

Adam Andrzejuk¹

Warsaw University of Life Sciences – SGGW, Poland

Classification of Agricultural Emissions Among OECD Countries with Unsupervised Techniques

Abstract. Agricultural emissions represent greenhouse gas emissions from crop and livestock production. There are various estimates on agricultural emissions, however on average about 14 to 25 percent of total global emissions comes from agriculture. The main goal of this paper was to present distribution of agricultural emissions among OECD countries with the help of clustering analysis. Clustering analysis is one of the tools used in the field of exploratory data mining. Two methods were used in the analysis: K-means and HDBSCAN algorithms. Both techniques are part of unsupervised learning tasks, which group data into multiple clusters. Finally, an appraisal of obtained classifications was performed.

Key words: agricultural economics, emissions, classification, cluster analysis, k-means, hdbscan

JEL Classification: C4, O13, Q16

Introduction

Greenhouse gases such as water vapor, carbon dioxide, methane, nitrous oxide and ozone, trap heat that would otherwise escape into space and radiate it back towards the earth's surface. This phenomenon is known as the 'greenhouse effect'. The concentration of greenhouse gases in the earth's atmosphere has grown mainly as a result of human activity. The growth of greenhouse gas emissions may be linked to rising temperatures, rising sea levels and more frequent weather anomalies, otherwise referred to as 'global warming' (EuroStat, 2018).

The estimates of agriculture's contribution to greenhouse gas emissions vary depending on a source of data. According to the OECD Trade and Agriculture Directorate, agriculture contributes to climate change and at the same time is affected by it. At least 14% of global greenhouse gas emissions come directly from the farm sector. (Legg et al., 2010). The United States Environmental Protection Agency (EPA), estimates that agriculture contributes to 24% of global greenhouse gas emissions worldwide, and 9% in the United States alone (EPA, 2018).

How does agriculture contribute to such emissions? One way is through direct farming activity. Ploughing fields releases carbon dioxide in the soil. Rice cultivation and livestock breeding both emit large quantities of methane. Another way is that farming uses fossil fuels and fertilizers. Agriculture also involves land-use changes, including deforestation and desertification of fragile grasslands. These changes alter the earth's ability to absorb or reflect heat and light (Legg et al., 2010).

¹ MA, Department of Agricultural Economics and International Economic Relations WULS-SGGW, ul. Nowoursynowska 166, 02-787 Warszawa, e-mail: adam_andrzejuk@sggw.pl; <https://orcid.org/0000-0002-9844-3285>

Data from OECD Stats reveal that in OECD countries, agriculture has significant impact on the environment, as it uses on average over 40% of available water and land. The impact occurs on and off farm. Most OECD countries are tracking the environmental performance of agriculture, which is informing policy makers and society on the state and trends in agri-environmental conditions, and can provide a valuable aid to policy analysis (OECD Stat, 2018).

The main goal of this paper is to classify OECD countries according to agricultural production and emissions. There have been many studies on agricultural emissions in specific countries, or between developed and developing economies (Tubiello et al., 2017; Wilkes et al, 2017; FAO, 2015). This article tries to elaborate on agricultural emissions in relation to agricultural production within the OECD members. In addition, two unsupervised classification methods will be tested to compare outcomes of clustering.

Material and methods

The data for the analysis was acquired from FAO Stat service for all 36 OECD member states. However, in the course of data preprocessing, United States was eliminated as the single biggest outlier. Due to lack of data, Latvia and Lithuania were not included in the research. In the end, the total of 33 OECD member states were taken into account. The dataset included two features per country: Agriculture Gross Production Value and Agriculture Total Emissions (in CO₂eq). According to FAO methodology, the value of gross production has been compiled by multiplying gross production in physical terms by output prices at farm gate. Thus, value of production measures production in monetary terms at the farm gate level. Since intermediate uses within the agricultural sector (seed and feed) have not been subtracted from production data, this value of production aggregate refers to the notion of "gross production" (FAO, 2018a).

Agriculture Total Emissions contains total Greenhouse Gas (GHG) emissions, and aggregated GHG emissions for each greenhouse gas (CH₄, N₂O), expressed in CO₂ equivalents. Total agricultural emissions include the following sub-domains: Enteric fermentation (CH₄), Manure management (CH₄, N₂O), Rice cultivation (CH₄), Synthetic fertilizers (N₂O), Manure applied to soils (N₂O) Manure applied to pastures (N₂O), Crop residues (N₂O), Cultivation of organic soils (N₂O), Burning-crop residues (CH₄, N₂O), Burning-savanna (CH₄, N₂O) (FAO, 2018b).

To ensure equal availability of data among different countries, 2016 was chosen as the year under investigation. The study was design to compare clustering results.

Clustering is a task of organizing a dataset of objects into groups (or clusters) of objects with similar traits. Sometimes clustering is also called - data segmentation. It is an example of unsupervised learning, and does not require each data object to be manually labeled. As a result, clustering is useful for exploratory data analysis, exploring datasets that are not yet well-understood. There are many methods that can be used to cluster a dataset, each with their own advantages and disadvantages. One popular method is K-means (Qiao et al., 2018).

K-means clustering is an unsupervised machine learning algorithm. On the other hand, an example of a supervised learning algorithm can be seen when looking at Neural Networks where the learning process involves both the inputs (x) and the outputs (y). During the learning process the error between the predicted outcome (predY) and actual

outcome (y) is used to train the system. In an unsupervised method such as K-means clustering the outcome (y) variable is not used in the training process (Stamford, 2015).

Another method is HDBSCAN, which stands for Hierarchical Density-Based Spatial Clustering of Applications with Noise. It's a clustering algorithm that overcomes many of the limitations of K-means. For example, it does not require a difficult-to-determine parameter to be set before it can be used (Qiao et al., 2018). HDBScan, like other clustering algorithms it is used to group similar data together.

All calculations were performed using a general purpose programming language Python with addition of supplementary scientific libraries such as Numpy, Scipy, Scikit and Pandas.

Results and discussion

Comparison of emissions

Before attempting to classify countries according to agricultural production and emissions. It's advisable to compare agricultural emissions with total emissions of OECD countries. Table 1, provides data for total greenhouse gas emissions in selected OECD countries, in years between 2010 and 2016.

It's worth mentioning that the majority of OECD countries have seen reduction in total greenhouse gas emissions from 2010 to 2016 which is in line with the Kyoto Protocol which entered into force on 16 February 2005. For all OECD members the reduction was 4,3%, whereas for the OECD European members, the reduction amounted to 7,5%. The United States also achieved reduction of GHG emissions of almost 6%.

Table 1. Total greenhouse gas emissions in selected countries in 2016 (tones of CO2 equivalent, Millions)

| Country | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 |
|---------------|--------|--------|--------|--------|--------|--------|--------|
| United States | 6 923 | 6 771 | 6 529 | 6 709 | 6 763 | 6 638 | 6 511 |
| Japan | 1 300 | 1 351 | 1 393 | 1 407 | 1 360 | 1 321 | 1 305 |
| Germany | 943 | 920 | 925 | 942 | 903 | 907 | 909 |
| Canada | 694 | 700 | 707 | 716 | 716 | 714 | 704 |
| Australia | 539 | 540 | 544 | 533 | 527 | 538 | 549 |
| Turkey | 403 | 431 | 446 | 439 | 452 | 470 | 496 |
| UK | 616 | 568 | 584 | 570 | 529 | 511 | 486 |
| France | 517 | 490 | 490 | 490 | 461 | 465 | 465 |
| Italy | 504 | 491 | 472 | 441 | 425 | 433 | 428 |
| Poland | 406 | 405 | 398 | 395 | 382 | 385 | 396 |
| OECD - Europe | 5 080 | 4 935 | 4 898 | 4 814 | 4 650 | 4 693 | 4 699 |
| OECD - Total | 16 020 | 15 828 | 15 602 | 15 734 | 15 566 | 15 467 | 15 330 |

Source: Author's own study based on OECDStat data: https://stats.oecd.org/Index.aspx?DataSetCode=AIR_GHG.

Table 2, provides data for agricultural greenhouse gas emissions in selected OECD countries between 2010 and 2016. Results of data analysis for agricultural emissions in that period is not as clear cut as the total emissions. In this case, the emissions for most countries have increased. It's worth mentioning that larger decrease of agricultural

greenhouse gas emissions can be observed on a longer time scale between 1990's and the current period. This according to EuroStat can be attributed to an overall reduction in livestock numbers, more efficient farming practices, the reduced application of nitrogen-based fertilizers, as well as better forms of manure management (EuroStat, 2018).

In 2016 the highest increase compared to the year 2010 was reported in Turkey (32%), Hungary (22%) and Czech Republic (15%). During the analyzed period, decrease of agricultural emissions occurred in 6 countries. The highest reduction was achieved in Greece (11%), Japan (6,5%) and Belgium (3,2%). It's noticeable that the highest reduction of GHG emissions in Greece happened during the period of its worst financial crisis.

Table 2. Agricultural greenhouse gas emissions in selected countries in 2016 (tones of CO2 equivalent, Millions)

| Country | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 |
|----------------|-------|------|------|-------|-------|-------|-------|
| United States | 550 | 536 | 520 | 543 | 540 | 567 | 563 |
| France | 77 | 77 | 77 | 76 | 78 | 78 | 77 |
| Australia | 66 | 71 | 72 | 73 | 73 | 70 | 69 |
| Germany | 63 | 64 | 64 | 65 | 66 | 67 | 65 |
| Canada | 56 | 55 | 57 | 59 | 58 | 59 | 60 |
| Turkey | 43 | 45 | 51 | 54 | 54 | 54 | 56 |
| United Kingdom | 42 | 42 | 41 | 41 | 43 | 42 | 42 |
| New Zealand | 38 | 38 | 39 | 39 | 40 | 39 | 39 |
| Spain | 34 | 33 | 32 | 32 | 34 | 35 | 34 |
| Japan | 36 | 35 | 35 | 35 | 34 | 34 | 34 |
| OECD - Europe | 448 | 450 | 454 | 459 | 466 | 467 | 470 |
| OECD - Total | 1 004 | 997 | 988 | 1 017 | 1 019 | 1 043 | 1 039 |

Source: Author's own study based on OECDStat data: https://stats.oecd.org/Index.aspx?DataSetCode=AIR_GHG.

Comparing the data on total emissions and agricultural emission for each OECD country in 2016, it is noticeable that the country with the highest percentage of agricultural emissions is New Zealand (49,2%), followed by Ireland (31,3%) and Denmark (20,5%). The outcome is a result of lower share of heavy industry in the economies, which are geared towards services rather than production by means of polluting technologies. Hence, it cannot be inferred that those countries rely on agriculture as a main driver of GDP. On the other hand, the country with the lowest ratio of agricultural emissions to total emissions is Japan (2,6%). The highly developed economy, with extensive supply chains which include every step from processing of raw materials to final assembly of products, dwarfs the output from agriculture. In this respect, Japan is followed by three former communist countries: Slovak Republic, Czech Republic and Estonia, with ratios of 6,5% for Slovakia and 6,6% for the last two countries.

On a nominal scale, the country with the highest agricultural emissions is the United States. According to OECD data its gas emissions was 562,590 thousand tones, of CO2 equivalent. In 2016 alone, agricultural emissions from the world's biggest economy were more than 7 times bigger than those of France, which is the second largest agricultural emissions producer in the world, with a value of 76,957 thousand tones, of CO2 equivalent. Next in hierarchy was Australia, Germany, Canada and Turkey. In nominal terms, the

lowest agricultural emissions in 2016 was in Iceland and amounted to 602 thousand tones, of CO2 equivalent.

The data from 2016 shows, that in a group of OECD countries the ratio of agricultural greenhouse gas emissions to total gas emissions is 8%, whereas in OECD European countries the ratio stands at 10%. The ratio for the United States is even lower than in Europe and stands at 8,6% which is in line with EPA estimates (EPA, 2018).

Clustering of countries in relation to emissions

According to Vivek, clustering is an unsupervised machine learning technique, where there are no defined dependent and independent variables. The patterns in the data are used to identify or group similar observations. The objective of any clustering algorithm is to ensure that the distance between data points in a cluster is very low compared to the distance between two clusters. In other words, members of a group are very similar, and members of different groups are as much as possible dissimilar. (Vivek, 2018) In other words, it is a task of organizing datasets of objects into groups of similar objects.

One of such clustering techniques is a popular K-means algorithm. The algorithm was used among many others by Kisielińska (2009) for cluster analysis in the agricultural economics, as well as Pietrzykowski (2006) and Kobus (2006) in portfolio analysis.

The main advantage of K-means is that it is fast and easy to understand. However K-means is not really a clustering algorithm in the strict sense of the word. It is a partitioning algorithm. Tholoz writes, that K-means doesn't "find clusters", it partitions the dataset into as many (assumed to be globular) chunks as asked for, by attempting to minimize intra-partition distances. That leads to the second problem: one needs to specify exactly how many clusters are expected. If a lot is known about the data then that is something that might be expected. If, on the other hand, one is simply exploring a new dataset then the number of clusters is a hard parameter to have any good intuition for. The usually proposed solution is to run K-Means for many different number of clusters values and score each clustering with certain measure such as Silhouette index. Finally K-means is also dependent upon initialization. Given multiple different random starts, multiple different clusterings can be obtained (Tholoz, 2018).

The mathematics behind K-means method is quite simple. It all comes down to minimizing the sum of square of distances between the cluster centroid and its associated data points.

$$\text{Minimize } \sum_{j=1}^k \sum_{i=1}^n (x_{ij} - c_j)^2$$

where: k is number of clusters, n is number of data points, c is centroid of cluster j and $(x_{ij} - c_j)$ is a distance between data point and centroid.

However, a question may arise as to how to pick the right number of clusters. Given that K-Means has no in-built preference for the right number of clusters, following are some of the common ways that k can be selected:

Domain Knowledge – This is one of most common method in practice. Segmentation very often doesn't exist in vacuum and is aimed towards solving a problem. Therefore if

there is a requirement or a researcher prefers certain number or range of clusters, then that can be used to select k.

Elbow-Method using Within-Cluster-Sum-of-Squares (WCSS). Elbow method is a method which looks at the percentage of variance explained as a function of the number of clusters. This method exists upon the idea that one should choose a number of clusters so that adding another cluster doesn't provide any better modelling of the data. The percentage of variance explained by the clusters is plotted against the number of clusters. The first clusters will add much information but at some point the marginal gain will drop dramatically. This gives an angle in the graph (Bholowalia et al., 2014).

Cluster using silhouette coefficient. Struyf interprets the value of silhouette coefficient according to the method originally provided by Kaufman and Rousseeuw (Table 3), however noticing that experience has led to the subjective interpretation of the silhouette coefficient (Struyf et al., 1996).

Table 3. Interpretation of the silhouette coefficient for partitioning methods

| Silhouette coefficient | Proposed interpretation |
|------------------------|---|
| 0.71 - 1.00 | A strong structure has been found |
| 0.51 - 0.70 | A reasonable structure has been found |
| 0.26 - 0.50 | The structure is weak and could be artificial, try additional methods |
| < 25 | No substantial structure has been found |

Source: Struyf et. al., Clustering in an Object-Oriented Environment.

As mentioned earlier, another one of such clustering techniques is a powerful algorithm called HDBSCAN, which stands for Hierarchical Density-Based Spatial Clustering of Applications with Noise. HDBSCAN method was used by Rahman (2017) to augment location based services on the Internet (Rahman et al., 2016). It was also used in many biological applications, for instance in clustering of temporal patterns in high-dimensional neuronal ensembles (Grossberger, 2018).

According to Tholoz, HDBSCAN is an algorithm developed by some of the same people who wrote the original DBSCAN paper. And is an improvement over the DBSCAN. Their goal was to allow varying density clusters. The algorithm starts off much the same as DBSCAN which means that it assumes clusters for dense regions. It doesn't require that every point be assigned to a cluster, like was the case with K-means algorithm. Hence it doesn't partition the data. Instead, it extracts the "dense" clusters and leaves sparse background classified as noise. As a first step DBSCAN transforms the space according to the density of the data: points in dense regions are left alone, while points in sparse regions are moved further away. Applying single linkage clustering to the transformed space results in a dendrogram, which is cut according to a distance parameter (called epsilon or eps in many implementations) to get clusters. Epsilon can be thought of as the radius those neighbors have to be in for the core to form. The different approach that the HDBSCAN method takes is that instead of taking an epsilon value as a cut level for the dendrogram, HDBSCAN has a parameter called minimum cluster size, which determines how big a cluster needs to be in order to form. HDBScan approaches this by throwing out the tiny off shoots, and instead keeping the biggest clusters as defined by the minimum cluster size parameter. This results in a more condensed dendrogram with fewer loose points. That tree can then be used to select the most stable or persistent clusters. This process allows the tree

to be cut at varying height, picking varying density clusters based on cluster stability. The immediate advantage of this is that there are varying density clusters (Tholoz, 2018). HDBSCAN is built for the real-world scenario of having data with varying density, it's relatively fast, and it lets the researcher define what clusters are important based on their size (Bailey, 2017).

Exploring data of agricultural emissions and production

As a data mining function, cluster analysis can be used as a standalone tool to gain insight into the distribution of data, to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis. Clustering is a challenging research field. Many clustering algorithms require users to provide domain knowledge in the form of input parameters such as the desired number of clusters. Parameters are often hard to determine (Han et al., 2012). Such is the case w K-means algorithm.

In order to determine the number of cluster two methods were used: Elbow-Method using Within-Cluster-Sum-of-Squares and Silhouette Coefficient. The plot of within-cluster sum of squares against number of clusters is shown below.

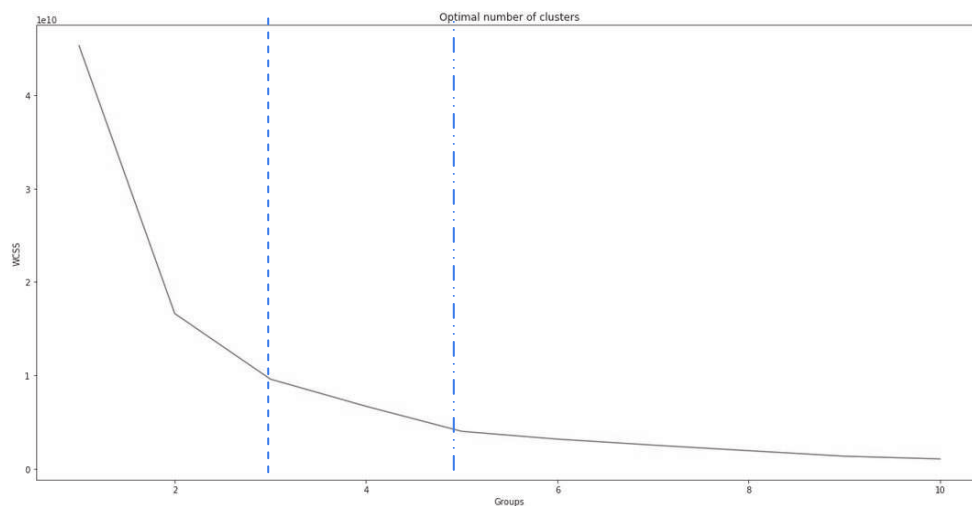


Fig. 1. Within-cluster sum of squares against number of clusters

Source: Author's own study based on the data from FAOStat.org. <http://www.fao.org/faostat/en/#data/GT>, (Access date: 19.09.2018).

Based on the domain knowledge and the elbow graph, it is initially suggested that three clusters should be used for the K-means algorithm. However, there appears to be the next bend around 5 groups, which leaves the case inconclusive. In order to obtain further proof on the number of clusters, a silhouette coefficient was calculated for several values of k.

The outcome of silhouette coefficient calculations shows that segmentation consisting of 3 groups has a much better silhouette coefficient (0,610) than 5 groups (0,570). And by using 3 groups it can be argued that a reasonable structure of groups has been found. Combining the knowledge from the elbow method and the silhouette calculations suggest, that three clusters should be used for further analysis.

Table 4. The silhouette coefficient for k-means method

| Number of clusters | Silhouette Coefficient |
|--------------------|------------------------|
| 2 | 0,626 |
| 3 | 0,610 |
| 4 | 0,523 |
| 5 | 0,570 |
| 6 | 0,491 |
| 7 | 0,549 |
| 8 | 0,438 |
| 9 | 0,452 |
| 10 | 0,394 |

Source: Author’s own study.

It can be seen from the graph (Fig. 2), that the clustering has yielded three groups of countries based on agricultural production and emissions. It is shown with triangles, circles and crosses. Additionally centroids were included as bigger dark diamonds. It is important to remember that K-means method is a partitioning algorithm that partitions the dataset into clusters and attaches all data points to a certain group. Therefore every country has its own group (Table 5). The first group consists of two countries with very high agricultural emissions and average agricultural production (triangles). The second group consists of eight countries with slightly lower than average agricultural emissions and slightly higher than average agricultural production (crosses). The largest group, consisting of 23 states and represented with circles, includes countries with lower than average agricultural emissions and lower than average agricultural production.

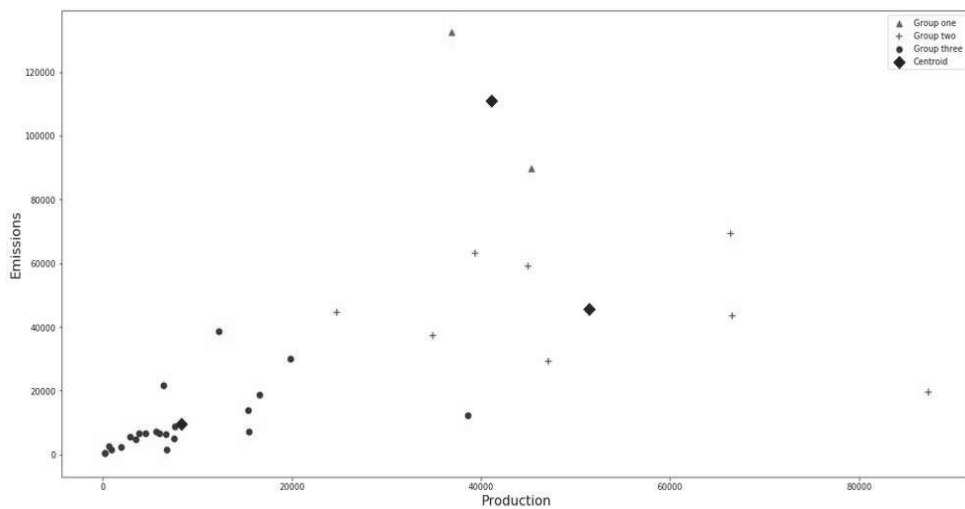


Fig. 2. Outcome of K-means clustering

Source: Author’s own study based on the data from FAOStat.org. <http://www.fao.org/faostat/en/#data/GT>, (Access date: 19.09.2018).

Table 5. Classification of OECD member states among different clusters with K-means method

| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 3 cont. | Cluster 3 cont. |
|-----------|-----------|-----------|-----------------|-----------------|
| Australia | Canada | Austria | Hungary | Poland |
| Mexico | France | Belgium | Iceland | Portugal |
| | Germany | Chile | Ireland | Rep. of Korea |
| | Italy | Czechia | Israel | Slovakia |
| | Japan | Denmark | Luxembourg | Slovenia |
| | Spain | Estonia | Netherlands | Sweden |
| | Turkey | Finland | New Zealand | Switzerland |
| | UK | Greece | Norway | |

Source: Author's own study.

To find clusters of arbitrary shape, one can model clusters as dense regions in the data space, separated by sparse regions. This is the main strategy behind density-based clustering methods, which can discover clusters of nonspherical or non-globular shape. (Han et al., 2012). One of the advantages of density-based method is that they do not require prior knowledge of number of clusters. A method, which represents one of the many density-based clustering methods and doesn't require input in the form of parameters is HDBSCAN. Another advantage is that it doesn't require that every data point be assigned to a cluster. HDBSCAN accepts the dense clusters and treats the sparse background as outliers.

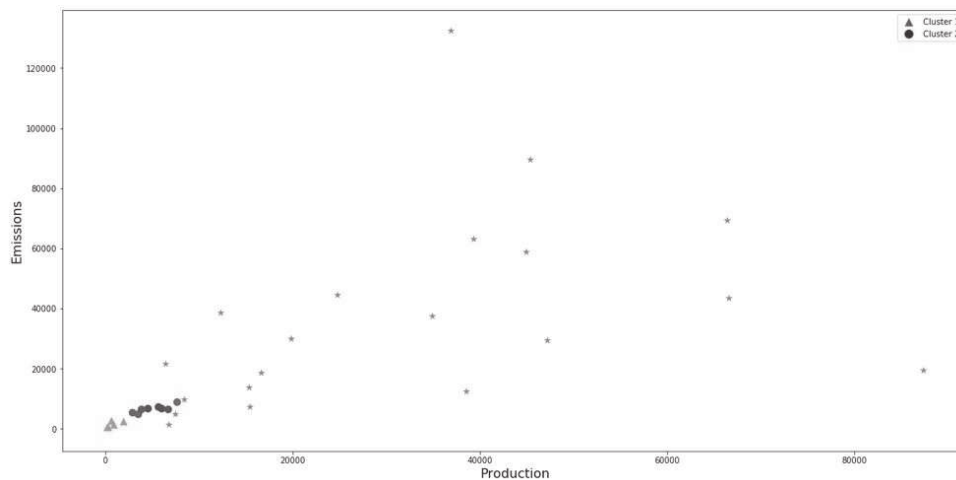


Fig. 3. Outcome of HDBSCAN clustering

Source: Author's own study based on the data from FAOStat.org. <http://www.fao.org/faostat/en/#data/GT>, (Access date: 19.09.2018).

The plot shows that the outcome of HDBSCAN algorithm produced two groups of countries based on agricultural emissions and production. Since the method does not require that all data points need to be included in a group, many countries were treated as outliers. It must be mentioned, however that HDBSCAN method also requires some user

input. Despite the fact of not needing prior input in the form of number of clusters, the method still requires user input in the form of minimum cluster size. For this research, since the number of explored countries is relatively low, the minimum cluster size parameter was set to three. Increasing the number of minimum cluster size parameter did not produce higher number of clusters. On the contrary, increasing the parameter beyond 5 did not produce any clusters.

Table 6. Classification of OECD member states among different clusters with HDBSCAN method

| Cluster 1 | Cluster 2 | Outliers | Outliers cont. |
|------------|-----------|-----------|----------------|
| Estonia | Austria | Australia | Japan |
| Iceland | Belgium | Canada | Mexico |
| Luxembourg | Czechia | Chile | Netherlands |
| Slovakia | Finland | Denmark | New Zealand |
| Slovenia | Hungary | France | Poland |
| | Norway | Germany | Rep. of Korea |
| | Portugal | Greece | Spain |
| | Sweden | Ireland | Switzerland |
| | | Israel | Turkey |
| | | Italy | UK |

Source: Author’s own study.

In yet another approach, where the minimum cluster size was set to two, the number of clusters has increased to three. However, the third cluster contained only two data points, separated from each other by a wide margin. In such circumstances, it was determined that the current minimum cluster size was the right choice.

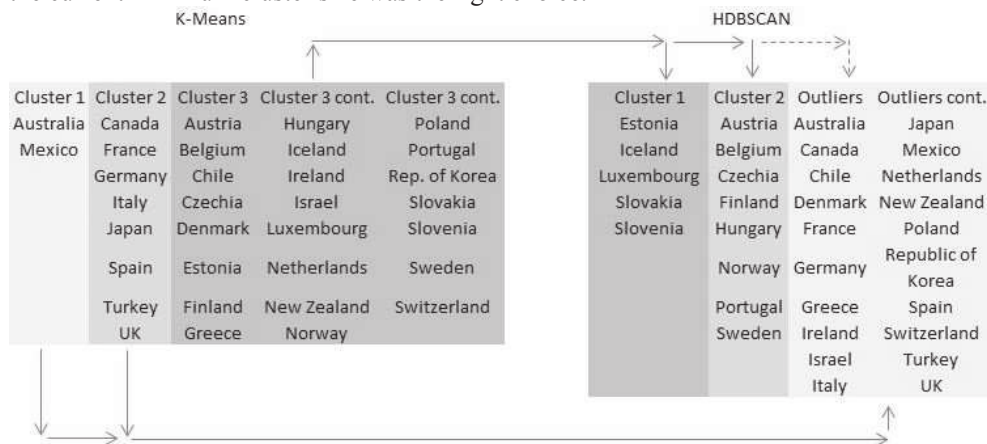


Fig. 4. Comparison of K-Means and HDBSCAN clustering

Source: Author’s own study.

The HDBSCAN algorithm ensured, that countries are classified according to their similarity with other group members in terms of agricultural emissions and production (Table 6). The first group include states with very little agricultural emissions and very low agricultural production. The second group falls into slightly higher territory of emissions and production, while still being considered low in general terms, compared to the states

classified as outlier. It has to be pointed out, that the clustering included countries with very little variance of those two variables among themselves. Countries with higher variance of either agricultural emissions or production were automatically left out as noise.

In order to put the analysis in perspective, it has to be stated that all countries from the first and second cluster of K-means partitioning were classified as outliers in the HDBSCAN method (Fig. 4). This comes as a result of high dispersion among their agricultural emissions and agricultural production values in relation to each other. On the other hand 13 out of 23 states from cluster three, in K-means method, which represents countries with the lowest agricultural emissions and the lowest agricultural production ended up in cluster 1 and 2 of the HDBSCAN method. That results from their dense concentration on the left side of the plot.

It can be reasoned that the HDBSCAN method, only investigates the densely concentrated region of the plot which includes countries from cluster three in K-means method, and then further discriminates among countries in that segment.

Conclusions

Major conclusion which can be drawn from the analysis is that the two models discussed, don't match exactly. Results obtained from the K-means clustering are very different than those obtained by applying the HDBSCAN algorithm. But such is the nature of unsupervised learning process. As Luxburg puts it, the difficulty with unsupervised clustering is that there are a huge number of possibilities regarding what will be done with it. And there is no abstraction akin to a loss function in supervised learning, which distills the end-user intent (Luxburg et al., 2012).

The K-means algorithm puts all countries into a certain group, whereas HDBSCAN method narrows down selection to only a few states and then discriminates among them.

It can be inferred from the analysis that cluster three in the K-means method represents countries with the lowest agricultural emissions and production among all OECD states. However this cluster can further be split into two segments which includes very small countries such as Estonia, Iceland, Luxembourg, Slovakia and Slovenia, and larger countries such as Austria, Belgium, Czech Republic, Finland, Hungary, Norway, Portugal and Sweden.

Taking into account data, which represent developed economies of OECD countries without drastic disproportions in agricultural greenhouse gas emissions and agricultural production, it cannot simply be argued that, just because K-means technique assigns all countries to a group, the K-means algorithm serves the purpose of classification better than HDBSCAN.

The two algorithms show complementary picture of the situation. However, an issue which also has to be accounted for, is a low number of explored samples. The research included only 33 states, whereas both the K-means and the HDBSCAN methods are well suited to handle much more data.

References

- Bailey, B. (2017). Lightning Talk: Clustering with HDBScan, <https://towardsdatascience.com/lightning-talk-clustering-with-hdbscan-d47b83d1b03a>. (Access date: 21.09.2018).
- Bholowalia, P., Kumar, A. (2014). EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. <https://pdfs.semanticscholar.org/5771/aa21b2e151f3d93ba0a5f12d023a0bfcf28b.pdf>. (Access date: 21.09.2018).
- EPA (2018). Global Greenhouse Gas Emissions Data, <https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data>. (Access date: 19.09.2018).
- EuroStat (2018). Agriculture - greenhouse gas emission statistics, https://ec.europa.eu/eurostat/statistics-explained/index.php/Archive:Agriculture_-_greenhouse_gas_emission_statistics. (Access date: 19.09.2018).
- FAO (2015). Estimating Greenhouse Gas Emissions In Agriculture. A Manual to Address Data Requirements for Developing Countries, <http://www.fao.org/3/a-i4260e.pdf>. (Access date: 19.09.2018).
- FAO Stat (2018a). Value of gross production, http://fenixservices.fao.org/faostat/static/documents/QV/QV_e.pdf. (Access date: 19.09.2018).
- FAO Stat (2018 b). Agriculture Total Emissions, http://fenixservices.fao.org/faostat/static/documents/GT/GT_e.pdf. (Access date: 19.09.2018).
- Grossberger, L., Battaglia, F.P., Vinck, M. (2018). A novel distance measure for the unsupervised clustering of temporal patterns in high-dimensional neuronal ensembles, <https://www.biorxiv.org/content/biorxiv/early/2018/01/24/252791.full.pdf?%3Fcollection=>. (Access date: 21.09.2018).
- Han, J., Kamber, M., Pei, J. (2012). Data mining concepts and techniques, Third edition, Morgan Kaufmann Publishers, Elsevier Inc.
- Legg, W., Huang, H. (2010). Climate change and agriculture, http://oecdobserver.org/news/archivestory.php/aid/3213/Climate_change_and_agriculture.html. (Access date: 19.09.2018).
- OECD Stat (2018). Environmental Performance of Agriculture in OECD countries since 1990, <https://stats.oecd.org/index.aspx?queryid=521>. (Access date: 19.09.2018).
- Luxburg, U., Wiliamson, R.C., Guyon, I. (2012). Clustering: Science or Art? <http://proceedings.mlr.press/v27/luxburg12a/luxburg12a.pdf>. (Access date: 23.09.2018).
- Qiao, A., Jackson, J. (2018). Scalable Clustering for Exploratory Data Analysis, <https://medium.com/Petuum/scalable-clustering-for-exploratory-data-analysis-60b27ea0fb06>. (Access date: 19.09.2018).
- Rahman, F., Liu, W., Suhaim, S., Thirumuruganathan, S., Zhang, N., Das, G., HDBSCAN: Density based clustering over location based services, <https://arxiv.org/pdf/1602.03730.pdf>. (Access date: 21.09.2018).
- Stamford, J. (2015). K Means Clustering in Python, <http://stamfordresearch.com/k-means-clustering-in-python/>. (Access date: 19.09.2018).
- Struyf, A., Hubert, M., Rousseeuw, P. (1996). Clustering in an Object-Oriented Environment. *Journal of Statistical Software*, 1(4) URL <http://www.jstatsoft.org/v01/i04/>. (Access date: 21.09.2018).
- Tholoz, P. (2018). Comparing Python Clustering Algorithms, https://github.com/scikit-learn-contrib/hdbscan/blob/master/docs/comparing_clustering_algorithms.rst. (Access date: 21.09.2018).
- Tubiello, F.N., Cerilli, S., Conchedda, G. (2017). Mapping IPCC greenhouse gas emissions categories to ISIC A in the SEEA AFF, https://seea.un.org/sites/seea.un.org/files/lg23_seea_aff_ghg_final.pdf. (Access date: 19.09.2018).
- Vivek, S. (2018). Clustering algorithms for customer segmentation, <https://towardsdatascience.com/clustering-algorithms-for-customer-segmentation-af637c6830ac>. (Access date: 19.09.2018).
- Wilkes, A., Reisinger, A., Wollenberg, E., van Dijk, S. (2017). Measurement, reporting and verification of livestock GHG emissions by developing countries in the UNFCCC: current practices and opportunities for improvement, https://cgspace.cgiar.org/bitstream/handle/10568/89335/CCAFS_Report17.pdf. (Access date: 19.09.2018).

For citation:

Andrzejuk A. (2018). Classification of Agricultural Emissions Among OECD Countries with Unsupervised Techniques. *Problems of World Agriculture*, 18(4), 80–91;
DOI: 10.22630/PRS.2018.18.4.99