for set No. 2 should be adopted as equal to 4 ($k = 4$). For four clusters ($k = 4$), an evident peak of the CHIndex is observed (*CHIndex*=801), and the curve of total within sum of squares (*wss*) flattens out to below 2000 after a rapid decrease from a level of approximately 5000. The cited CHIndex and wss values cannot be referred to values obtained in the case of set No. 1 (Figure 7). Orders of magnitude in both cases are different, as results from different abundance of sets No. 1 and No. 2, and as accounted for by the structure of formulas (3) and (4).

The process of k-means clustering of set No. 2 of normalised cumulative hyetographs from Kraków into four clusters ended with the designation of clusters meeting the criteria of internal coherence and external isolation (Gordon, 1999). The confirmation of the above was obtaining values of the bootmean parameter substantially exceeding the threshold of 0.6 for each of the clusters. The parameter reached: 0.93; 0.91; 0.96, and 0.91, respectively for clusters from 1 to 4. As the final result of clustering, subsequent clusters were ascribed, respectively: 510, 506, 232, and 245 storm rainfalls. This corresponded with the respective share of 34%, 34%, 16%, and 16% in the entire population of analysed rainfalls in set No. 2. The cited percent share of particular clusters in set No. 2 proved approximate, like in the case of the previously discussed set No. 1. For particular clusters, averaged cumulative storm rainfall hyetographs were also determined, presented in figure 11. The diagrams of averaged normalised cumulative hyetographs in the figure pointed to very high compatibility in terms of temporal distribution with the previously discussed results for set No. 1 (Figure 8).

Figure 11. Diagrams of averaged dimensionless cumulative rainfall hyetographs for four clusters designated by means of the k-means clustering method for Kraków based on set No. 2 for 1493 storm rainfalls

The evident divergence of results of clustering of storm rainfall hyetographs for sets No. 1 and No. 2 became an impulse for undertaking analysis of set No. 3, constituting a combination of the aforementioned sets. As expected, the obtained results proved to be virtually identical to results obtained previously for sets No. 1 and No. 2. Curves of variability of values of the Caliński and Harabasz index (CHIndex) and total within sum of squares (wss) developed for this large set of 1806 storm rainfalls are presented in figure 12, and raise no doubts as for their interpretation. The evident maximum of the CHIndex value (*CHIndex*=1105) is obtained for four clusters (*k* = 4). For four clusters, a very steep gradient of decrease in *wss* values from more than 5500 to approximately 2000 is also rapidly levelled almost to zero. Clustering by means of the k-means method of hyetographs included in set No. 3 ended with creating four clusters meeting the criteria of internal coherence and external isolation. The obtained bootmean parameter values were practically approximate to 1.0, and for subsequent clusters they reached: 0.94; 0.93; 0.96, and 0.94, respectively. Clusters No. 1, 2, 3, and 4 were ascribed: 613, 605, 288, and 300 rainfalls, respectively, corresponding to a respective share of 34%, 33%, 16%, and 17% in the entire set No. 3.

Figure 12. Value of the CHIndex and total within sum of squares (wss), and for a set 3 of 1806 rainfalls from Kraków, depending on the adopted number k of clusters

For particular clusters of set No. 3, averaged hyetographs of dimensionless cumulative storm rainfalls were also determined, collectively presented in figure 13.

Figure 13. Diagrams of averaged dimensionless cumulative rainfall hyetographs for four clusters designated by means of the k-means clustering method for Kraków based on set No. 3 for 1806 storm rainfalls. For comparison, the diagram also shows Median Time Distributions of Heavy Storm Rainfall at a Point developed by Huff (1990)

The figure also presents dimensionless hyetographs developed by Huff (1990), constituting the subject of a separate discussion. Pursuant to the expectations, the diagram in figure 13 proved strongly similar to the diagrams of model hyetographs for clusters from 1 to 4 determined for sets No. 1 and No. 2 (Figures 8 and 11). For better assessment of the scale of similarity of the model hyetographs obtained for particular clusters in the case of clustering sets No. 1, No. 2, and No. 3, figure 14 was additionally prepared.

Figure 14. Comparison of averaged dimensionless cumulative rainfall hyetographs for four clusters designated by means of the k-means clustering method based on sets No. 1, 2, and 3: 313, 1493, and 1806 storm rainfalls recorded in Kraków. The horizontal axes present percent increase in rainfall duration, and the vertical axes show percent shares in total rainfall depth

The similarity of averaged normalised cumulative rainfall hyetographs obtained for all three sets of storm rainfalls from Kraków within particular clusters is unquestionable. In the case of sets No. 2 and No. 3, the diagrams of normalised cumulative hyetographs even overlap. In the case of cluster No. 4, hyetographs resulting from the analysis of all three sets overlap. This provides the basis for the presumption that the developed methodology of application of cluster analysis for the designation of model hyetographs shows repeatability in the scope of obtained results within the urban precipitation field. In engineering practice, it can be therefore used for the determination of local sets of model hyetographs based on the analysis of approximately 30-year-long rainfall series even from single rain gauges located near city boundaries. They will be able to supply computer models for simulation of drainage systems in the centre of a city as large as Kraków with satisfactory precision.

For a better understanding of the obtained results, particularly including the determination of characteristic features of particular clusters that differentiate them from one another, their profiling was performed. Such profiling is called simplified, because it is only based on general variables describing storm rainfalls in the form of their total depths, total durations, and mean intensities. For this purpose, table 2 presents mean values of these variables within particular clusters for sets No. 1 and No. 2. Figures 15a and 15b present histograms of the frequency of occurrence of storm rainfalls with different mean intensities within different clusters, respectively for sets No. 1 and No. 2. Values of rainfall intensity in diagrams in figures 15a and 15b and in table 2 are expressed in a standard unit of $dm^3/(s \cdot ha)$ applied in urban hydrology.

Table 2. Mean values of total depths, durations, and intensities of storm rainfalls included in particular clusters for sets No. 1 and No. 2 from Kraków

Figure 15 a. Histograms of mean rainfall intensities for four designated clusters in rainfall set No. 1 from Kraków (313 rainfalls); **b.** Histograms of mean rainfall intensities for four designated clusters in rainfall set No. 2 from Kraków (1493 rainfalls)

Data included in table 2 and in diagrams in figures 15a and 15b suggest that rainfalls classified to cluster 3 usually showed mean intensities higher than those of rainfalls included in other clusters. For the majority of storm rainfalls included in cluster 3 in sets No. 1 and No. 2, mean intensities in sets No. 1 and No. 2 were usually within the range from 5 to 20 $dm^3/(s \cdot ha)$, and their mean value within sets No. 1 and No. 2 exceeded 14 $dm^3/(s \cdot ha)$. High mean rainfall intensities in cluster 3 did not result from high total rainfall depths, but from evidently shorter durations. In set No. 2, mean rainfall depths for cluster 3 were even evidently lower than for the three remaining clusters. Mean rainfall durations for cluster 3, however, were approximately 6 h, whereas for the remaining clusters they were considerably longer, within a range from approximately 8 h to 12 h. This suggests that cluster n = 3 included short but very intensive rainfalls, probably with convection genesis. This hypothesis would also explain the greatest variability of point rainfall depths for unitary duration intervals shown on the model hyetograph of cluster 3 in figure 9. Completely different profiling results were obtained for cluster 2 which had a considerably more equalised course of the model hyetograph (Figure 9). Rainfalls classified to this cluster had not only longer durations within the clusters in particular sets No. 1 and No. 2, but usually also very low intensities within a range from 0 to 5 $dm^3/(s \cdot ha)$. This encourages a hypothesis that this cluster primarily included frontal rainfalls with long durations, but low and considerably more even in time intensities. The research hypotheses stated here regarding the division of storm rainfalls into clusters by rainfall genesis should be verified in further research. It will however require access to synoptic records permitting more precise profiling of the designated clusters.

In the summary of the entire discussion of results, it is also necessary to refer to the classic papers by Huff (1967, 1990). Although the research on Dimensionless, Cumulative Rainfall Hyetographs was implemented in different climate conditions (Illinois, USA), with a different approach to the identification of single rainfalls, and with the application of a considerably simpler method of their classification to the first, second, third, or fourth quartile (depending on whether the highest percent of cumulative rainfall occurred in the first, second, third, or fourth quarter of its duration, respectively), relatively high correspondence of the courses of medians is observed (50th-percentile) between dimensionless hyetograph curves derived from point rainfall values derived by Huff (1967, 1990) and model hyetographs in figures 8, 11 and 13. In the case of hyetographs from Kraków, cluster 3 corresponds with first-quartile storms, cluster 1 with second-quartile storms, cluster 2 with third-quartile storms, and cluster 4 with fourth-quartile storms. Further analogies can be sought in general characteristics of particular quartiles. For designing and modelling drainage systems, Huff (1990) recommended the application of first-quartile storm hyetographs for time scales of about 6 hours or less, and second-quartile storm hyetographs for time scales of about 6 to 12 hours. These recommendations overlap with mean durations determined for clusters No. 3 and No. 1 in sets No. 1 and 2, respectively (Table 2). Nonetheless, at a closer investigation of the study by Huff (1967), differences are observed in terms of frequencies of occurrence of rainfalls with the adopted model hyetographs. In the case of research from Illinois, the relative frequencies of the storms were 30, 36, 19, and 15 percent for the first, second, third, and fourth quartiles, respectively. They are not in accordance with frequencies of 16, 34, 33, and 15 percent obtained for the analogical model hyetographs designated from set No. 3, corresponding to subsequent clusters No. 3, 1, 2, and 4, respectively. The determined divergence, however, does not undermine results from Kraków, because the analogical divergence has already been signalled and discussed by Pani and Haragan (1981). Analysing a set of 117 rainfalls from Texas (USA), recorded in months with the highest probability of occurrence of convection rainfalls, the authors obtained a median (50th-percentile) dimensionless hyetograph curve with shapes fully corresponding with results by Huff (1967), but the determined relative frequencies of the storms were 13, 41, 32, and 14 percent for the first, second, third, and fourth quartiles, respectively. The latter frequencies are considerably more approximate to results from Kraków, despite obvious differences in the location of both research polygons and approach to processing rainfall records and determination of model hyetographs. The qualitative compatibility of study results from Kraków with classic papers by Huff (1967, 1990) and Pani and Haragan (1981) is an additional premise confirming the accuracy of the methodology of identification of model hyetographs of storm rainfalls based on complete cluster analysis and quality indices.

4. Summary and final conclusions

Progress in the scope of atmospheric precipitation measurements techniques, and the occurring municipal rain gauge networks expansion requires simultaneous modernisation of the methodology of recorded precipitation series processing. A necessary element of such a methodology are certainly modern methods of objective and automatic search of rainfalls groups with similar courses in time that can be described in a general way by means of model hyetographs. The practical application area of such methods can currently exceed processing sets of local model hyetographs, and find implementation in the practical rainwater runoff control systems operation. Due to the growing number of rain gauges and rapidly increasing sets of rainfall records, also in the case of model hyetographs identification, it becomes justified and necessary to reach for data mining tools, primarily including the cluster analysis.

This paper is not pioneer in terms of the very idea of the cluster analysis application in storm rainfall hyetographs classification. The application of the cluster analysis in research on temporal distributions of storm rainfalls has already been postulated by Licznar et al. (2017), and then tested for several locations in Poland (Licznar, 2018; Wartalska et al., 2020). The primary objective of the paper was the improvement of the research methodology to meet the requirements of the complete cluster analysis methodology covering seven stages (Milligan, 1996), including: selection of objects and variables; formula selection of variable values normalisation; selection of distance measure; selection of the classification method; the number of classes determination; assessment of classification results; class description and profiling. The aforementioned objective covered the three primary detailed objectives involving: objectivization of the number of clusters determination, the internal coherence and external isolation of clusters verification, and profiling of the retrieved clusters.

Another substantial objective of the paper was to demonstrate the developed methodology at the scale of a large precipitation field in Poland. This objective covered two detailed objectives, namely testing the methodology on a large measurement set, and equally importantly, analysing its repeatability at the scale of a large urban precipitation field. The basic question was also to what extent model hyetographs developed based on records from nearby rain gauges (located e.g., at an airport or in suburbs) correspond with the shapes of model hyetographs for rain gauges of the urban rain monitoring network. Owing to the collaboration with the Municipal Water Supply and Sewerage Company (MWSSC) in Kraków, Poland, it was possible to apply the developed complete cluster analysis methodology to a large measurement set of 1806 storm rainfalls (Set No. 3), composed of set No. 1 – 313 storm rainfalls designated from two nearby rain gauges belonging to the countrywide network of IMGW, and set No. 2 – 1493 storm rainfalls designated from 23 rain gauges belonging to the municipal rain monitoring network of MWSSC. Three applications of the complete cluster analysis methodology for sets No. 1, No. 2, and No. 3 permitted its thorough testing, designation of a set of model hyetographs for practical application in modelling of drainage systems in Kraków, and drawing the following final conclusions:

1) The complete methodology of the storm rainfall hyetographs cluster analysis should cover tools for both hierarchical and non-hierarchical analysis of the sets structure. Before the application of the cluster of normalised (dimensionless) hyetographs, key in terms of the final products, i.e., model hyetographs, by means of the k-means method, the hierarchical agglomeration method should be applied to prepare dendrograms of similarity of temporal courses of rainfalls in the analysed sets. The dendrograms should be subject to expert analysis not only in terms of determination of a potential number of clusters in the analysed sets, but more importantly the identification of particularly peculiar rainfall patterns. Like in the case of set No. 2 and rainfall recorded by rain gauge No. 12 in Płaszów in 2014, such records should be removed before the division of the set using a predefined number k of clusters by means of non-hierarchical methods;
2) The diagrams analysis of correlation of values of the Caliński and Harabasz index (*CHIndex*) and total within sum of squares (*wss*) with number *k* of clusters permits completely objective determination of the correct number of clusters for which the division of storm rainfall sets should be performed from the similarity point of view of their normalised (dimensionless) hyetographs. For the accurate, optimal number of clusters, maximisation of *CHIndex* values is observed combined with an evident decrease in the gradient of the decrease in *wss* values. In the case of all three analysed sets of storm rainfalls from Kraków, based on analyses of *CHIndex* and *wss* values, the adopted optimum number of clusters was four (k = 4), and the choice was positively verified in all further research through obtaining clusters meeting the requirements of internal coherence and external isolation;

3) The fundamental element for the credibility of the obtained results of storm rainfalls divisions and model hyetographs identification is the assessment of classification results. The study conducted on three sets of rainfalls from Kraków justified the repeated launching of the clustering algorithm with the application of the bootstrap method. Although this undoubtedly complicates the computing algorithms and prolongs the time of calculations, it allows for calculating the bootmean parameter corresponding to the mean value of the Jaccard Index (Jaccard similarity coefficient) for each of the designated clusters. The bootmean parameter permits drawing objective conclusions on whether particular clusters meet the criteria of internal coherence and external isolation (Gordon, 1999). For all clusters designated from three sets of rainfalls from Kraków, the bootmean parameter usually exceeded the adopted threshold of 0.6, confirming that the designated subclusters included no clusters with random character, i.e., those including rainfall patterns deviating from the remaining clusters, but also evidently mutually different. Relatively lowest values of the bootmean parameter were obtained for the least abundant set No. 1 (in a range from 0.65 to 0.84), whereas for approximately five or six times more abundant sets No. 2 and 3, they were higher than 0.9, or even approximate to 1.0. The latter observation suggests that the k-means clustering method is predestined for the analysis of very large sets, and is more reliable in their case;

4) The developed complex cluster analysis methodology for the division of sets of storm rainfalls and identification of model hyetographs implemented in the case of all three sets of storm rainfalls from Kraków generated coherent final results. For each of the three sets, the optimum number of clusters was four, and the resulting averaged normalised cumulative hyetographs for particular clusters showed no mutual differences within the three analysed sets. The coherence of the obtained results also concerned the

frequency of storm rainfalls occurrence included to particular clusters. For all the three sets, storm rainfalls were distributed in proportions of approximately: 1/3, 1/3, 1/6, and 1/6 for clusters No. 1, 2, 3, and 4, respectively. All the aforementioned observations suggest the possibility of development of model hyetographs based on multiannual records from suburban stations (e.g., from the rain gauge at the nearby airport), and like in the case of set No. 1, their application in practice throughout the city in the case of lack of the possibility of hyetographs development based on records from the territory of the city itself (based on set No. 2).

5) The obtained set of model hyetographs for Kraków does not include hyetographs with a shape corresponding to that of synthetic hyetographs developed based on IDF (Intensity-Duration-Frequency) or DDF (Depth-Duration-Frequency) models, adopted *a priori* for hydrodynamic modelling of drainage systems in Poland, such as: model rainfall according to Euler type II, block rainfall, or model rainfall according to DVWK. Nonetheless, the shapes of the developed hyetographs point to high similarity to classic medians (50th-percentile) of dimensionless hyetograph curves derived from point rainfall values derived by Huff (1967, 1990) and Pani and Haragan (1981). In results from Kraków, model hyetographs for subsequent clusters No. 3, 1, 2, and 4 correspond to medians (50th-percentile) of dimensionless hyetographs for: first-quartile storms, second-quartile storms, third-quartile storms, and fourth-quartile storms. Profiling of clusters results of storm rainfalls from Kraków also remain in complete accordance with earlier research by Huff (1990) according to which first-quartile storms (storms from cluster 3) usually correspond with time scales of about 6 hours or less, whereas second-quartile storms (storms from cluster 1) usually have longer durations within a range from 6 to 12 hours.

The development of a complex methodology of storm rainfall hyetographs analysis and its successful testing in a polygon of a large rain gauge network in Kraków offers the possibility of its implementation at a considerably broader scale in the scope of implementation of the WaterFolder Connect project. The practical objective here is to develop credible sets of local model hyetographs in a network of 100 rain gauges in Poland for the purpose of their later use in practice for supplying a digital platform dedicated for designing and modelling drainage systems throughout Poland. Further research, however, must be also undertaken due to new hypotheses that appeared as a result of the study. Firstly, the research hypothesis assuming the correlation of the division of storm rainfalls into particular clusters with the genesis of rainfalls needs to be verified. Moreover, in the context of the determined lack of variability of model hyetographs at the single urban precipitation field scale, it is important to verify the thesis on the regionalisation of study results possibility from more mutually distant rain gauges, and the practical application of common sets of model hyetographs in larger areas of the country. In the future, the developed methodology of hyetograph analysis could be also implemented in other countries around the globe.

Acknowledgments: This work was carried out as part of the fourth edition of the programme implementation doctorate conducted by the Ministry of Science and Higher Education, Republic of Poland, and within the project entitled '*WaterFolder Connect – an integrated platform for design and modelling of drainage systems* – POIR.01.01.01-00-0119/21', financed by the National Centre for Research and Development under the Smart Growth Operational Programme 2014-2020, Priority axis: Support for R&D in enterprises; Measure: R&D projects in enterprises; Sub-measure: Industrial research and development conducted by enterprises.

Precipitation data were partially provided by the Municipal Water Supply and Sewerage Company (MWSSC) in Kraków, Poland.

References

1. Kundzewicz, Z., Licznar, P., 2021 (2022). Climate change adjustments in engineering design standards. European perspective. Water Policy (in print).
2. Nix, S.J., 1994. Urban stormwater modelling and simulation. Lewis Publishers, CRC Press.
3. de Lima, M.I.P., 1998. Multifractals and the temporal structure of rainfall. Doctoral dissertation, Wageningen Agricultural University, Wageningen.
4. Deidda, R., Benzi, R., Siccardi, F., 1999. Multifractal modeling of anomalous scaling laws in rainfall. Water Resources Research 35, 1853–1867.
5. Licznar, P., 2009. Generatory syntetycznych szeregów opadowych do modelowania sieci kanalizacji deszczowych i ogólnospławnych. Monografie LXXVII. UP we Wrocławiu, pp 180.
6. Licznar, P., 2008. Obliczenia częstotliwości nadpiętrzania sieci kanalizacji deszczowej. Gaz, Woda i Technika Sanitarna 7-8, 16-21.
7. Gires, A., Onof C., Maksimovic C., Schertzer D., Tchiguirinskaia I., and Simoes N., 2012. Quantifying the impact of small scale unmeasured rainfall variability on urban hydrology through multifractal downscaling: a case study, J. Hydrol. 442–443, 117–128.

8. Gires, A., Tchiguirinskaia, I., Schertzer, D., and Lovejoy, S., 2013. Multifractal analysis of an urban hydrological model on a Seine-Saint-Denis study case, Urban Water J. 10, 195–208.ightlights. Remote Sensing of Environment 147, 173-185.
9. Zhou, Y., Smith, S.J., Elvidge, C.D., Zhao, K., Thomson, A., Imhoff, M., 2014. A cluster-based method to map urban area from DMSP/OLS n
10. Licznar, P., 2013. Stormwater reservoir dimensioning based on synthetic rainfall time series, Ochrona Srodowiska, 35, 27–32.
11. Jakubiak, P., Licznar, P., Malinowski, S., 2014. Rainfall estimates from radar vs. raingauge measurements. Warsaw case study. Environment Protection Engineering 40, 162-170.
12. Nayak, M.A., Ghosh, S, 2013. Prediction of extreme rainfall event using weather pattern recognition and support vector machine classifier. Theor Appl Climatol 114, 583–603.
13. Molnar, P., Lüscher, R., Hausherr, R., 2006. Impact of storm rainfall variability on urban drainage system performance. Proc. 7th International Workshop on Precipitation in Urban Areas: Extreme Precipitation, Multisource Data Measurement and Uncertainty, 7-10 Dec. 2006, St. Moritz, Switzerland.
14. Güntner, A., Olsson, J., Calver, A., Gannon, B., 2001. Cascade-based disaggregation of continuous rainfall time series: the influence of climate. Hydrology & Earth System Sciences 5, 145–164.
15. Hingray, B. and Ben Haha, M., 2005. Statistical performances of various deterministic and stochastic models for rainfall series disaggregation, Atmos. Res. 77, 152–175.
16. Licznar, P., Łomotowski, J., Rupp, D.E., 2011a. Random cascade driven rainfall disaggregation for urban hydrology: An evaluation of six models and a new generator. Atmospheric Research 99, 563-578.

17. Licznar, P., Schmitt, T.G., Rupp, D.E., 2011b. Distributions of microcanonical cascade weights of rainfall at small timescales. Acta Geophysica 59, 1013-1043.
18. Over, T.M., Gupta, V.K., 1994. Statistical analysis of mesoscale rainfall: dependence of a random cascade generator on large-scale forcing. Journal of Applied Meteorology 33, 1526–1542.
19. Rupp, D.E., Licznar, P., Adamowski, W., Leśniewski, M., 2012. Multiplicative cascade models for fine spatial downscaling of rainfall: parameterization with rain gauge data. Hydrol. Earth Syst. Sci. 16, 671–684.
20. Deidda, R., 2000. Rainfall downscaling in a space-time multifractal framework. Water Resources Research. https://doi.org/10.1029/2000WR900038.
21. Schmitt, T.G., 2000. Kommentar zum Arbeitsblatt A 118 Hydraulische Bemessung und Nachweis von Entwässerungssystemen. DWA, Hennef.
22. Sifalda, V., 1973. Entwicklung eines Berechnungsregens für die Bemessung von Kanalnetzen. GWFWasser Abwasser 114, 435–440.
23. Yen, B.C., Chow, V.T., 1980. Design hyetographs for small drainage structures. J. Hydraul. Eng. Div. ASCE 106, 1055–1976.
24. Desbordes, M., 1978. Urban runoff and design storm modeling. In Proceedings of the First International Conference on Urban Drainage, London, UK, April 1978; pp. 353–361.
25. Peyron, N., Nguyen, V.T.V., Rivard, G., 2002. An optimal design storm pattern for urban runoff estimation in southern Québec. In Proceedings of the 30th Annual Conference of the Canadian Society for Civil Engineering, Montréal, QC, Canada, 5–8 June 2002.
26. Lee, K.T., Ho, J.Y., 2008. Design hyetograph for typhoon rainstorms in Taiwan. J. Hydrol. Eng. 13, 647–651.

27. Keifer, C.J., Chu, H.H., 1957. Synthetic storm pattern for drainage design. J. Hydr. Eng. Div. 83, 1–25.
28. DVWK. Arbeitsanleitung zur Anwendung Niederschlag-Abflub-Modellen in kleinen Einzugsgebieten. Regeln 113 (Teil II: Synthese); Verlag Paul Parey: Hamburg, Germany, 1984.
29. Huff, F.A., 1967. Time Distribution of Rainfall in Heavy Storms. Water Resources Research, 3, 1007-1019.
30. Huff, F.A., 1970. Time Distribution Characteristics of Rainfall Rates, Water Resources Research 6, 447-454.
31. Huff, F.A., 1990. Time Distributions of Heavy Rainstorms in Illinois; Circular. Illinois StateWater Survey 173: Champaign, IL, USA, 1990; Available online: http://hdl.handle.net/2142/94492
32. Pani, E.A., Haragan, D.R., 1981. A comparison of Texas and Illinois temporal rainfall distributions. In Proceedings of the Fourth Conference on Hydrometeorology, Reno, NV, USA, 7–9 October 1981; American Meteorological Society: Boston, MA, USA, 1981; pp. 76–80.
33. Bonta, J.V., Rao, A.R., 1987. Factors affecting development of Huff curves. Trans. ASAE 30, 1689–1693.
34. Bonta, J.V., 2004. Development and utility of Huff curves for disaggregating precipitation amounts. Appl. Eng. Agric. 20, 641–652.
35. Terranova, O.G., Iaquinta, P., 2011. Temporal properties of rainfall events in Calabria (southern Italy). Nat. Hazard. Earth Syst. 11, 751–757.
36. Elfeki, A.M., Ewea, H.A., Al-Amri, N.S., 2014. Development of storm hyetographs for flood forecasting in the Kingdom of Saudi Arabia. Arab. J. Geosci. 7, 4387–4398.

37. Pan, C., Wang, X., Liu, L., Huang, H., Wang, D., 2017. Improvement to the Huff Curve for Design Storms and Urban Flooding Simulations in Guangzhou, China. Water 9, 411.
38. McCuen, R. H., 1986. Hydrologic Analysis and Design. Prentice Hall.
39. Licznar, P., Szeląg B., 2014. Analiza zmienności czasowej opadów atmosferycznych w Warszawie. Ochrona Środowiska 36, 23–28.
40. Licznar, P., Burszta-Adamiak, E., Łomotowski, J., Stańczyk, J., 2017. Modern proposal of methodology for retrieval of characteristic synthetic rainfall hyetographs. E3S Web of Conferences 22, 00104 (2017) DOI: 10.1051/e3sconf/20172200104 (E3S Web of Conferences, Volume 22 (2017), International Conference on Advances in Energy Systems and Environmental Engineering (ASEE17), Wrocław, Poland, July 2-5, 2017, B. Kaźmierczak, M. Kutyłowska, K. Piekarska, H. Jouhara and J. Danielewicz (Eds.))
41. Licznar, P., 2018. Analiza opadów atmosferycznych na potrzeby projektowania systemów odwodnienia. Monografie Komitetu Inżynierii Środowiska PAN 137, Wrocław 2018, pp 209.
42. Wartalska, K., Kaźmierczak, B., Nowakowska, M., Kotowski, A., 2020. Analysis of Hyetographs for Drainage System Modeling. Water 12, 149.
43. Burszta-Adamiak, E., Licznar, P., Zaleski, J., 2019. Criteria for identifying maximum rainfalls determined by the peaks-over-threshold (POT) method under the Polish Atlas of Rainfalls Intensities (PANDa) project. Meteorology, Hydrology and Water Management 7, 3-13.
44. Licznar, P., Siekanowicz, K., Stach, A., Zaleski, J., 2020. Atlasy opadowe. W: Metodyka opracowania polskiego atlasu natężeń deszczów (PANDa) / pod red. Pawła Licznara i Janusza Zaleskiego. Warszaw: Instytut Meteorologii i Gospodarki Wodnej. Państwowy Instytut Badawczy 99-137.

45. Larose, D.T, 2005. Discovering knowledge in data. An Introduction to Data Mining. John Wiley & Sons, Inc., Hoboken, New Jersey and Canada.
46. Stanisz, A., 2007. Przystępny Kurs Statystyki z Zastosowaniem STATISTICA PL na Przykładach z Medycyny. Tom I-III; StatSoft Polska Sp. z o.o.: Kraków, Poland.
47. Licznar, P., De Michele, C., Adamowski, W., 2015. Precipitation variability within an urban monitoring network via microcanonical cascade generators, Hydrol. Earth Syst. Sci. 19, 485-506.
48. Gordon, A.D., 1999. Classification. Chapman and Hall/CDC, London.
49. Everitt, B.S., Landau, S., Leese, M., 2001. Cluster analysis. Edward Arnold, London.
50. Dzimińska, P., Drzewiecki, S., Ruman, M., Kosek, K., Mikołajewski, K., Licznar, P., 2021. The Use of Cluster Analysis to Evaluate the Impact of COVID-19 Pandemic on Daily Water Demand Patterns. Sustainability 13, 5772.
51. Walesiak, M., Gatnar, E., 2009. Analiza Skupień. Statystyczna Analiza Danych z Wykorzystaniem Programu R; Wydawnictwo Naukowe PWN: Warszawa, Poland.

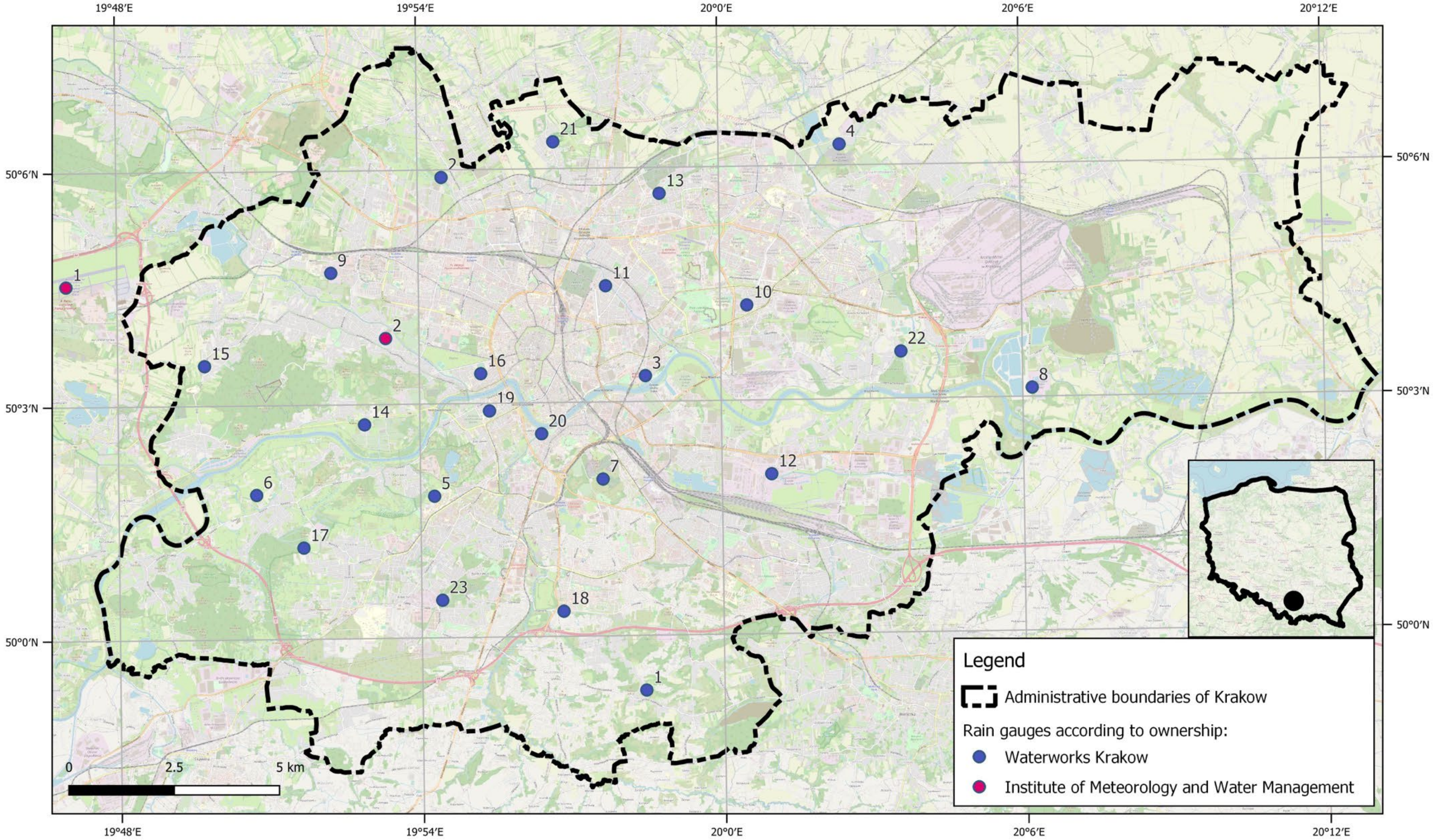


Figure 1. Location of rain gauges belonging to Waterworks Kraków and the Institute of Meteorology and Water Management

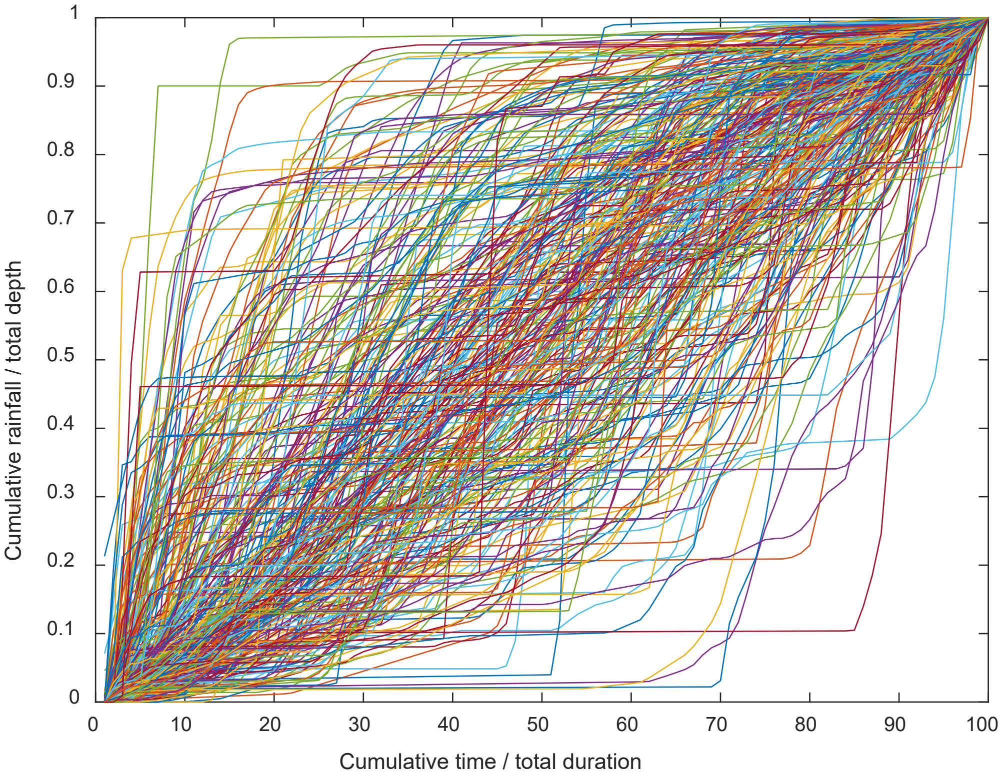


Figure 2. Cumulative dimensionless hyetographs of 313 storm rainfalls from Kraków (set No. 1)

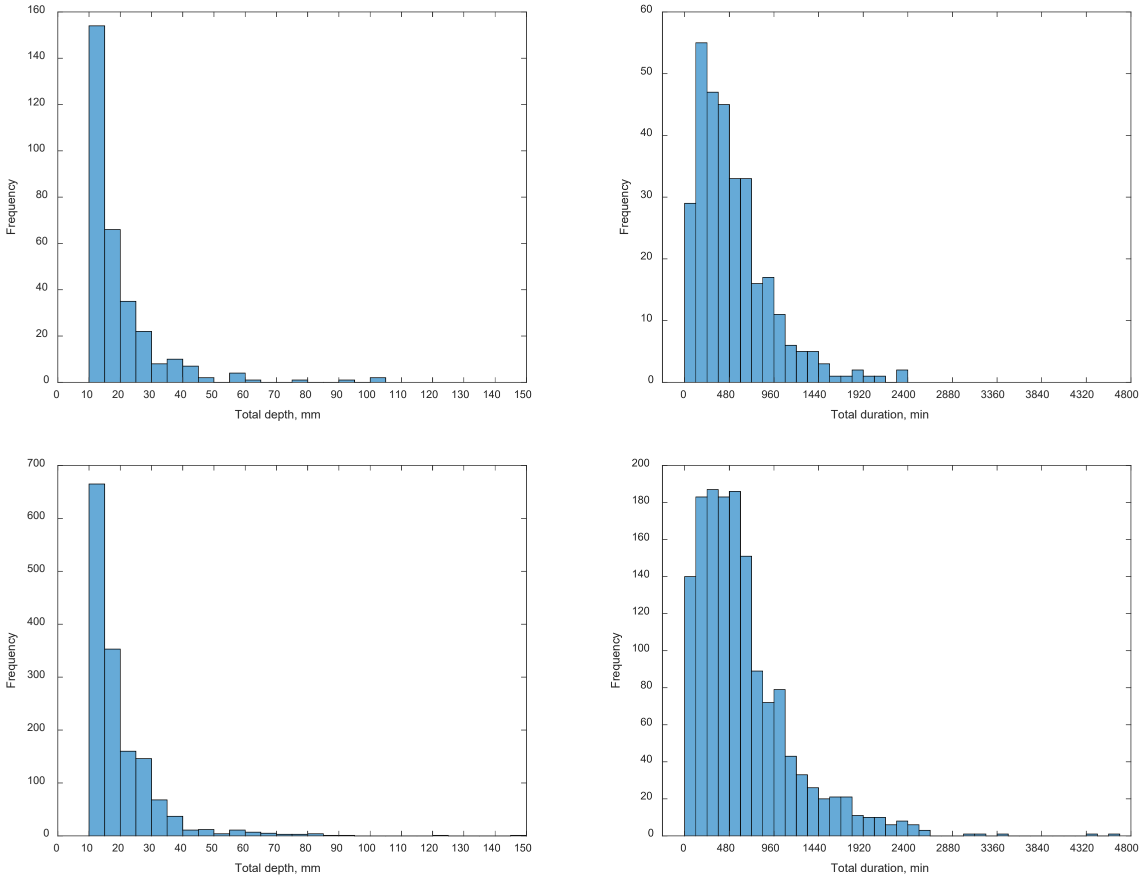
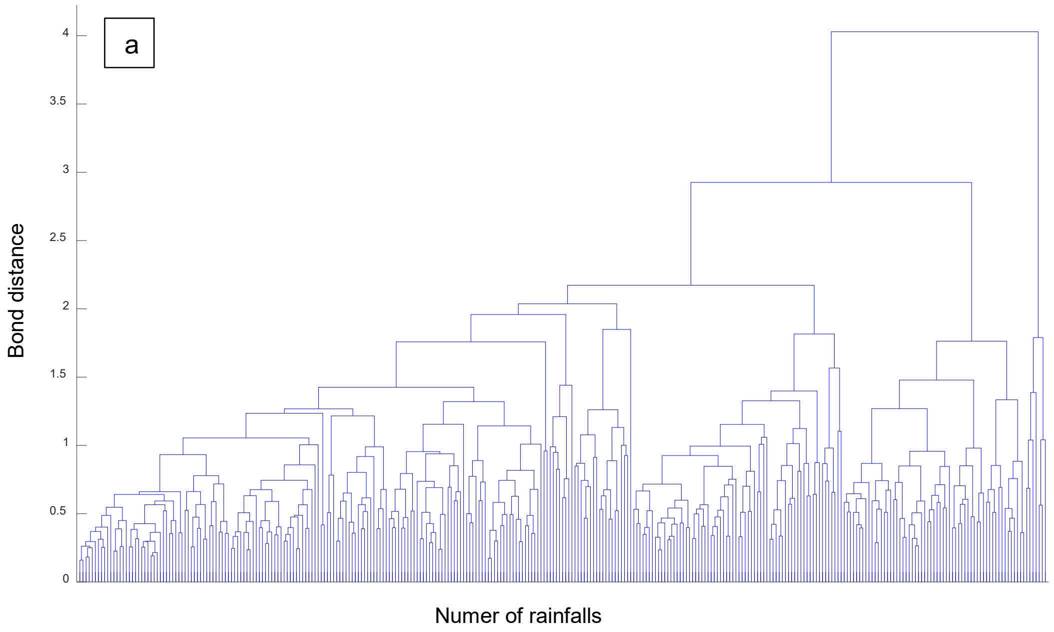
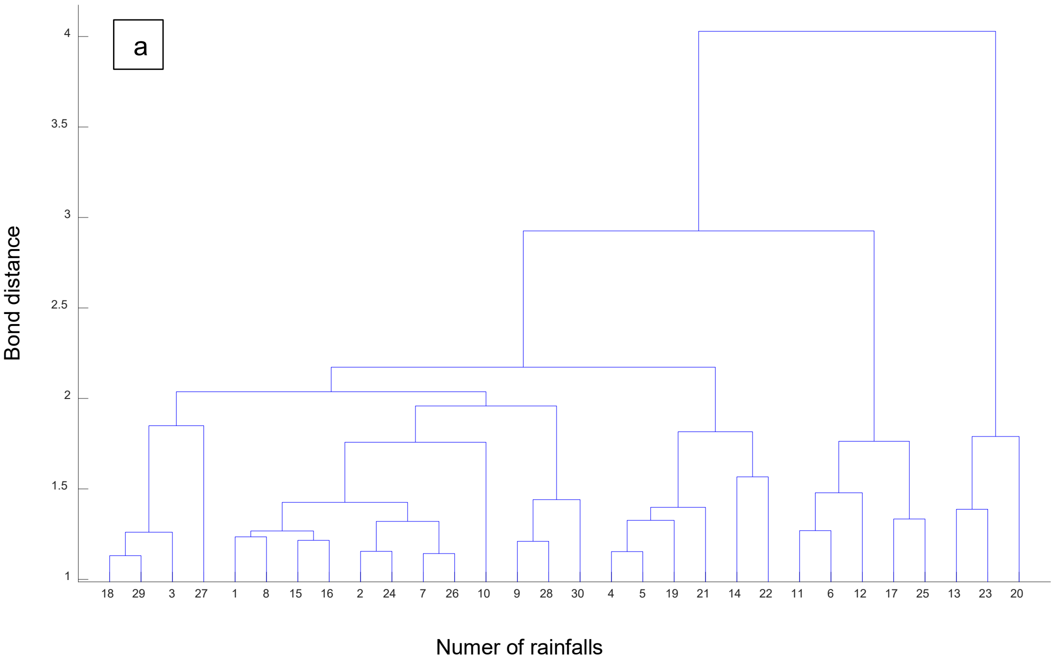


Figure 3. Histograms of rainfall depths and durations for the designated sets of rainfalls from Kraków, diagrams in the top row for set No. 1 (313 rainfalls), diagrams in the bottom row for set No. 2 (1493 rainfalls)


a
Bond distance
4
3.5
3
2.5
2
1.5
1
0.5
0
Numer of rainfalls


a
Bond distance
4
3.5
3
2.5
2
1.5
1
18 29 3 27 1 8 15 16 2 24 7 26 10 9 28 30 4 5 19 21 14 22 11 6 12 17 25 13 23 20
Numer of rainfalls

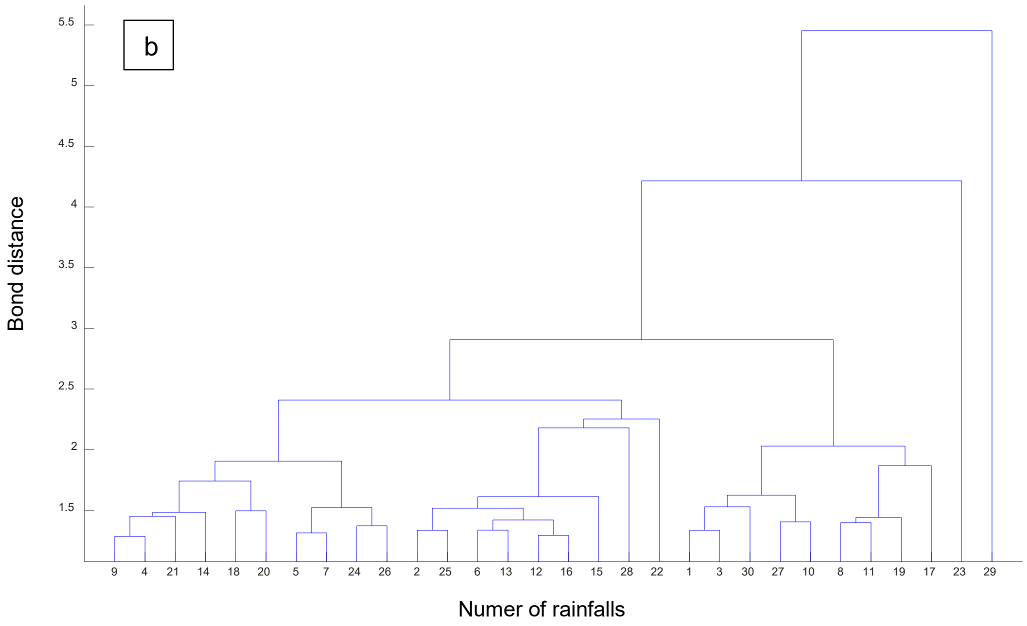


Figure 4 a. Dendrograms obtained for set No. 1 composed of 313 dimensionless cumulative rainfall hyetographs from Kraków (top panel). Below find the same dendrogram prepared for the reduced number of 30 leaf nodes (bottom panel). In the diagrams, vertical axes show bond distances for particular rainfalls and rainfall clusters. The horizontal axis of the dendrogram on the bottom panel shows numbers of rainfalls in particular clusters; **b.** Dendrogram obtained for set No. 2 composed of 1494 dimensionless cumulative rainfall hyetographs from Kraków. The dendrogram was prepared for the reduced number of 30 leaf nodes. The vertical axis of the diagram shows bond distances for particular rainfall clusters, and the horizontal axis shows their numbers

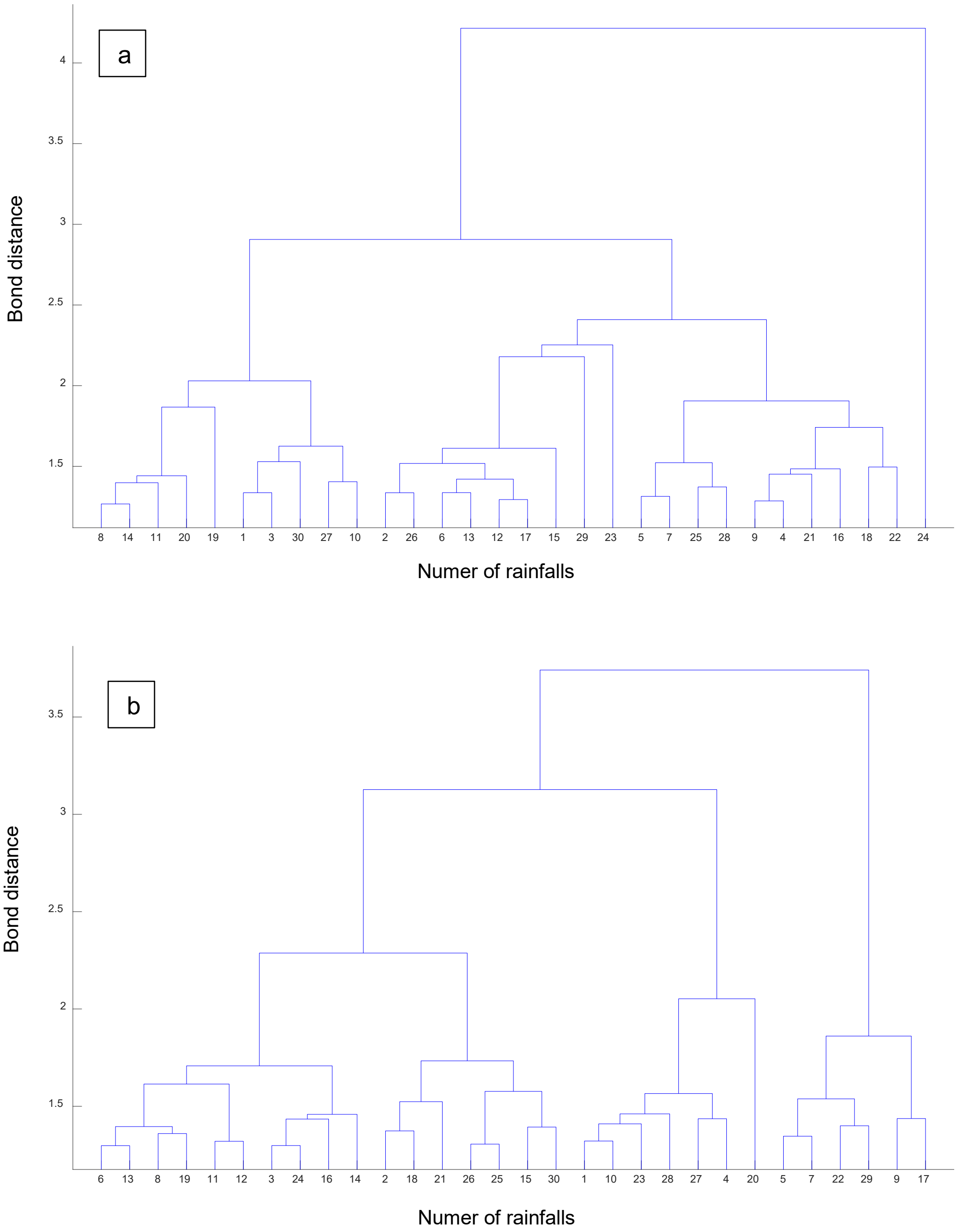


Figure 5 a. Dendrogram obtained for adjusted set No. 2 composed of 1493 dimensionless cumulative rainfall hyetographs from Kraków. The dendrogram was prepared for the reduced number of 30 leaf nodes. The vertical axis of the diagram shows bond distances for particular rainfall clusters, and the horizontal axis shows their numbers; **b.** Dendrogram obtained for adjusted set No. 3 composed of 1806 dimensionless cumulative rainfall hyetographs from Kraków. The dendrogram was prepared for the reduced number of 30 leaf nodes. The vertical axis of the diagram shows bond distances for particular rainfall clusters, and the horizontal axis shows their numbers

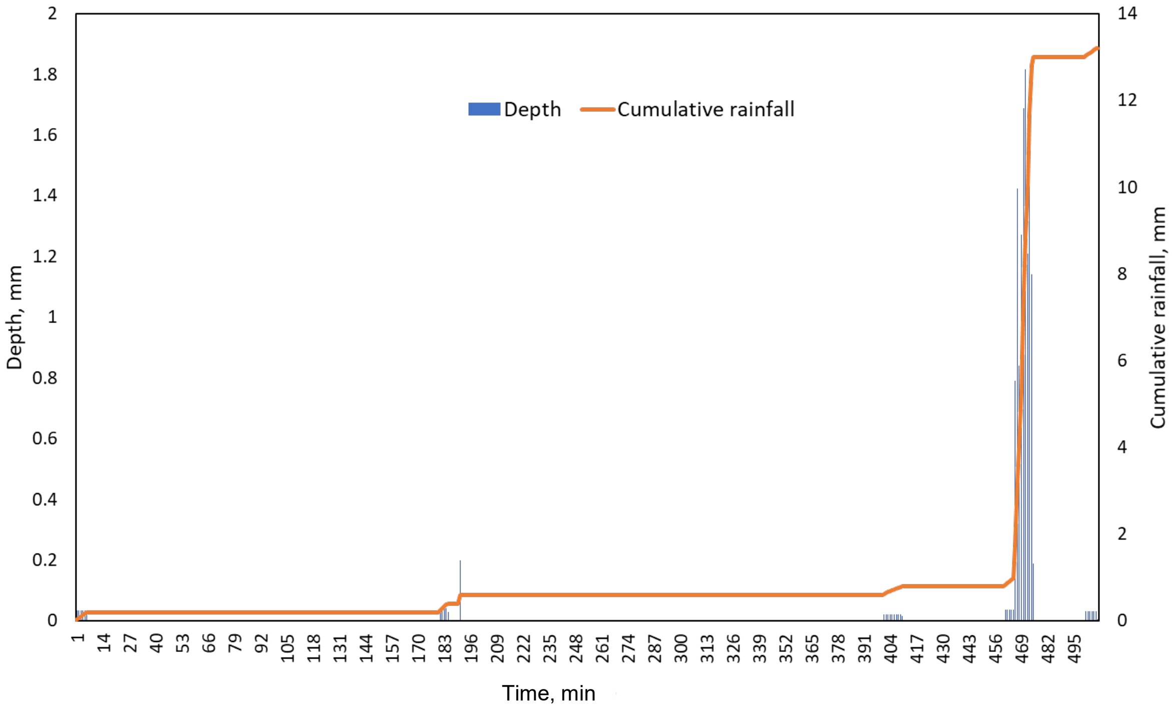


Figure 6. Hyetograph of a rainfall event recorded on station Kraków-Płaszów in 2014

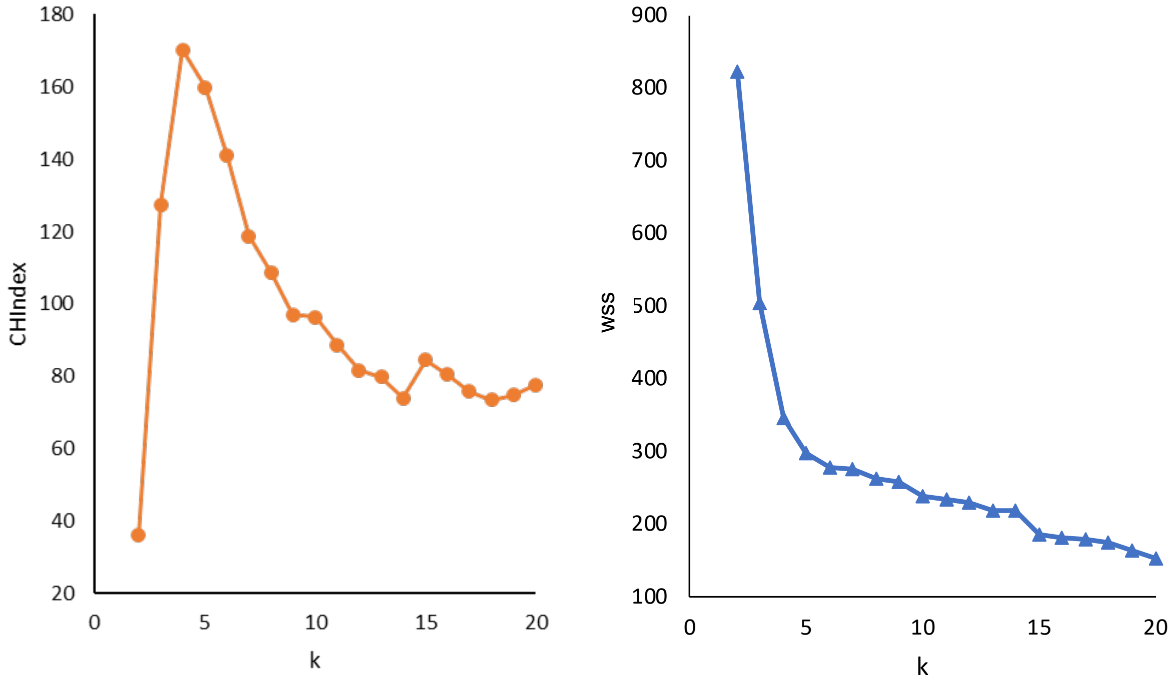


Figure 7. Value of the CHIndex and total within sum of squares (wss), for a set 1 of 313 rainfalls from Kraków, depending on the adopted number of clusters k

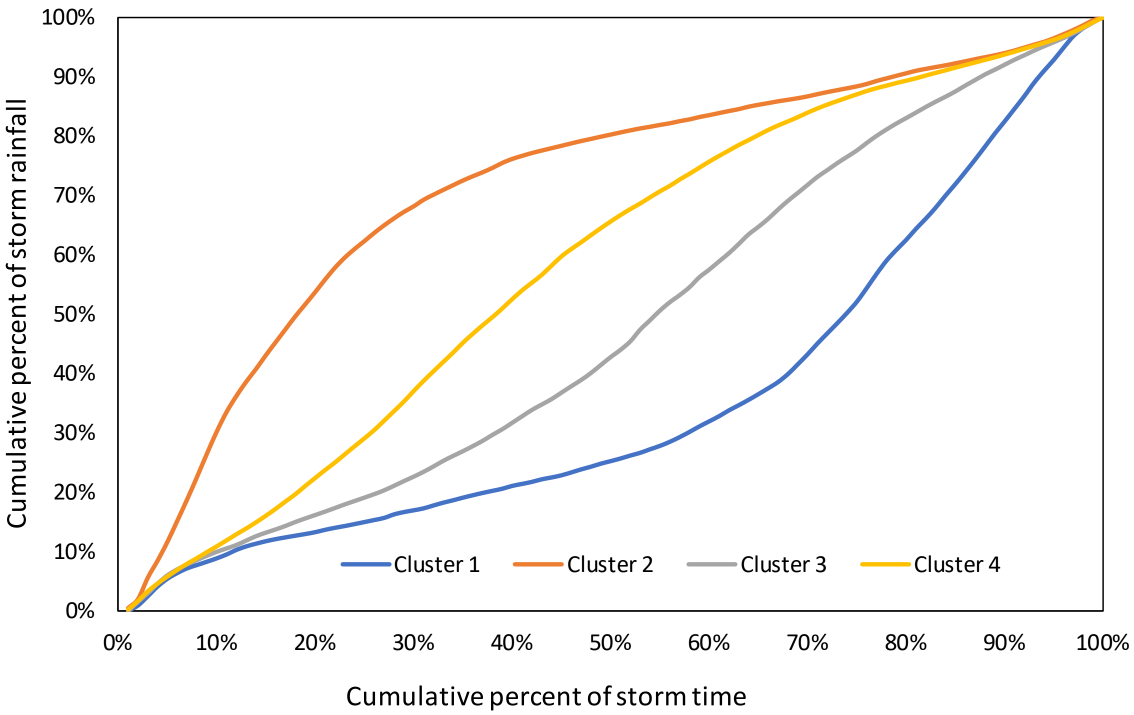


Figure 8. Diagrams of averaged dimensionless cumulative rainfall hyetographs for four clusters designated by means of k-means clustering for Kraków based on set No. 1 for 313 storm rainfalls

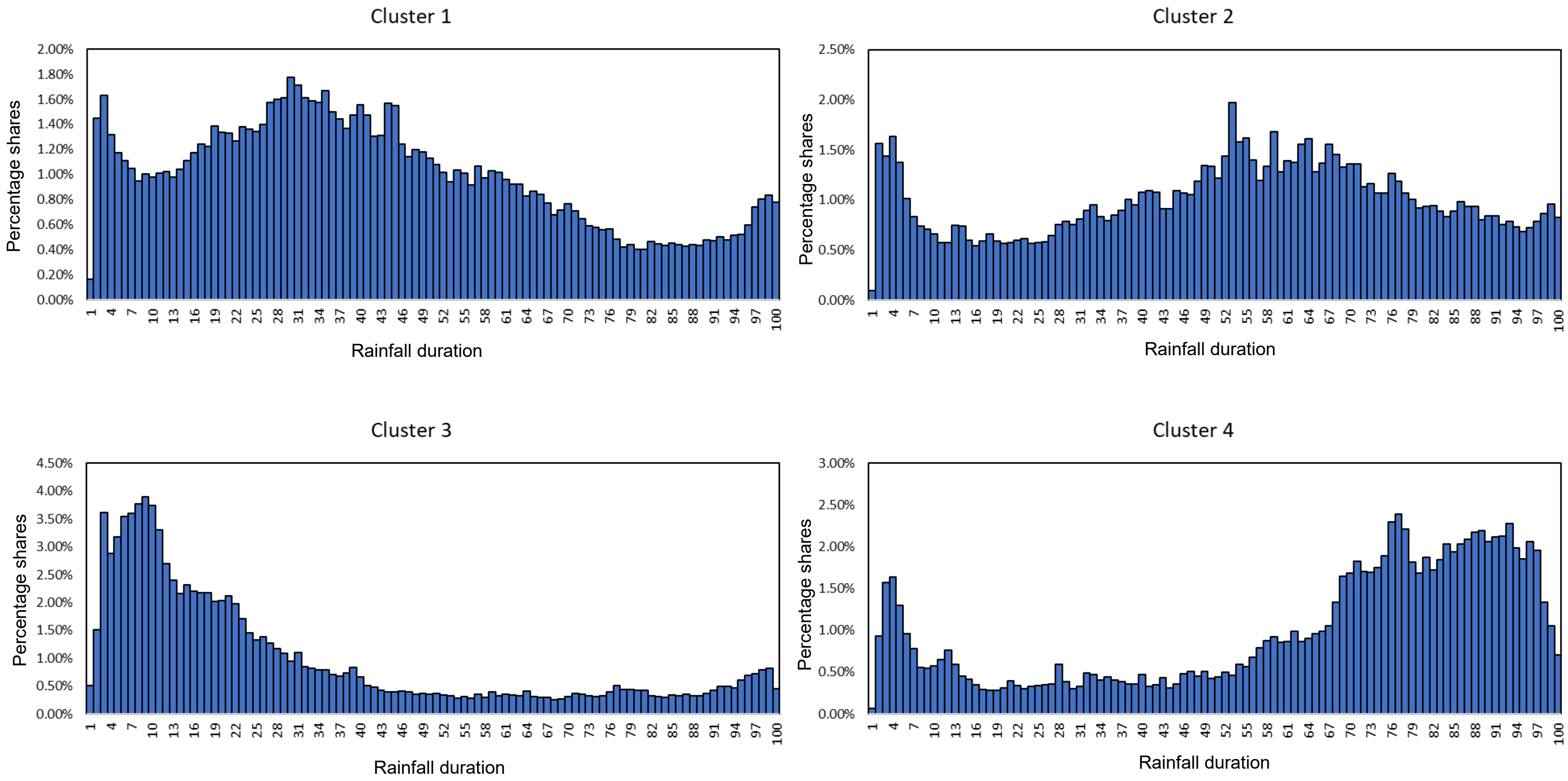


Figure 9. Model dimensionless rainfall hyetographs developed by means of the k-means clustering method for set No. 1 for 313 rainfalls from Kraków. The horizontal axis shows percent increase in rainfall duration, and the vertical axis shows percent shares in total precipitation amount

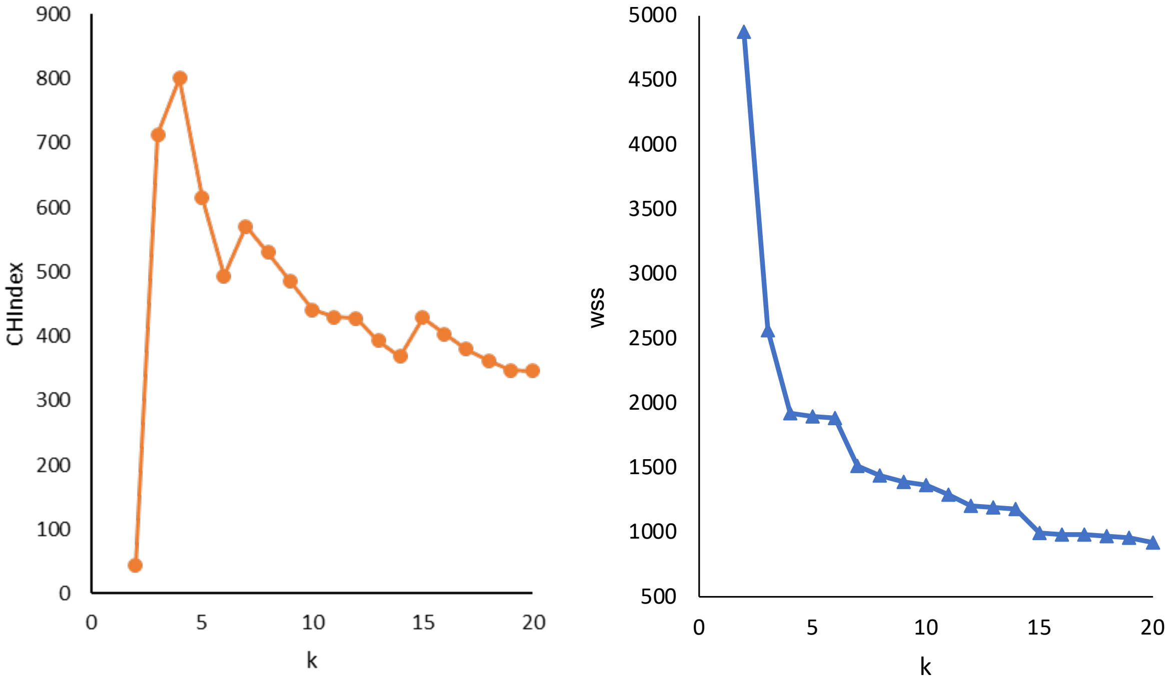


Figure 10. Value of the CHIndex and total within sum of squares (wss), for set No. 2 of 1493 rainfalls from Kraków, depending on the adopted number of clusters k

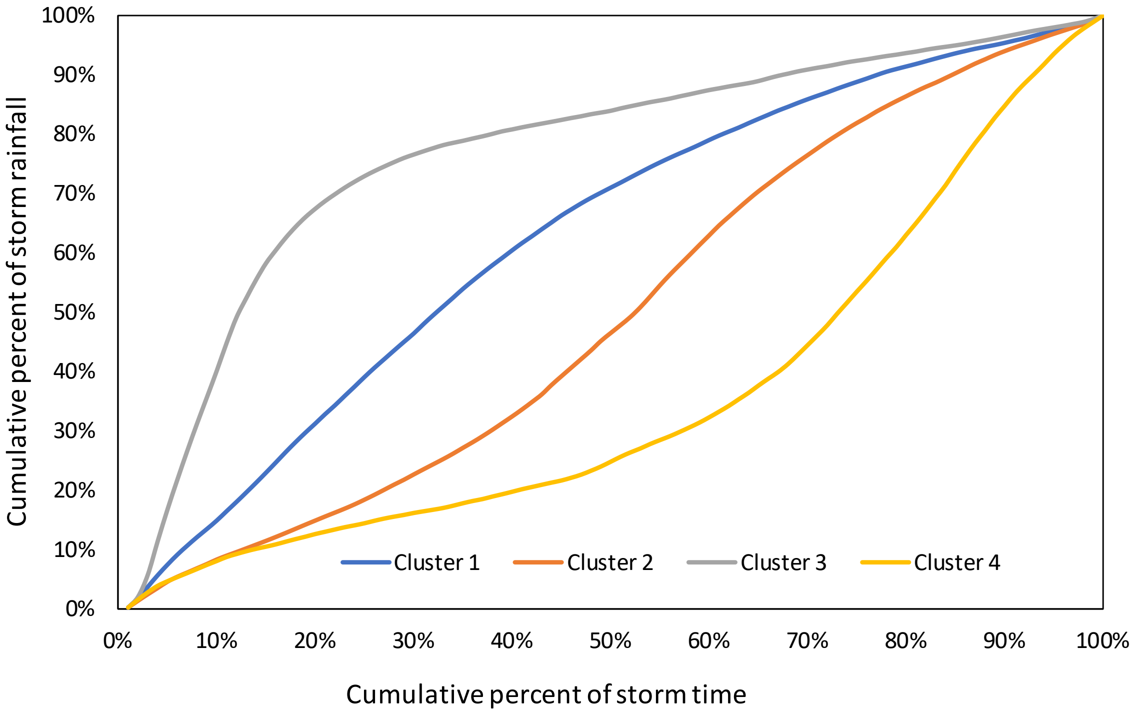


Figure 11. Diagrams of averaged dimensionless cumulative rainfall hyetographs for four clusters designated by means of the k-means clustering method for Kraków based on set No. 2 for 1493 storm rainfalls

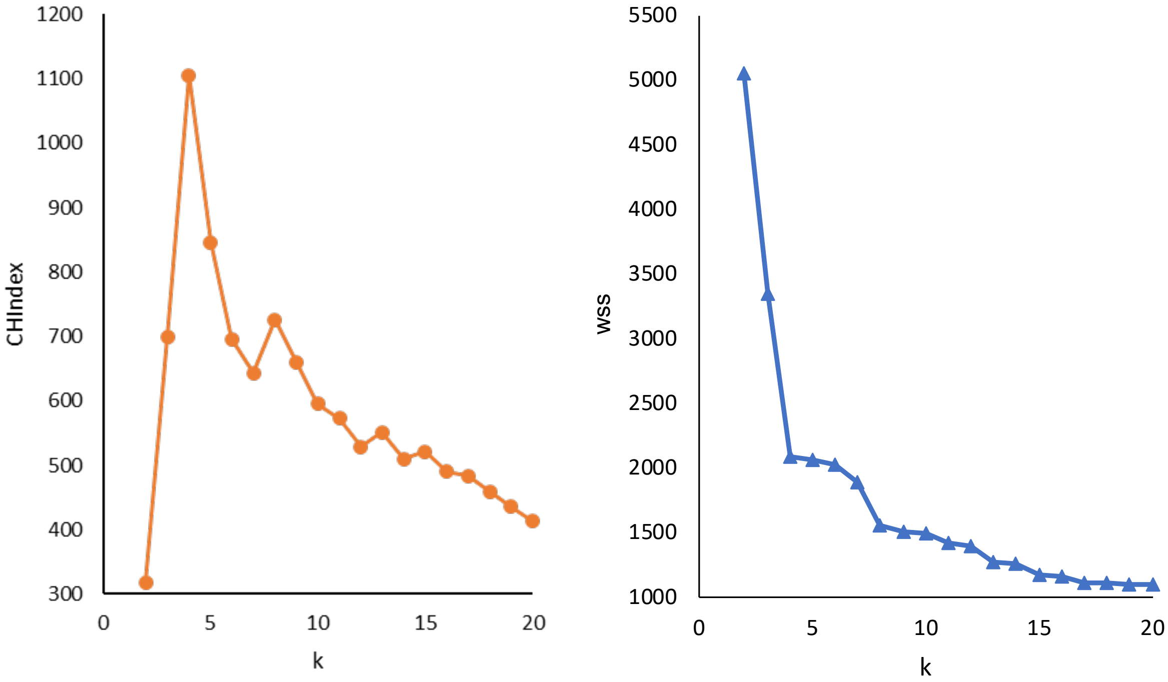


Figure 12. Value of the CHIndex and total within sum of squares (wss), for a set 3 of 1806 rainfalls from Kraków, depending on the adopted number k of clusters

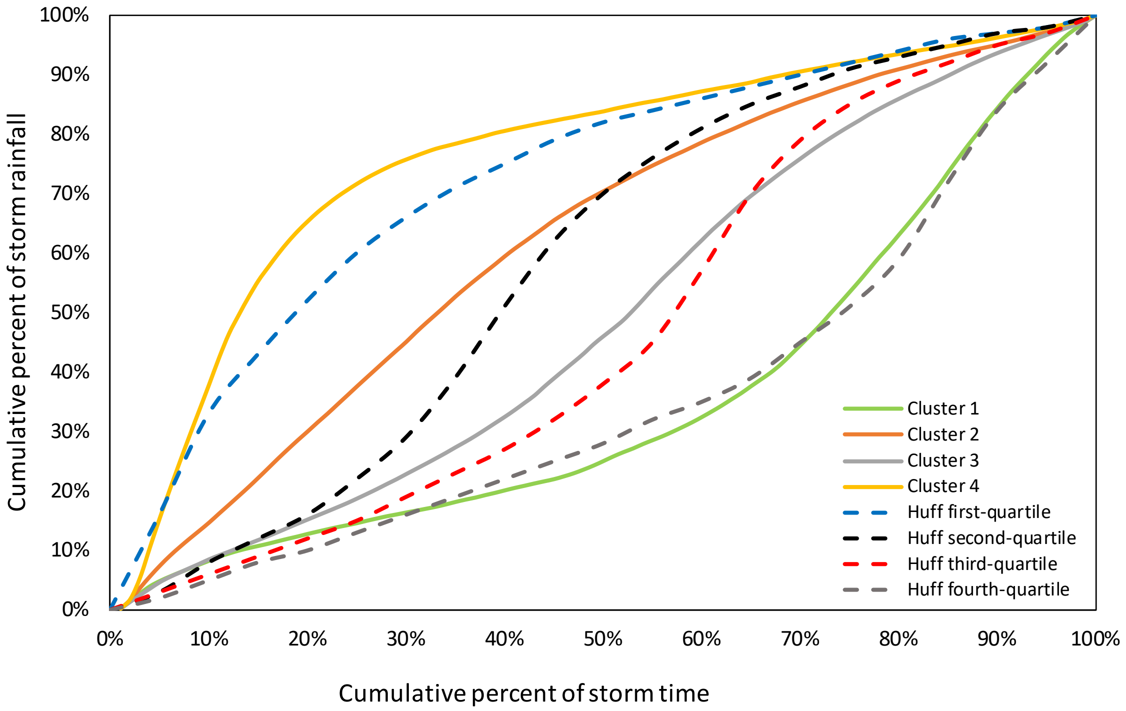


Figure 13. Diagrams of averaged dimensionless cumulative rainfall hyetographs for four clusters designated by means of the k-means clustering method for Kraków based on set No. 3 for 1806 storm rainfalls. For comparison, the diagram also shows Median Time Distributions of Heavy Storm Rainfall at a Point developed by Huff (1990)

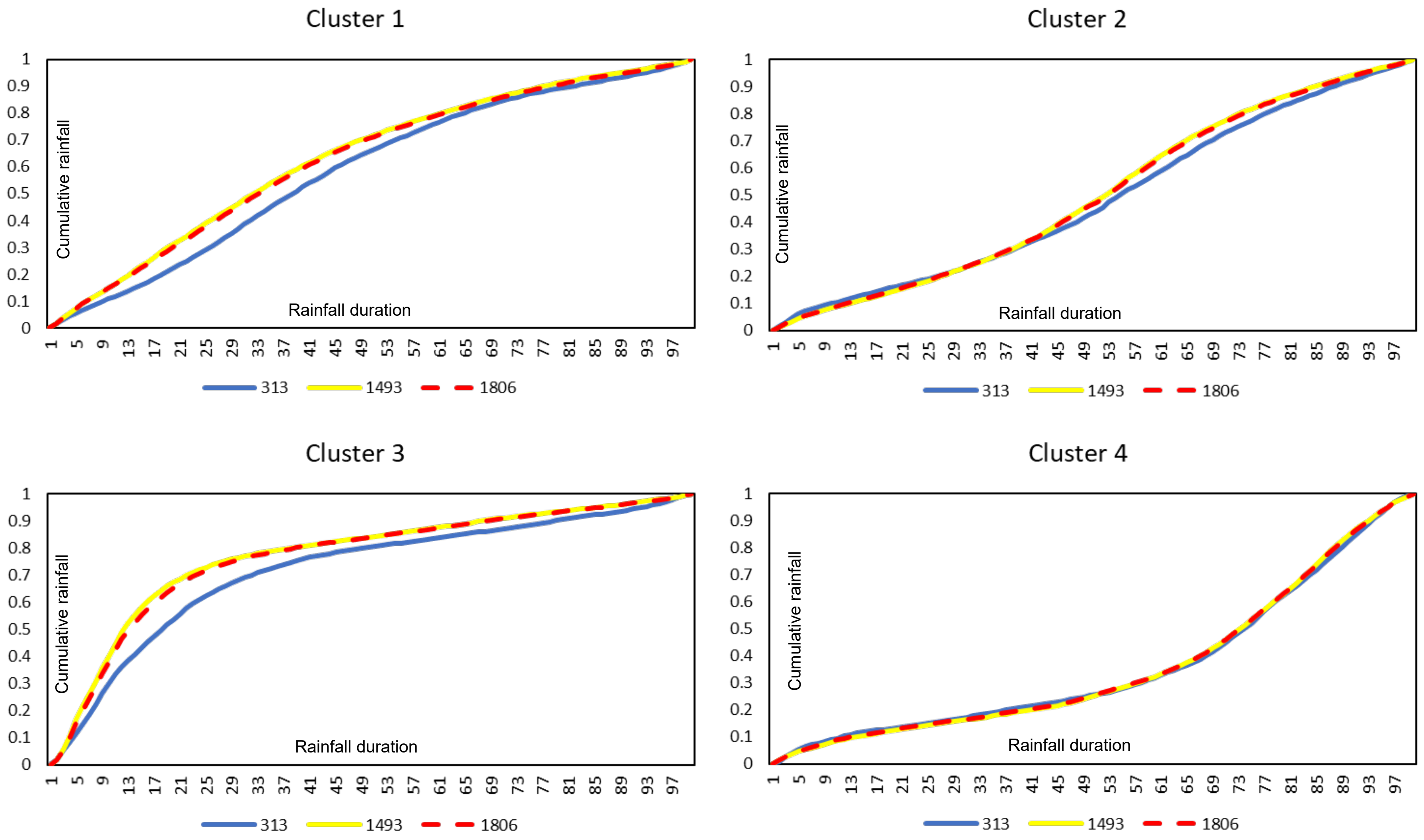
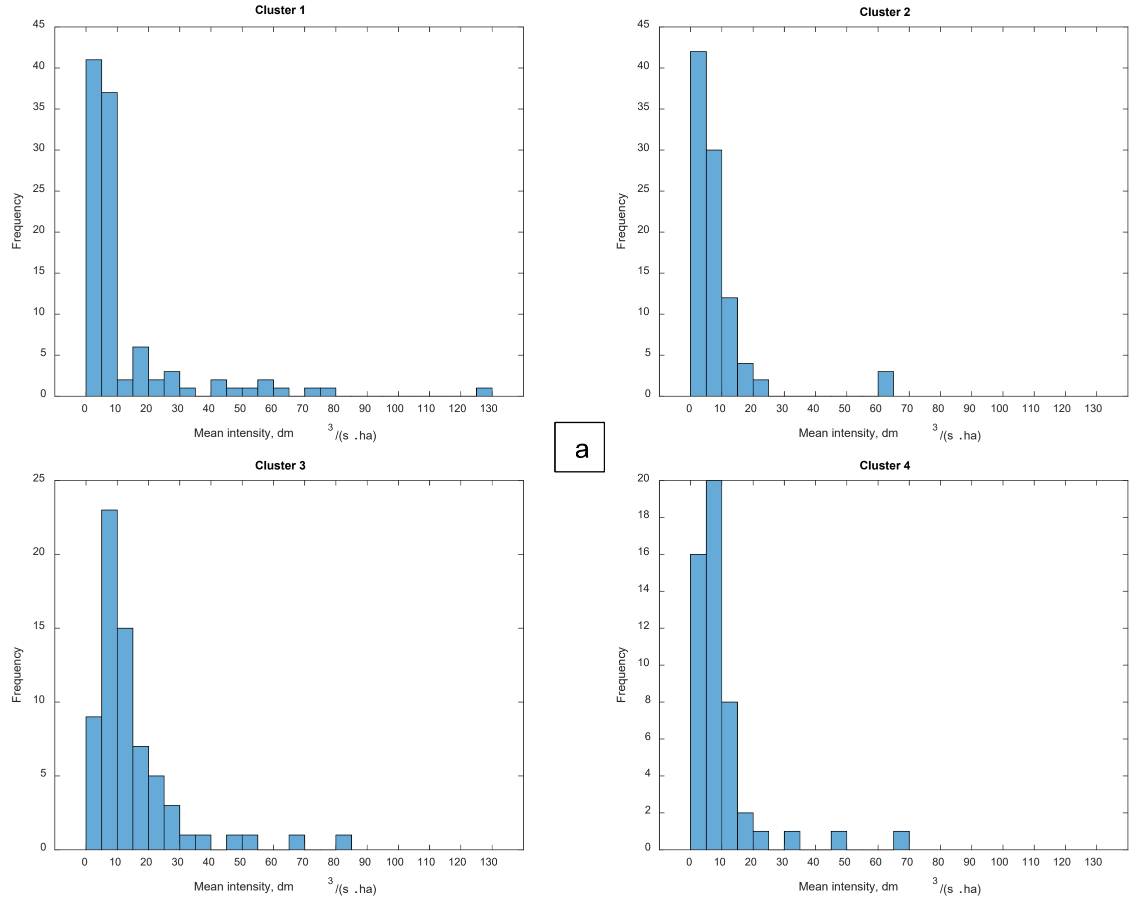

Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cumulative rainfall
Rainfall duration
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1
5
9
13
17
21
25
29
33
37
41
45
49
53
57
61
65
69
73
77
81
85
89
93
97
313
1493
1806

Figure 14. Comparison of averaged dimensionless cumulative rainfall hyetographs for four clusters designated by means of the k-means clustering method based on sets No. 1, 2, and 3: 313, 1493, and 1806 storm rainfalls recorded in Kraków. The horizontal axes present percent increase in rainfall duration, and the vertical axes show cumulative rainfall depth


Cluster 1
Cluster 2
a
Cluster 3
Cluster 4
Frequency
Mean intensity, dm^3/(s .ha)
0
10
20
30
40
50
60
70
80
90
100
110
120
130

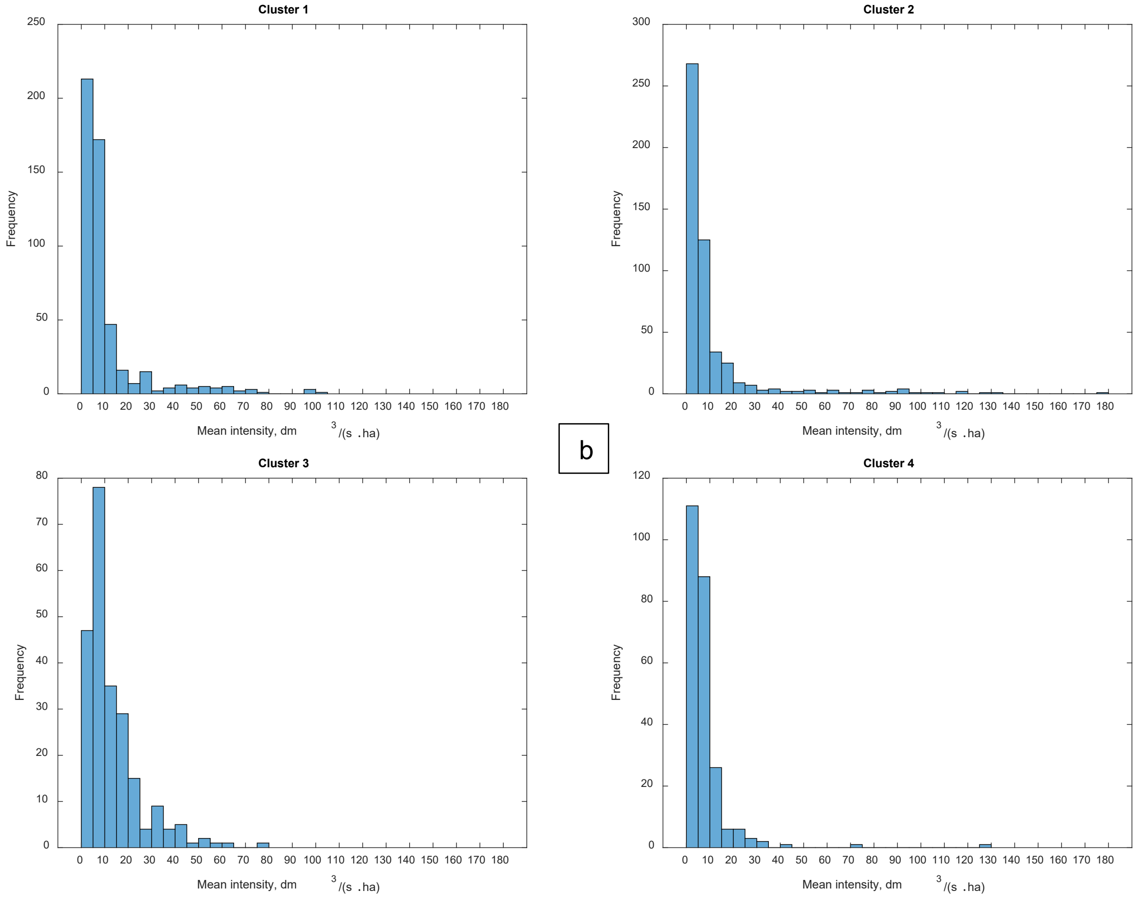

Cluster 1
Cluster 2
b
Cluster 3
Cluster 4
Frequency
Mean intensity, dm^3/(s .ha)
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180

Figure 15 a. Histograms of mean rainfall intensities for four designated clusters in rainfall set No. 1 from Kraków (313 rainfalls); **b.** Histograms of mean rainfall intensities for four designated clusters in rainfall set No. 2 from Kraków (1493 rainfalls)

Table 1. List of rain gauges belonging to sets No. 1 and 2 with characteristics of designated storms

No.	Rain gauge location	Observation period	Number of storm rainfalls	Minimum duration, min.	Maximum duration, min.	Minimum precipitation amount, mm	Maximum precipitation amount, mm
Set No. 1							
1	Kraków – Balice	1986 – 2006	313	24	2329	10.0	105.0
2	Kraków – Wola Justowska	2007 – 2015					
Set No. 2							
1	Bełzy	2016 ÷ 2018	36	89	1639	10.6	74.0
2	Chabrowa	2014 ÷ 2018	41	22	2206	10.2	52.8
3	Miedziana	2010 ÷ 2018	128	20	4359	10.15	121.6
4	Jeziorany	2013 ÷ 2018	86	59	2615	10.0	81.2
5	Kampus UJ	2014 ÷ 2018	67	14	1940	10.2	67.6
6	Kostrze	2013 ÷ 2016	33	49	2377	10.4	43.8
7	Krzemionki	2015 ÷ 2018	59	39	1927	10.0	68.2
8	Kujawy	2013 ÷ 2016	40	32	2258	10.2	93.6
9	Lindego	2008 ÷ 2018	106	21	3015	10.0	60.0
10	Narciarska	2014 ÷ 2018	49	17	2324	10.0	76.0
11	Olsza	2011 ÷ 2012	11	38	1065	10.4	19.4
12	Płaszów	2011 ÷ 2018	100	26	2499	10.0	69.6
13	Reduta	2015 ÷ 2018	49	47	2517	10.0	62.4
14	Rybna	2015 ÷ 2018	43	22	1697	10.2	61.8
15	Rzepichy	2012 ÷ 2018	80	22	2500	10.0	57.2
16	Senatorska	2013 ÷ 2018	81	34	2134	10.2	67.2
17	Skotniki	2009	5	67	839	10.2	24.2
18	Stojałowskiego	2011 ÷ 2018	92	32	3239	10.2	64.8
19	Szwedzka	2008 ÷ 2018	49	21	4576	10.0	145.4
20	Wilga	2013 ÷ 2018	102	25	2518	10.0	73.6
21	Witkowice	2013 ÷ 2018	60	31	2121	10.2	86.0
22	Żaglowa	2012 ÷ 2018	86	22	2405	10.0	82.4

23	Zawiła	2013 ÷ 2018	91	65	2595	10.2	82.2

Table 2. Mean values of total depths, durations, and intensities of storm rainfalls included in particular clusters for sets No. 1 and No. 2 from Kraków

Cluster	Total depth, mm	Total duration, min	Mean intensity, dm^3/(s·ha)
Set No. 1			
1	20.0	577	13.0
2	19.3	598	8.5
3	19.5	348	14.7
4	18.8	505	9.7
Set No. 2			
1	19.6	665	10.8
2	20.4	712	11.3
3	17.5	365	14.2
4	19.9	668	7.7