Marcin Rabiza

# Clustering Case Study

**Table of contents:**

# 1. Introduction

The main goal of this task was to discover an inner structure and regularities in data concerning size of the population of selected counties of Poland with regard to gender and age groups. The regularities of population structure in 2018 and predicted population structure in 2050 were to be compared.

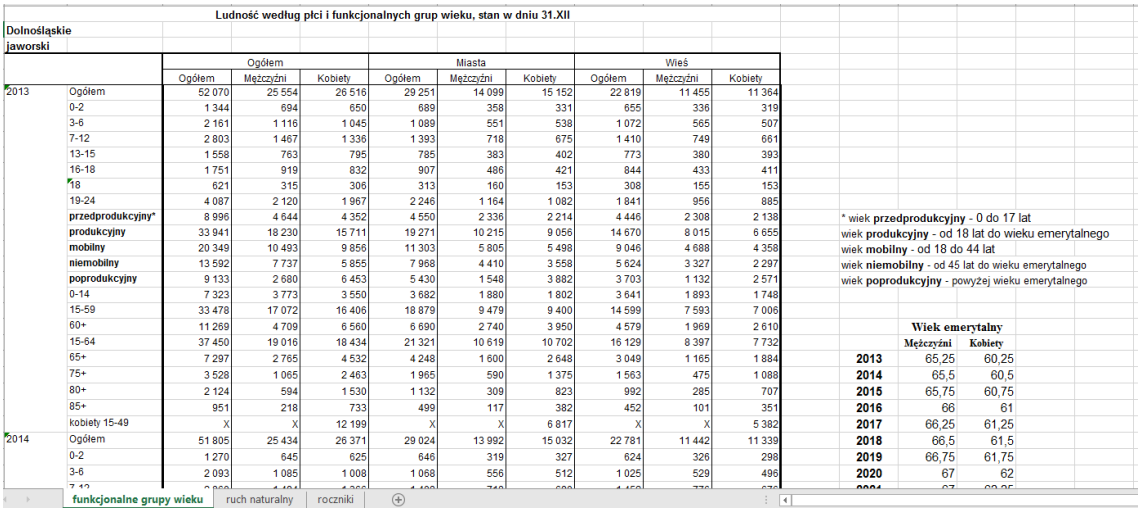
For this task two datasets with the description of population size in different counties were given. One dataset concerns population size for 2018, another is focused on predicted population size for 2050. The datasets contain 8 attributes, among which two are categorical – *county*, *age* (kind of interval also), and the rest is numerical: *city-gen*, *city-men*, *city-women*, *country-gen*, *country-men*, *coutry-women.* As the one may see, the datasets consist information about population age groups, gender and differentiation for cities and country/villages inhabitants. There were 7 polish districts chosen for the study: *powiat poznański, powiat bydgoski, powiat krakowski, powiat gdański, powiat wrocławski, powiat warszawski zachodni, powiat rzeszowski*.

In order to discover "natural" groupings in data and organize instances them into similar clusters, an unsupervised learning algorithms of cluster analysis were performed.

# 2. Data preprocessing

As a first step of the analysis, a basic data preprocessing was performed in order to check the quality of the data and clean the data if necessary.

Original datasets (eg. on Fig. 1) comes from Central Office of Statistics of Poland. The datesets consist information for measured and predicted population features of each county since 2013 till 2050. From this range, only 2018 and 2050 were chosen for comparison. From the datasets, missing records, functional age groups records (*przedprodukcyjny*, *produkcyjny*) and general statistics of cities and counties inhabitants altogether were removed manually.



Ludność według płci i funkcjonalnych grup wieku, stan w dniu 31.XII

Dolnośląskie
jaworski

| | | Ogółem | | | Miasta | | | Wieś | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Ogółem | Mężczyźni | Kobiety | Ogółem | Mężczyźni | Kobiety | Ogółem | Mężczyźni | Kobiety |
| 2013 | Ogółem | 52 070 | 25 554 | 26 516 | 29 251 | 14 099 | 15 152 | 22 819 | 11 455 | 11 364 |
| | 0-2 | 1 344 | 694 | 650 | 689 | 358 | 331 | 655 | 336 | 319 |
| | 3-6 | 2 161 | 1 116 | 1 045 | 1 089 | 551 | 538 | 1 072 | 565 | 507 |
| | 7-12 | 2 803 | 1 467 | 1 336 | 1 393 | 718 | 675 | 1 410 | 749 | 661 |
| | 13-15 | 1 558 | 763 | 795 | 785 | 383 | 402 | 773 | 380 | 393 |
| | 16-18 | 1 751 | 919 | 832 | 907 | 486 | 421 | 844 | 433 | 411 |
| | 18 | 621 | 315 | 306 | 313 | 160 | 153 | 308 | 155 | 153 |
| | 19-24 | 4 087 | 2 120 | 1 967 | 2 246 | 1 164 | 1 082 | 1 841 | 956 | 885 |
| | **przedprodukcyjny*** | 8 996 | 4 644 | 4 352 | 4 550 | 2 336 | 2 214 | 4 446 | 2 308 | 2 138 |
| | **produkcyjny** | 33 941 | 18 230 | 15 711 | 19 271 | 10 215 | 9 056 | 14 670 | 8 015 | 6 655 |
| | **mobilny** | 20 349 | 10 493 | 9 856 | 11 303 | 5 805 | 5 498 | 9 046 | 4 688 | 4 358 |
| | **niemobilny** | 13 592 | 7 737 | 5 855 | 7 968 | 4 410 | 3 558 | 5 624 | 3 327 | 2 297 |
| | **poprodukcyjny** | 9 133 | 2 680 | 6 453 | 5 430 | 1 548 | 3 882 | 3 703 | 1 132 | 2 571 |
| | 0-14 | 7 323 | 3 773 | 3 550 | 3 682 | 1 880 | 1 802 | 3 641 | 1 893 | 1 748 |
| | 15-59 | 33 478 | 17 072 | 16 406 | 18 879 | 9 479 | 9 400 | 14 599 | 7 593 | 7 006 |
| | 60+ | 11 269 | 4 709 | 6 560 | 6 690 | 2 740 | 3 950 | 4 579 | 1 969 | 2 610 |
| | 15-64 | 37 450 | 19 016 | 18 434 | 21 321 | 10 619 | 10 702 | 16 129 | 8 397 | 7 732 |
| | 65+ | 7 297 | 2 765 | 4 532 | 4 248 | 1 600 | 2 648 | 3 049 | 1 165 | 1 884 |
| | 75+ | 3 528 | 1 065 | 2 463 | 1 965 | 590 | 1 375 | 1 563 | 475 | 1 088 |
| | 80+ | 2 124 | 594 | 1 530 | 1 132 | 309 | 823 | 992 | 285 | 707 |
| | 85+ | 951 | 218 | 733 | 499 | 117 | 382 | 452 | 101 | 351 |
| | kobiety 15-49 | X | X | 12 199 | X | X | 6 817 | X | X | 5 382 |
| 2014 | Ogółem | 51 805 | 25 434 | 26 371 | 29 024 | 13 992 | 15 032 | 22 781 | 11 442 | 11 339 |
| | 0-2 | 1 270 | 645 | 625 | 646 | 319 | 327 | 624 | 326 | 298 |
| | 3-6 | 2 093 | 1 085 | 1 008 | 1 068 | 556 | 512 | 1 025 | 529 | 496 |

* wiek **przedprodukcyjny** - 0 do 17 lat
wiek **produkcyjny** - od 18 lat do wieku emerytalnego
wiek **mobilny** - od 18 do 44 lat
wiek **niemobilny** - od 45 lat do wieku emerytalnego
wiek **poprodukcyjny** - powyżej wieku emerytalnego

| Wiek emerytalny | | |
|---|---|---|
| | Mężczyźni | Kobiety |
| **2013** | 65,25 | 60,25 |
| **2014** | 65,5 | 60,5 |
| **2015** | 65,75 | 60,75 |
| **2016** | 66 | 61 |
| **2017** | 66,25 | 61,25 |
| **2018** | 66,5 | 61,5 |
| **2019** | 66,75 | 61,75 |
| **2020** | 67 | 62 |

funkcjonalne grupy wieku | ruch naturalny | roczniki

**Figure 1**. Exemplary dataset with population structure prediction from Central Office of Statistics.

Attributes can be correlated – linearly related to each other. Particular associations may lead to data redundancy, which may be detected by correlation analysis. Handling redundant data is beneficial due to fact that larger number of redundant data may slow-down or confuse knowledge discovery process. In this case, correlation analysis may give an answer whether all conditional attributes are necessary to describe the decision attribute.

In order to detect correlated attributes, the chi-squared dependency test was performed, as it resolves the matter of the independence of two variables. Performing chi-square dependency test also enables the analyst to determine a measure of association between two variables, giving a value between 0 and +1 (inclusive) using Cramér's V formula.

Below a null ($H_0$) and an alternative ($H_1$) hypothesises are formulated.

$H_0$: Variables are independent.

$H_1$: Variables are dependent.

The one reject null hypothesis if a calculated $\chi^2$ value is bigger than the upper-tail Critical Value of chi-square distribution for certain number of Degrees of Freedom (df) and accepted Significance Level $\alpha$. In other case, $H_0$ is accepted. An example of chi-squared distribution was given on fig. 1 below.

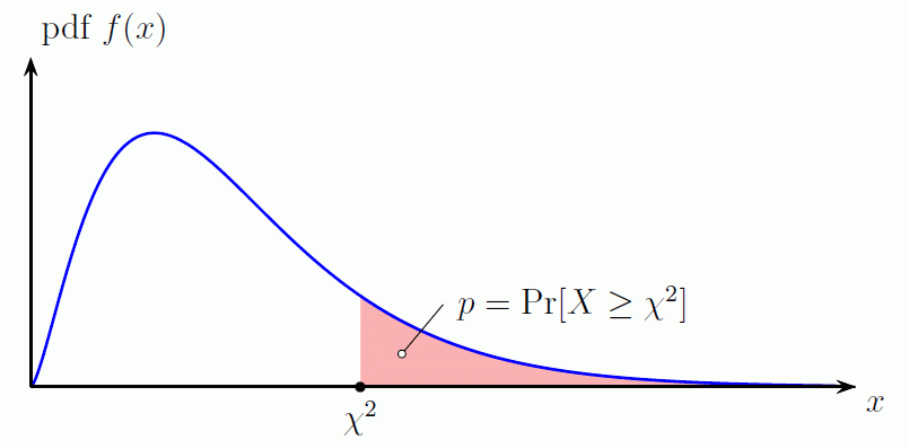



**Fig. 1.** Exemplary $\chi^2$ distribution [2].

The one may see, that for tested variables every chi-squared independence test is associated with a very large number of degrees of freedom. However, for large numbers of degrees of freedom the chi-squared distribiution is almost indistinguishable from normal distribution (fig. 2 below). In that case the one may calculate Critical Values for $\alpha$ based on normal distribution with a good approximation without loosing generality.

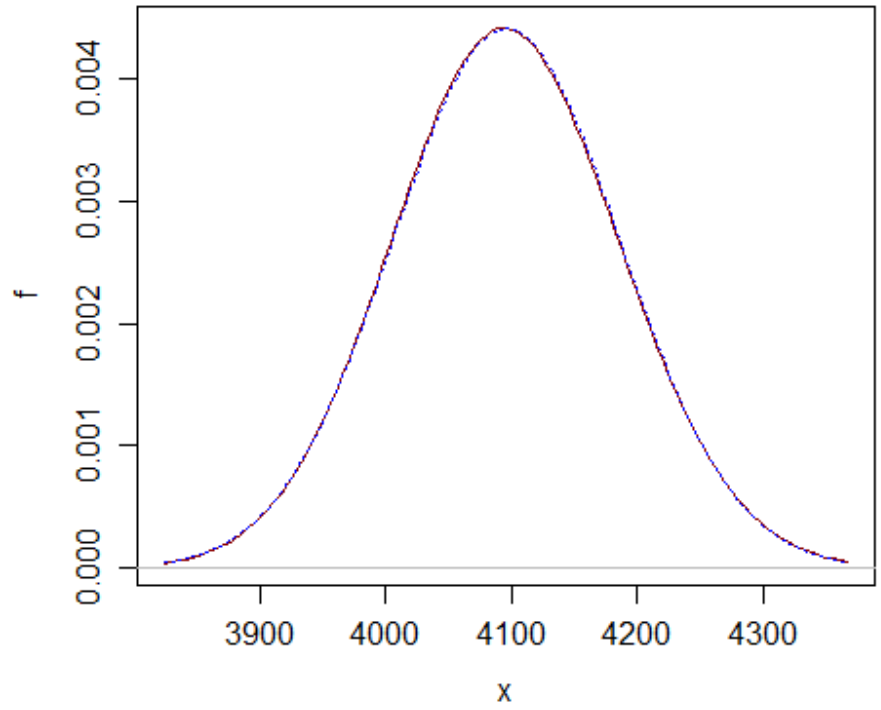



**Fig. 2.** Chi-squared distribution (solid dark red curve) and normal distribution (dotted blue curve) for $2^{12}$ degrees of freedom.

For the needs of this task, chi-square independency tests for pairs of different conditional attributes and one decision attribute and Cramér's V calculations were performed using R programming language and software environment for statistical computing. Simplified R code is given below.

**Simplified R code:**

```
data <- read.csv(file="e:/Powiaty2018.csv", header=TRUE, sep=";")      # Load data
chisq.test(data$City.gen, data$City.men)                               # Calculate χ2 test statistics
DescTools::CramerV(data$City.gen, data$City.men)                       # Calculate Cramér's V
```

Computed statistics are given in table 1 below. The Significance Level for this task was arbitrary set as $\alpha = 0{,}05$. Critical Value for numbers of Degrees of Freedom of different attributes from the given dataset was read from chi-square distribution tables.

**Table 1.** Independence tests statistics for dependent attributes.

| Attributes | $\chi^2$ | df | p-Value | Crit. V. | Reject $H_0$? | Cramér's V |
|---|---|---|---|---|---|---|
| **city-gen / city-men** | 10815 | 10712 | 0,2401 | 1,64 | Yes | 1 |
| **city-gen / city-women** | 10710 | 10609 | 0,2433 | 1,64 | Yes | 0,995 |
| **country-gen / country-men** | 10920 | 10816 | 0,239 | 1,64 | Yes | 1 |
| **country-gen / country-women** | 10815 | 10712 | 0,2401 | 1,64 | Yes | 1 |

Since (1) *city-gen* and (2) *country-gen* attributes were strongly correlated with (1) *city-men*, *city-women* and (2) *country-men*, *country-women*, there were removed from a dataset. Then, the final dataset prepared for cluster analysis consists of 6 attributes and 105 records.

# 3. Cluster analysis

"Cluster analysis" stands for a type of unsupervised learning technique - it finds "natural" grouping of instances given unlabeled data. Clustering is the process of grouping a set of objects in such a way that objects in the same group (called a *cluster*) are more similar[1] to each other than to those in other groups (clusters). Clustering task comes down to partitioning a set of data into a set of meaningful, yet not predefined, sub-classes. The main advantage of using clustering is that it enables analysts to find comprehensible patterns, structures and regularities in raw data.

In terms of evaluation, the one concerns cluster analysis outcome as a "good clustering" if it produces high quality clusters in which:

- the intra-class similarity is high,
- the inter-class similarity is low.

Similarity is usually expressed in terms of a distance function, which is typically metric $d(i, j)$. As such, distance functions are also used in clustering quality evaluation, which estimates the "goodness" of a cluster. Distance functions may vary depending of a type variables and type of application. Exemplary distance measures include Euclidean distance: $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$ or Manhattan distance: $\sum_{i=1}^{k}|x_i - y_i|$.

**HAC Algorithm**

The first cluster analysis was performed using *Hierarchical Agglomerative Clustering* (HAC) algorithm. Hierarchical clustering are methods of cluster analysis which seek to build a hierarchy of clusters – they perform hierarchical decomposition of the set of data using some criterion.

Hierarchical clustering algorithms generate dendrograms – tree-based hierarchical taxonomies – out of a set of unlabeled examples. Dendrogram shows data decomposed into several levels of nested partitioning. Exemplary dendrogram is shown on figure 3 below.

[1] Eg. similar in terms of a distance: Euclidean, Manhattan etc.

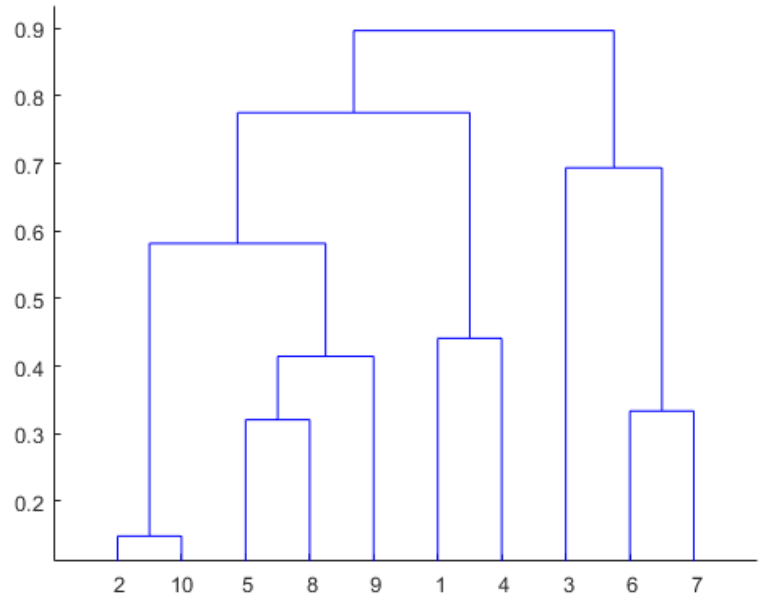



**Fig. 3.** Exemplary dendrogram

A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster. The main advantage of using dendrograms is that they are easily readable and inspectable for humans.

For this task the HAC algorithm is used. The algorithm is *agglomerative*, what indicates "bottom-up" approach – each observation starts in its own cluster and in the next step pairs of clusters are merged as one moves up the hierarchy. Using HAC the one may choose different linkage criteria, which determines the distance between sets of instances as a function of the pairwise distances between instances. Choosing linkage method determines how distance between clusters is measured. The one may choose:

- *single linkage* – distance between clusters is a minimum distance between instances belonging to different clusters,
- *complete linkage* – distance between clusters is a maximum distance between instances belonging to different clusters,
- mean distance – distance between clusters is a distance between means of different clusters
- average distance.

Outcomes of single and complete linkage HAC algorithms generated using Statistica software for the case study are presented below.

**a) HAC algorithm with single linkage and Euclidean distance measure for 2018 and 2050 data comparison.**

On figure 4 the one may inspect dendrograms build for 2018 and 2050 counties population size.

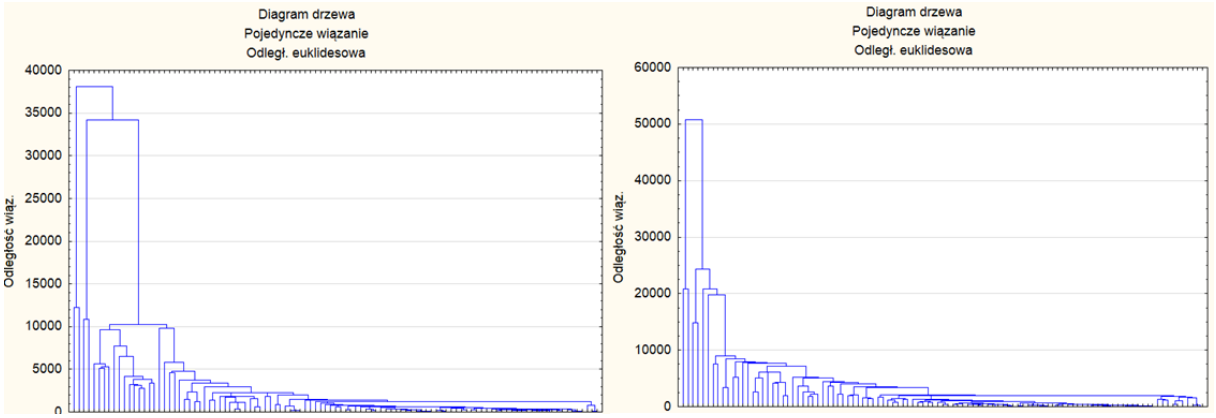



**Fig. 4.** Scaled dendrogram for 2018 (left) and predicted 2050 (right).

While using HAC algorithm, it is important to find a cut point which determines individual cluster. It is advisable to follow the "knee" rule. Cutting points for both 2018 and 2050 data according to "knee" rule are about 5000 distance units. The points are presented on agglomeration charts below.

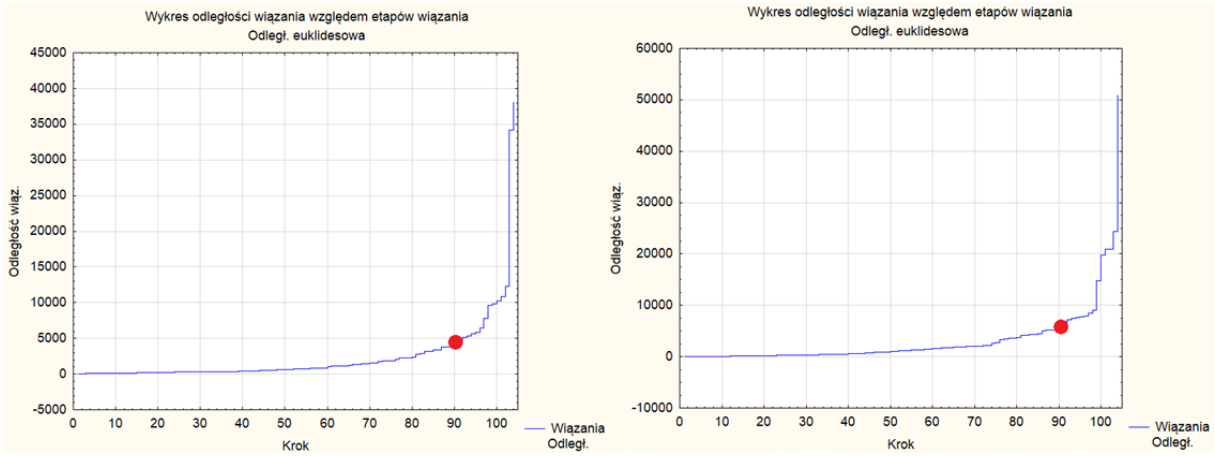



**Fig. 5.** Agglomeration chart for 2018 (left) and 2050 (right) data.

The "knee" points are indicated with red dot.

Clusters discovered by cutting dendrograms at the "knee"-based distance units are listed on figure 6 below.

Przynależność do skupień (Powiaty2018)
Odległość łączenia = 4993,67
Pojedyncze wiązanie
Odległ. euklidesowa

| | Przynależność do skupień | County | Age | City-men | City-women | Country-men | Country-woman |
|---|---|---|---|---|---|---|---|
| C_9 | 1 | poznański | 15-59 | 41530 | 42759 | 76485 | 78814 |
| C_10 | 2 | poznański | 60+ | 12704 | 17637 | 18822 | 23881 |
| C_11 | 3 | poznański | 15-64 | 45849 | 47818 | 83350 | 86457 |
| C_12 | 4 | poznański | 65+ | 8385 | 12578 | 11957 | 16238 |
| C_38 | 5 | krakowski | 0-14 | 3549 | 3251 | 19660 | 18465 |
| C_39 | 6 | krakowski | 15-59 | 13610 | 13704 | 72584 | 72189 |
| C_40 | 7 | krakowski | 60+ | 4842 | 6901 | 21849 | 27719 |
| C_41 | 8 | krakowski | 15-64 | 15158 | 15503 | 80044 | 79757 |
| C_69 | 9 | wrocławski | 15-59 | 6611 | 6905 | 37898 | 37743 |
| C_71 | 10 | wrocławski | 15-64 | 7301 | 7721 | 41536 | 41487 |
| C_99 | 11 | rzeszowski | 15-59 | 9007 | 9037 | 45700 | 43622 |
| C_101 | 12 | rzeszowski | 15-64 | 9956 | 10027 | 49578 | 47461 |
| C_8 | 13 | poznański | 0-14 | 11862 | 11292 | 26033 | 24819 |
| C_24 | 13 | bydgoski | 15-59 | 8526 | 8468 | 28991 | 28517 |
| C_26 | 13 | bydgoski | 15-64 | 9393 | 9494 | 31820 | 31282 |
| C_54 | 13 | gdański | 15-59 | 9319 | 9889 | 27360 | 26941 |
| C_56 | 13 | gdański | 15-64 | 10174 | 10893 | 29755 | 29429 |
| C_84 | 13 | awski zach. | 15-59 | 11936 | 12445 | 22482 | 22878 |
| C_86 | 13 | awski zach. | 15-64 | 13302 | 14082 | 25091 | 25636 |
| C_1 | 14 | poznański | 0-2 | 2210 | 2079 | 4555 | 4333 |
| C_2 | 14 | poznański | 3-6 | 3103 | 2946 | 6730 | 6435 |
| C_3 | 14 | poznański | 7-12 | 5089 | 4892 | 11565 | 11078 |
| C_4 | 14 | poznański | 13-15 | 2105 | 2037 | 4573 | 4343 |
| C_5 | 14 | poznański | 16-18 | 2063 | 1922 | 4074 | 3712 |
| C_6 | 14 | poznański | 18 | 724 | 652 | 1371 | 1203 |
| C_7 | 14 | poznański | 19-24 | 4242 | 4152 | 7612 | 7347 |
| C_13 | 14 | poznański | 75+ | 2540 | 4777 | 3121 | 5786 |
| C_14 | 14 | poznański | 80+ | 1397 | 2929 | 1664 | 3609 |
| C_15 | 14 | poznański | 85+ | 546 | 1370 | 662 | 1722 |
| C_16 | 14 | bydgoski | 0-2 | 368 | 339 | 1332 | 1245 |
| C_17 | 14 | bydgoski | 3-6 | 528 | 490 | 1942 | 1793 |
| C_18 | 14 | bydgoski | 7-12 | 971 | 897 | 3676 | 3466 |

Przynależność do skupień (Powiaty2050)
Odległość łączenia = 4975,07
Pojedyncze wiązanie
Odległ. euklidesowa

| | Przynależność do skupień | County | Age | City-men | City-women | Country-men | Country-woman |
|---|---|---|---|---|---|---|---|
| C_9 | 1 | poznański | 15-59 | 35730 | 35734 | 106529 | 108461 |
| C_10 | 2 | poznański | 60+ | 24346 | 31485 | 56272 | 69313 |
| C_11 | 3 | poznański | 15-64 | 40859 | 41320 | 119702 | 122791 |
| C_12 | 4 | poznański | 65+ | 19217 | 25899 | 43099 | 54983 |
| C_13 | 5 | poznański | 75+ | 8532 | 13617 | 16844 | 25077 |
| C_24 | 6 | bydgoski | 15-59 | 6643 | 6266 | 31317 | 30827 |
| C_39 | 7 | krakowski | 15-59 | 10600 | 10491 | 71086 | 68921 |
| C_40 | 8 | krakowski | 60+ | 8134 | 9796 | 46120 | 54964 |
| C_41 | 9 | krakowski | 15-64 | 12264 | 12235 | 81230 | 79459 |
| C_42 | 10 | krakowski | 65+ | 6470 | 8052 | 35976 | 44426 |
| C_56 | 11 | gdański | 15-64 | 11627 | 12050 | 39111 | 38683 |
| C_71 | 12 | wrocławski | 15-64 | 9512 | 10555 | 53090 | 52551 |
| C_84 | 13 | awski zach. | 15-59 | 11168 | 11304 | 23734 | 23243 |
| C_85 | 14 | awski zach. | 60+ | 8278 | 10327 | 14602 | 18470 |
| C_86 | 15 | awski zach. | 15-64 | 12775 | 13045 | 26865 | 26586 |
| C_70 | 16 | wrocławski | 60+ | 5032 | 6180 | 26665 | 32290 |
| C_100 | 16 | rzeszowski | 60+ | 5702 | 6449 | 25680 | 29900 |
| C_69 | 17 | wrocławski | 15-59 | 8339 | 9244 | 46785 | 45953 |
| C_101 | 17 | rzeszowski | 15-64 | 9118 | 8651 | 45039 | 43136 |
| C_25 | 18 | bydgoski | 60+ | 4804 | 5778 | 18847 | 22276 |
| C_43 | 18 | krakowski | 75+ | 3043 | 4287 | 15961 | 22744 |
| C_55 | 18 | gdański | 60+ | 6096 | 7769 | 18697 | 21243 |
| C_72 | 18 | wrocławski | 65+ | 3859 | 4869 | 20360 | 25692 |
| C_102 | 18 | rzeszowski | 65+ | 4437 | 5268 | 19722 | 24167 |
| C_8 | 19 | poznański | 0-14 | 10503 | 9910 | 34036 | 32330 |
| C_26 | 19 | bydgoski | 15-64 | 7675 | 7292 | 35603 | 35188 |
| C_54 | 19 | gdański | 15-59 | 10249 | 10492 | 34563 | 34130 |
| C_99 | 19 | rzeszowski | 15-59 | 7853 | 7470 | 39081 | 37403 |
| C_1 | 20 | poznański | 0-2 | 2046 | 1924 | 6443 | 6105 |
| C_2 | 20 | poznański | 3-6 | 2842 | 2684 | 9162 | 8721 |
| C_3 | 20 | poznański | 7-12 | 4240 | 4006 | 13950 | 13254 |
| C_4 | 20 | poznański | 13-15 | 2054 | 1936 | 6674 | 6331 |

**Fig. 6.** Clusters and their instances for 2018 (left) and 2050 (right) data.

For 2018 population in different counties 14 clusters were found. Many of them have only one instance: clusters 1-12. 7 instances belong to cluster 13, and all the rest of a 105-item itemset belong to cluster 14. For 2050 predicted population in different counties 20 clusters were found. Some of them have only one instance: clusters 1-15. The majority of instances belong to cluster 20. The one may observe that for both 2018 and 2050 data some single-instance clusters contain similar corresponding records.

**b) HAC algorithm with complete linkage and Euclidean distance measure for 2018 and 2050 data comparison.**

On figure 7 the one may inspect dendrograms build for 2018 and 2050 counties population size.

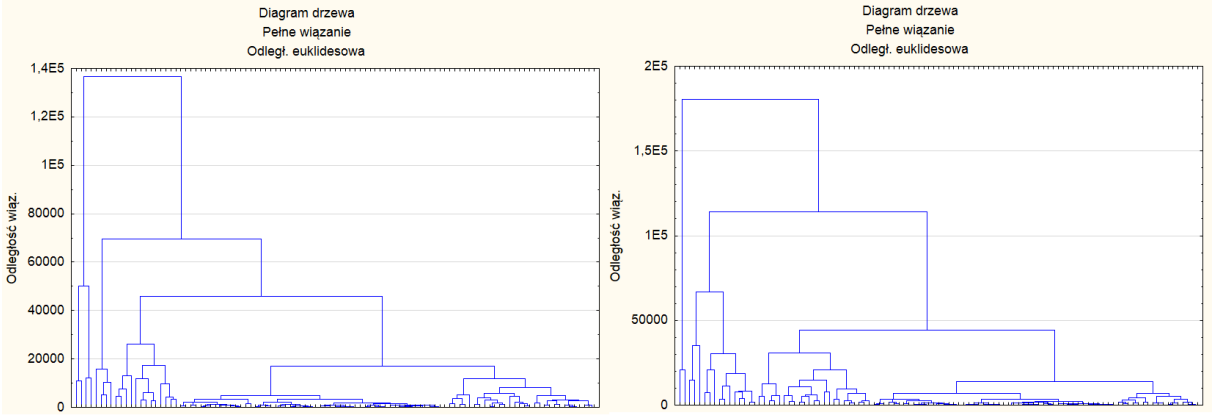



**Fig. 7.** Scaled dendrogram for 2018 (left) and predicted 2050 (right).

In this case we also attempt to find a cut point which determines individual clusters using "knee" rule. Cutting points for both 2018 and 2050 data according to "knee" rule are about 10000 distance units. The points are presented on agglomeration charts below.

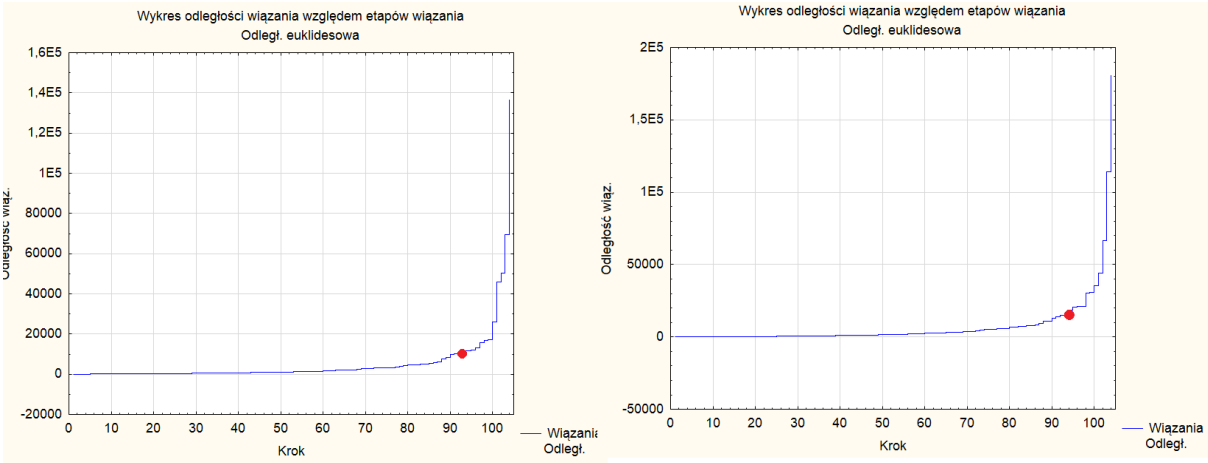



**Fig. 8.** Agglomeration chart for 2018 (left) and 2050 (right) data.
The "knee" points are indicated with red dot.

Clusters and their instances for 2018 and 2050 (right) are not presented, because of higher diversity of discovered clusters. For the 2018 data there were 15 clusters found. Among them clusters 1-8 contain only one instance, clusters 10-13 and 15 contain a few instances and the majority of the data is assigned to clusters 9 and 14. For the 2050 data there 18 clusters found. Among them cluster 1-7 contain only one instances and the majority of data is assigned to cluster 13. It should be noted that for both 2018 and 2050 data single-instance clusters contain nearly identical corresponding 2018/2050 records in terms of county and age group (fig. 9).

Przynależność do skupień (Powiaty2018)
Odległość łączenia = 10022
Pełne wiązanie
Odległ. euklidesowa

| | Przynależność do skupień | County | Age | City-men | City-women | Country-men | **Country-woman** |
|---|---|---|---|---|---|---|---|
| C_9 | 1 | poznański | 15-59 | 41530 | 42759 | 76485 | 78814 |
| C_11 | 2 | poznański | 15-64 | 45849 | 47818 | 83350 | 86457 |
| C_12 | 3 | poznański | 65+ | 8385 | 12578 | 11957 | 16238 |
| C_39 | 4 | krakowski | 15-59 | 13610 | 13704 | 72584 | 72189 |
| C_40 | 5 | krakowski | 60+ | 4842 | 6901 | 21849 | 27719 |
| C_41 | 6 | krakowski | 15-64 | 15158 | 15503 | 80044 | 79757 |
| C_69 | 7 | wrocławski | 15-59 | 6611 | 6905 | 37898 | 37743 |
| C_101 | 8 | rzeszowski | 15-64 | 9956 | 10027 | 49578 | 47461 |

Przynależność do skupień (Powiaty2050)
Odległość łączenia = 9990,66
Pełne wiązanie
Odległ. euklidesowa

| | Przynależność do skupień | County | Age | City-men | City-women | Country-men | **Country-woman** |
|---|---|---|---|---|---|---|---|
| C_9 | 1 | poznański | 15-59 | 35730 | 35734 | 106529 | 108461 |
| C_10 | 2 | poznański | 60+ | 24346 | 31485 | 56272 | 69313 |
| C_11 | 3 | poznański | 15-64 | 40859 | 41320 | 119702 | 122791 |
| C_12 | 4 | poznański | 65+ | 19217 | 25899 | 43099 | 54983 |
| C_39 | 5 | krakowski | 15-59 | 10600 | 10491 | 71086 | 68921 |
| C_41 | 6 | krakowski | 15-64 | 12264 | 12235 | 81230 | 79459 |
| C_42 | 7 | krakowski | 65+ | 6470 | 8052 | 35976 | 44426 |
| C_69 | 8 | wrocławski | 15-59 | 8339 | 9244 | 46785 | 45953 |
| C_101 | 8 | rzeszowski | 15-64 | 9118 | 8651 | 45039 | 43136 |

**Fig. 9.** Single-instance clusters for 2018 and 2050 data.

Generally, the one may observe that using complete linkage agglomeration results in more diverse clusters – instances appear to be more evenly distributed into different groups.

### K-means algorithm

After HAC cluster analysis an alternative analysis was performed using *K-Means Clustering* algorithm. *K*-means clustering is a method of partitioning algorithms paradigm, which generally construct various partitions and then evaluate them by some criterion. All partitioning techniques come down to tasks:

> Construct a partition of a database *D* of *n* objects into a set of *k* clusters.
>
> Given a *k*, find a partition of k clusters that optimizes the chosen partitioning criterion.

*K*-means algorithm represents cluster as its centre – mean of all instances the cluster contains. The centre of the cluster is called *centroid. K*-means clustering algorithm might be represented as a set of instructions:

1) Pick a number *k* of cluster centres – centroids.
2) Choose centroids values randomly.
3) Assign every item to its nearest cluster centre (e.g. using Euclidean distance)
4) Move each cluster centre to the mean of its assigned items
5) Repeat steps 2,3 until convergence (or change lesser than a threshold).

Exemplary illustration of k-means cluster analysis done with Python is shown on figure 10 below.


**Fig. 10.** Visualization of cluster analysis done with *k*-means algorithm.
Stars represents cluster centroids.

It should be noted that *k*-means clustering results may vary significantly depending on initial choice of seeds. The algorithm can also get trapped in local minimum, which is why it is advisable to iterate it with different random seeds. Moreover it is also very sensitive to outliers.

However, when terminated at global optimum, the algorithm produces simple, understandable and sometimes easy to visualize information of data inner structure. *K*-means time complexity (linear) is also lower than the HAC one (polynomial). Using *k*-means user is also able to determine how many cluster the algorithm have to produce.

The *k*-means algorithm cluster analysis was performed using Statistica software for the case study main task. The number of centroids has been arbitrarily adopted as $k = 3$. Centroids were selected to maximize cluster distance. Means of each cluster are presented on figure 11 below.

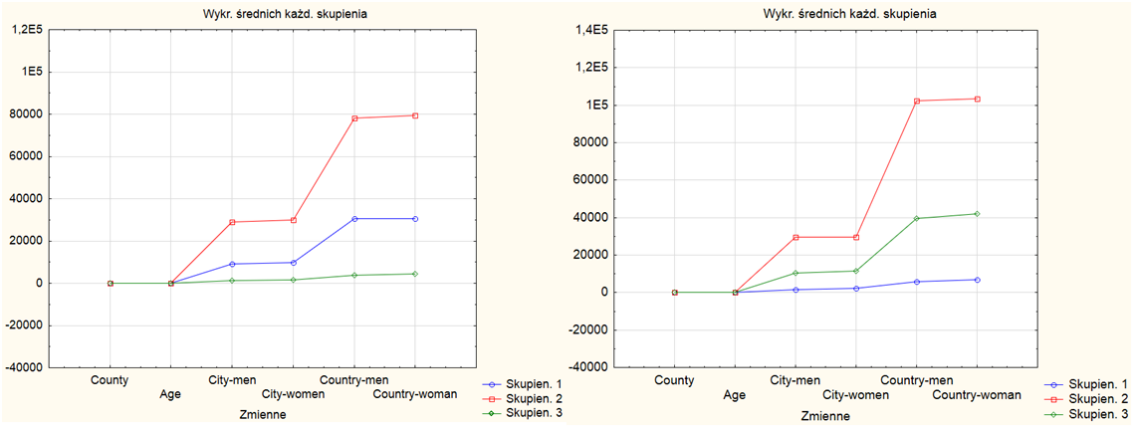



**Fig. 10.** Means of each cluster for 2018 (left) and 2050 (right) data.

The majority of dataset instances were assigned to cluster 1 for both 2018 and 2050. Cluster 2 contains only 3-4 instances for both datasets. Among cluster 2 instances the one may observe *poznański* – age groups *15-59*, *15-64* and *krakowski* – age group *15-64* (*15-59*, *15-64* for 2018).

| | Powiaty2018 | | | | | | | | Powiaty2050 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 County | 2 Age | 3 City-men | 4 City-women | 5 Country-men | 6 Country-woman | 7 GRUPA | | 1 County | 2 Age | 3 City-men | 4 City-women | 5 Country-men | 6 Country-woman | 7 GRUPA |
| C_83 | warszawski | 0-14 | 3402 | 3412 | 6605 | 6240 | 1 | C_83 | warszawski | 0-14 | 3152 | 3124 | 6647 | 6204 | 1 |
| C_85 | warszawski | 60+ | 4404 | 6211 | 7261 | 9387 | 1 | C_85 | warszawski | 60+ | 8278 | 10327 | 14602 | 18470 | 1 |
| C_87 | warszawski | 65+ | 3038 | 4574 | 4652 | 6629 | 1 | C_87 | warszawski | 65+ | 6671 | 8586 | 11471 | 15127 | 1 |
| C_88 | warszawski | 75+ | 1061 | 1941 | 1408 | 2682 | 1 | C_88 | warszawski | 75+ | 3135 | 4609 | 5031 | 7755 | 1 |
| C_89 | warszawski | 80+ | 594 | 1188 | 786 | 1708 | 1 | C_89 | warszawski | 80+ | 1774 | 2952 | 2690 | 4863 | 1 |
| C_90 | warszawski | 85+ | 237 | 581 | 325 | 898 | 1 | C_90 | warszawski | 85+ | 940 | 1837 | 1295 | 2817 | 1 |
| C_91 | rzeszowski | 0-2 | 437 | 419 | 2132 | 2019 | 1 | C_91 | rzeszowski | 0-2 | 393 | 373 | 1796 | 1696 | 1 |
| C_92 | rzeszowski | 3-6 | 648 | 587 | 3019 | 2885 | 1 | C_92 | rzeszowski | 3-6 | 559 | 533 | 2547 | 2424 | 1 |
| C_93 | rzeszowski | 7-12 | 999 | 976 | 5065 | 4788 | 1 | C_93 | rzeszowski | 7-12 | 861 | 826 | 4016 | 3800 | 1 |
| C_94 | rzeszowski | 13-15 | 458 | 437 | 2366 | 2226 | 1 | C_94 | rzeszowski | 13-15 | 427 | 410 | 2039 | 1933 | 1 |
| C_95 | rzeszowski | 16-18 | 465 | 431 | 2471 | 2363 | 1 | C_95 | rzeszowski | 16-18 | 425 | 407 | 2067 | 1967 | 1 |
| C_96 | rzeszowski | 18 | 167 | 142 | 838 | 808 | 1 | C_96 | rzeszowski | 18 | 142 | 136 | 695 | 661 | 1 |
| C_97 | rzeszowski | 19-24 | 955 | 990 | 5576 | 5320 | 1 | C_97 | rzeszowski | 19-24 | 876 | 840 | 4344 | 4109 | 1 |
| C_98 | rzeszowski | 0-14 | 2402 | 2275 | 11795 | 11172 | 1 | C_98 | rzeszowski | 0-14 | 2098 | 2006 | 9717 | 9207 | 1 |
| C_100 | rzeszowski | 60+ | 2787 | 3524 | 12249 | 16408 | 1 | C_102 | rzeszowski | 65+ | 4437 | 5268 | 19722 | 24167 | 1 |
| C_102 | rzeszowski | 65+ | 1838 | 2534 | 8371 | 12569 | 1 | C_103 | rzeszowski | 75+ | 2048 | 2787 | 9093 | 12899 | 1 |
| C_103 | rzeszowski | 75+ | 633 | 1117 | 3292 | 6291 | 1 | C_104 | rzeszowski | 80+ | 1163 | 1849 | 5132 | 8200 | 1 |
| C_104 | rzeszowski | 80+ | 360 | 719 | 1859 | 4122 | 1 | C_105 | rzeszowski | 85+ | 619 | 1130 | 2442 | 4530 | 1 |
| C_105 | rzeszowski | 85+ | 156 | 356 | 773 | 2104 | 1 | C_9 | poznański | 15-59 | 35730 | 35734 | 106529 | 108461 | 2 |
| C_9 | poznański | 15-59 | 41530 | 42759 | 76485 | 78814 | 2 | C_11 | poznański | 15-64 | 40859 | 41320 | 119702 | 122791 | 2 |
| C_11 | poznański | 15-64 | 45849 | 47818 | 83350 | 86457 | 2 | C_41 | krakowski | 15-64 | 12264 | 12235 | 81230 | 79459 | 2 |
| C_39 | krakowski | 15-59 | 13610 | 13704 | 72584 | 72189 | 2 | C_8 | poznański | 0-14 | 10503 | 9910 | 34036 | 32330 | 3 |
| C_41 | krakowski | 15-64 | 15158 | 15503 | 80044 | 79757 | 2 | C_10 | poznański | 60+ | 24346 | 31485 | 56272 | 69313 | 3 |
| C_8 | poznański | 0-14 | 11862 | 11292 | 26033 | 24819 | 3 | C_12 | poznański | 65+ | 19217 | 25899 | 43099 | 54983 | 3 |
| C_10 | poznański | 60+ | 12704 | 17637 | 18822 | 23881 | 3 | C_24 | bydgoski | 15-59 | 6643 | 6266 | 31317 | 30827 | 3 |
| C_24 | bydgoski | 15-59 | 8526 | 8468 | 28991 | 28517 | 3 | C_26 | bydgoski | 15-64 | 7675 | 7292 | 35603 | 35188 | 3 |
| C_26 | bydgoski | 15-64 | 9393 | 9494 | 31820 | 31282 | 3 | C_39 | krakowski | 15-59 | 10600 | 10491 | 71086 | 68921 | 3 |
| C_38 | krakowski | 0-14 | 3549 | 3251 | 19660 | 18465 | 3 | C_40 | krakowski | 60+ | 8134 | 9796 | 46120 | 54964 | 3 |
| C_40 | krakowski | 60+ | 4842 | 6901 | 21849 | 27719 | 3 | C_42 | krakowski | 65+ | 6470 | 8052 | 35976 | 44426 | 3 |
| C_54 | gdański | 15-59 | 9319 | 9889 | 27360 | 26941 | 3 | C_54 | gdański | 15-59 | 10249 | 10492 | 34563 | 34130 | 3 |
| C_56 | gdański | 15-64 | 10174 | 10893 | 29755 | 29429 | 3 | C_56 | gdański | 15-64 | 11627 | 12050 | 39111 | 38683 | 3 |
| C_69 | wrocławski | 15-59 | 6611 | 6905 | 37898 | 37743 | 3 | C_69 | wrocławski | 15-59 | 8339 | 9244 | 46785 | 45953 | 3 |
| C_71 | wrocławski | 15-64 | 7301 | 7721 | 41536 | 41487 | 3 | C_70 | wrocławski | 60+ | 5032 | 6180 | 26665 | 32290 | 3 |
| C_84 | warszawski | 15-59 | 11936 | 12445 | 22482 | 22878 | 3 | C_71 | wrocławski | 15-64 | 9512 | 10555 | 53090 | 52551 | 3 |
| C_86 | warszawski | 15-64 | 13302 | 14082 | 25091 | 25636 | 3 | C_84 | warszawski | 15-59 | 11168 | 11304 | 23734 | 23243 | 3 |
| C_99 | rzeszowski | 15-59 | 9007 | 9037 | 45700 | 43622 | 3 | C_86 | warszawski | 15-64 | 12775 | 13045 | 26865 | 26586 | 3 |

**Fig. 11.** Clusters and their instances for 2018 (left) and 2050 (right) data.

# 4. Conclusions

- Generally, population in *powiat poznański, powiat bydgoski, powiat krakowski, powiat gdański, powiat wrocławski, powiat warszawski zachodni*, *powiat rzeszowski* predicted for 2050 is going to be bigger than in 2018, and higher population growth will likely be observed in villages rather than in cities. It may indicated progressive deurbanization process.
- In studied counties population tend to grow, but mostly within older age groups, what corresponds with thesis about Polish society aging.
- Male-female inhabitants proportion will remains more or less at the same level in 2050 as it was in 2018. Among cities inhabitants, sex disparity aims at an even level in 2050, while

among villages inhabitants women will still outnumber men at the similar level as it was in 2018.

**HAC algorithm**

- More clusters (20) found in 2050 data than in 2018 data (14) suggests increasing diversification of the population of the studied counties.
- Single-instance clusters may be treated as outliers. Similar single-instance clusters in both 2018 and 2050 data indicate counties and age groups that tend to be outside of population - changing processes in terms of structure rather than number (figure 9).

**K-means algorithm**

- Cluster 1 appear to group decreasing population, especially strongly in its rural part.
- Cluster analysis cluster 2 appear to group increasing population that tend to stay at the same proportion of city-village inhabitants in 2050 as it was in 2018.
- Cluster 3 appear to group increasing population, especially strongly in its rural part.
- Small change in cluster 2 instances in 2018 and 2050 may be observed – age group *15-59* for *krakowski* county is within the cluster for 2018, but not for 2050, in which case age group age group *15-59* for *krakowski* county is within cluster 3. The change indicates that population of people of so-called “mobile age” or “production age” from *powiat krakowski* will not stay at the same level, but rather will increase.
- A change in cluster 3 itemset indicates huge growth in village population size mostly in terms of inhabitants of older age in *powiat poznański, krakowski, wrocławski* and *rzeszowski*. In *powiat warszawski zachodni* and *bydgoski* village population size growth will be significant for *15-59* age groups.