OPEN GOVERNMENT DATA PORTALS IN THE EUROPEAN UNION: CONSIDERATIONS, DEVELOPMENT, AND EXPECTATIONS

De Juana-Espinosa, Susana.

Business Organization Department, Universidad de Alicante. Susana.Espinosa@ua.es

Lujan-Mora, Sergio.

Languages and Information Systems Department. Universidad de Alicante.

Sergio.Lujan@ua.es

Abstract

The goal of open government data (OGD) initiatives is to promote transparency, efficiency and public participation in public management policies. To do so, public organizations must consider which elements might help the development of their open government data portals (OGDP). This paper studies the evolution of OGDP in the 28 countries of the European Union (EU) in a multidisciplinary setting. Whereas the comparative frameworks in the literature are mostly based only on technological parameters, this exploratory research aims to uncover which factors might uphold the successful development of OGDP through the analysis of the relationships between a number of technical and socioeconomical indicators over a period of three years (2015-2017), using a clustering methodology. The results show that EU countries are slowly homogenizing their OGD approaches into two currents/speeds, based mainly on economic factors and open government development status. The originality of this research lies in the sense that it provides not only a technical benchmark, but also a longitudinal and multidisciplinary perspective that will add to the current formulation of OGD policies and practices in any international setting.

Keywords: *Open Government Data (OGD), European Union, Socioeconomic Indicators, Longitudinal Analysis, Cluster Analysis.*

1. Introduction

According to Open Knowledge Foundation (n.d.), open data refers to data that may be "…freely accessed, used, modified, and shared by anyone for any purpose". Open government data (OGD) then, refers to that public sector data that are freely available for access and exploitation (Kalampokis, Tambouris, & Tarabanis, 2011) by public service stakeholders, namely, politicians, data collectors, data processors, data publishers, infrastructure providers, companies and firms,

infomediaries, citizens and policy makers (Charalabidis, et al., 2018). Research on OGD is important because public sector organizations (PSOs) are one of the major producers and users of information (Aichholzer & Burkert, 2004).

OGD is the combination of linked data, open data, big data and government data (Charalabidis, et al., 2018). Although the first OGD ideas can be dated from the 1950s in the United States of America (USA) (Parks, 1957), the European Union (EU) was a pioneer in this movement, with the delivery of the EU Public Sector Information Directive in 2003 (European Parliament and Council of the EU, 2003) which was revised in 2013 (European Parliament and Council of the EU, 2013). In the USA, during Obama's tenure, the USA Open Government Directive of December 8, 2009 (Office of the President, 2009) was launched, which Canada followed in 2011 with a pilot national open data site, followed in 2014 with the Directive on Open Government (Government of Canada, 2014). Another milestone was the inauguration of the Open Government Partnership (https://www.opengovpartnership.org) in 2011, a voluntary initiative whereby governments pledged to empower citizens and fight corruption by means of procuring and developing adequate and valuable OGD initiatives. By September 2019, a total of 79 countries and 20 local initiatives had signed up as members of this organization.

The uses of OGD should be centered on creating public value and constructing public policies based on an open culture (Zuiderwijk & Janssen, 2014a), such that governments should move from open data to open service (Chan, 2013; Yang, Lo, & Shiang, 2015), deploy collaborative solutions among e-government stakeholders to foster innovation (Edelmann, Höchtl, & Sachs, 2012; Veljković, Bogdanović-Dinić, & Stoimenov, 2014; Yang & Kankanhalli, 2013), and enforce government efficiency, transparency and accountability (Attard, Orlandi, Scerri, & Auer, 2015; Janssen, Matheus, Longo, & Weerakkody, 2017; Lourenço, 2013). In addition, the geopolitical context is a crucial factor, since the effectiveness of open government policies is influenced by cultural, geographic or regulatory factors tied to the country under consideration (Reale, 2014; Yang, et al., 2015).

It should not be forgotten that open government policies also have a dark side (Zuiderwijk & Janssen, 2014b), and therefore only those open data policies that offer a clear contribution to the decision making processes of the different stakeholders should be deployed. Data protection issues arise when collecting and sharing open data among organizations, an action which may be in conflict with the goals of the data users and holders. Therefore, it is necessary to determine how this collaboration takes place, in a way that is compliant with data protection legislation (van den Broek & van Veenstra, 2018). To be considered useful and valuable, OGD initiatives must comply, at least, with the eight principles of open data (Open Government Working Group, 2007). These

principles are: complete, primary (as in collected at the source), timely, accessible, machine processable, non-discriminatory, non-proprietary and license-free.

Despite the importance of this topic, in their bibliometric research Zhang, Hua, and Yuan (2018) state that OGD is still an emerging topic in the field of open data. Thorsby, Stowers, Wolslegel, and Tumbuan (2017) affirm that more empirical and systematic research needs to be carried out regarding OGD portals (OGDPs), while Yang, et al. (2015) assert that there are still no appropriate metrics to evaluate the success of open data initiatives. In particular, as Charalabidis, Alexopoulos, and Loukis (2016) posited, there is a gap regarding interdisciplinary research, including social and economic aspects, in terms of designing models to face the challenge of anticipating unexpected crises. In addition, socio-technical variables (Davies & Frank, 2013) should be included when strictly technology indicators are deemed insufficient to assess the public value of the portal (Sandoval-Almazan & Gil-Garcia, 2014).

The aim of this research is to bridge this gap, by using technical and socioeconomic indicators to measure the development of the OGDPs of the 28 countries of the EU (EU-28) over three years (2015-2017), and then to carry out a cluster analysis to examine the antecedents of their development. Our contribution to the OGD literature, therefore, lies in providing a new multidisciplinary analysis framework for OGDPs that attempts to advance knowledge with regard to the understanding of national OGDPs' expectations and future considerations. Finally, the implications of this study will benefit policymakers, open data users and researchers when working on the construction of valuable OGDPs, and formulating public policies to empower open data stakeholders.

This paper is structured as follows. Following this introduction, a review of the main literature is presented. The research methodology is then described. The next section contains the results and discussion. The paper finishes with some conclusions, research limitations and considerations for stakeholders.

2. Related works

Since its inception, a number of researchers have developed evaluation models for OGDP, focusing on different aspects such as the maturity of the portal and data usability (Veljković, et al., 2014) or the level of implication for stakeholders (Sayogo, Pardo, & Cook, 2014), all of which relates to public efficiency and openness. dos Santos Brito, Silva Costa, Cardoso Garcia, & Romero de Lemos Meira (2014) and Lourenço (2015) focused on the transparency and accountability features of the data in the portals, which they believed would translate into government

transparency and efficiency. Metadata quality, and data content are also often considered as benchmarking tools (Martin, Foulonneau, & Turki, 2013; Reiche & Höfig, 2013; Younsi Dhabi, Lamharhar, & Chiadmi, 2018). Máchová and Lnénicka (2017) posit that the level of sophistication of the portal itself is, in addition to the number of datasets and the quality of the data, positively related to the quality level of the portal, contributing to an increase in its social and economic value. The degree of compliance with the eight OGD principles has also been a way to evaluate national OGDPs using technical indicators to reflect a qualitative measure of OGD benefits (Gomes & Soares, 2014; Sanad, et al., 2018).

Other works have widened the scope and introduced non-technical features into their benchmarking analysis. Zuiderwijk and Janssen (2014a) consider a mix of socioeconomic factors such as environmental context, policy content, performance indicators and compliance with public values, always within the Dutch context. Huijboom and van den Broek (2011), which have also considered a mixed approach to the analysis of OGD, have focused on specific successful initiatives to reveal good practices. Moreover, Jetzek, Avital, and Bjorn-Andersen (2019) posit that to assess OGD initatives, both social and market-related mechanisms should be considered in order to create sustainable value for their users.

In addition, several non-academic OGDP rankings/indexes/barometers are available for use. Each of them uses a different algorithm for establishing their benchmarks or rankings, depending on the main points they want to stress. The most relevant ones are, on the one hand, the Open Data Barometer (https://opendatabarometer.org), a product of The Web Foundation, which is based on a survey addressed to experts that considers not only OGDPs, but many other aspects related to open data policies; on the other hand, there is the European Open Data Portal (https://www.europeandataportal.eu/) which offers an annual overview of the maturity of the open data initiatives of the EU-28 from 2016.

In sum, the existent literature has focused on studying the features of the OGDPs, usually by analyzing their technical features; or in analyzing the quality of the portals by introducing social, institutional or economic indicators but from a single country perspective, looking at the OGD best practices to extract lessons for open data policy improvement in a static picture. The systematic analysis of Hossain, Dwidedi, and Rana (2016) indicates that only 10% of the current literature on open data offers a mixed or integrated approach, and most research is of a qualitative nature. In this research we expand on the work of Thorsby, et al. (2017) but a) considering the national OGDPs of the EU-28, b) over a period of three years, and c) using cluster analysis.

3. Research methodology

3.1 Research context

The research method applied in this work was a cross-sectional analysis of the open data portals of the countries of the EU, which is considered an area of continuous growth and high interest in the European political agenda (Capgemini Consulting, 2015). A multiple-country assessment was carried out based on a comparative single-country approach. The initial list of countries was that of the EU-28 countries on July 1st, 2015. This way, we offer a characterization of national portals based on technical and socioeconomic indicators under the same global policy frame (European Parliament and Council of the EU, 2013), because legislation is one the factors that influences most the sharing of information between agencies (Yang, et al., 2015), in spite of internal differences between EU countries (Reale, 2014).

This study combines two main sources of data, the data about the OGDPs and the socioeconomic statistics about the countries included in the research. The analysis does not consider the content of the datasets or applications, but focuses instead on the structural elements of the portals. Since these structures are continuously evolving (Lourenço, 2015), the portals were observed over a period of three years, from July 2015 to December 2017, obtaining data every 13 months for a longitudinal analysis. We have chosen a 13 month period because this allows the data to be de-seasoned and is a period long enough to allow us to observe significant changes.

The information used in the analysis was gathered from the websites of the official OGDPs maintained by a designated agent, regardless of the language of publication and the number of published datasets.

In order to ensure that the portals conformed to the open government quality principles, it was required that they met the following criteria:

- Open data should be offered in a reusable digital format, such as xls, xlm, doc, pdf…
- All datasets (either specifically or in general) should display their license to reuse, republish and/or replicate the data (Open Knowledge Foundation, n.d.).
- Open data should be provided by governments and other PSOs, or in collaboration with groups and individual contributors and developers. That is, content does not have to be generated by a public organization, but its management as part of an open platform should be public.

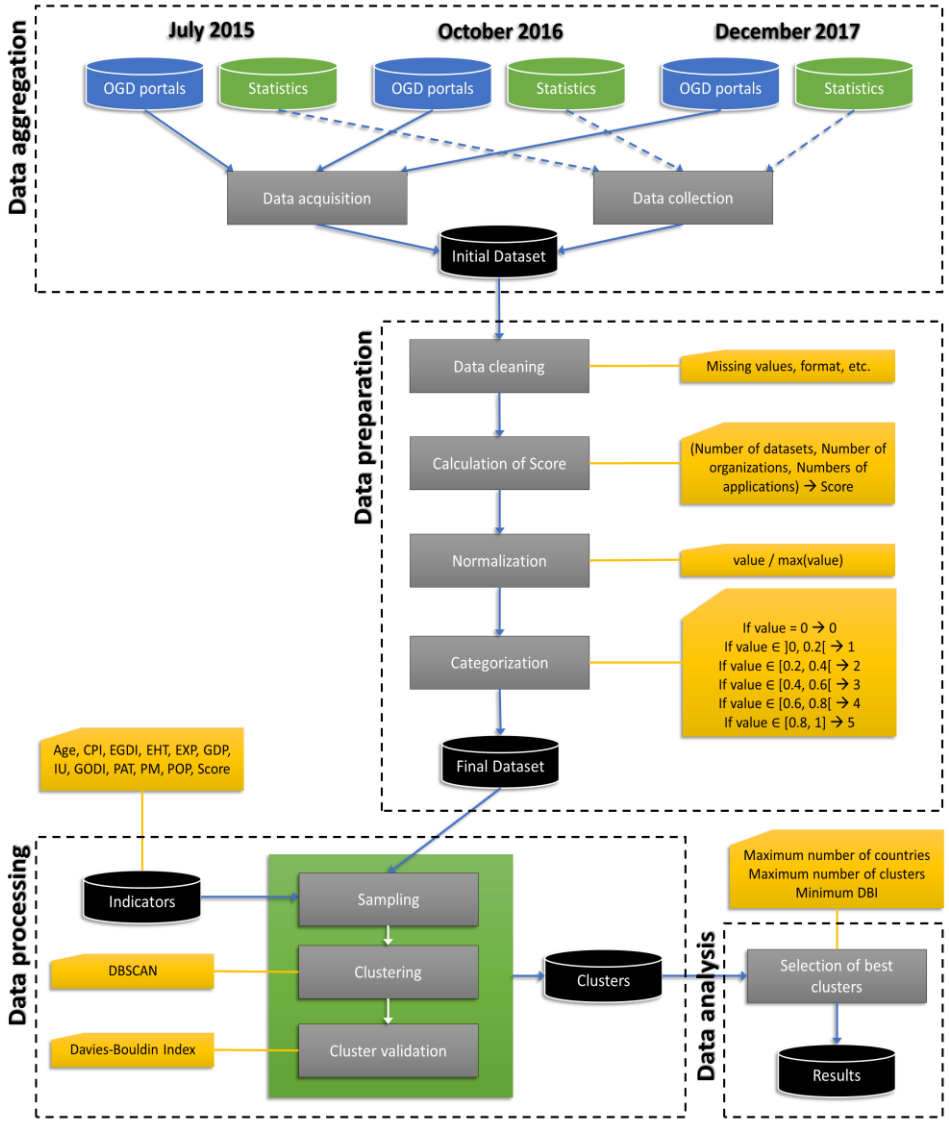
The portals should have a nation-wide, official character: only strictly national portals were considered, in order to be able to carry out the analysis at a comparable level. Unofficial portals (i.e.

open data portals offering datasets related to a country, but not managed by an officially-designated public organization) have been discarded due to the lack of governmental control.

3.2 Cluster analysis

The methodology adopted in this research comprises four main stages: data aggregation, data preparation, data processing and data analysis. A graphical description of the method, with the steps that compose each stage, is depicted in Figure 1. These stages are now described in detail.

Fig. 1. Methodology structured in four stages: data aggregation, data preparation, data processing, and data analysis



Source: own

3.2.1 Data aggregation

Considering the literature on OGD indexes, three parameters were chosen as part of our model to signal the evolution of the portal:

- Number of datasets per 100,000 inhabitants: Following Thorsby, et al. (2017) and Yang and Wu (2016), the number of datasets in relation to the population could be a measure of the possible usefulness of the portal.
- Data reuse (number of applications available): Reuse of data is considered a crucial indicator of OGD success, since there is a symbiotic relationship between users and producers of OGD (Saxena, 2018; Sharon, 2010).
- Organizations participating: This is an indicator that demonstrates the engagement of the country's PSOs in providing content to the portal and co-creating public knowledge as part of the ecosystem, which is largely dependent on the country (Styrin, Luna-Reyes, & Harrison, 2017; Yang & Wu, 2016). Also, most interactions within OGD seem to be on an inter-organizational basis (Yang, et al., 2015).

All data were collected manually by means of an online search of the OGDP of each country[1]. Unfortunately, collecting and normalizing the features of an OGDP is quite challenging, because there does not exist a standard that all countries accept and apply. Therefore, in this study, we only considered the data that could be extracted from all the OGDPs and could be quantified: launch date, number of datasets, number of organizations, and number of applications.

Regarding socioeconomic indicators, or possible antecedents of OGD development, they were chosen in such a way as to evaluate different aspects of the socioeconomic context of the EU-28, and were manually collected from their primary sources. The list of chosen socioeconomic indicators is as follows:

- EGDI: e-Government Data Index. Index 0-1. https://publicadministration.un.org/egovkb/Portals/egovkb/Documents/un/2017-Survey/E-Gov_Complete_Survey-2017.pdf)
- CPI: Corruption Perceptions Index. Index 0-1. https://www.transparency.org/cpi2015#results-table
- GDP: Gross Domestic Product. Millions of euros. http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=nama_10_gdp&lang=en

[1] Data available at http://hdl.handle.net/10045/93447

- PAT: Number of patents per 1,000,000 inhabitants. https://ec.europa.eu/eurostat/web/science-technology-innovation
- EHT: Employment in high- tech, manufacturing and knowledge-intensive service sectors (as % of total employment). https://ec.europa.eu/eurostat/web/science-technology-innovation
- IU: Internet connections over the total population. https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx
- EXP: Expenditure (national budget). Millions of euros. http://appsso.eurostat.ec.europa.eu/nui/submitViewTableAction.do
- POP: Population. Number of inhabitants. https://ec.europa.eu/eurostat/tgm/refreshTableAction.do?tab=table&plugin=1&pcode=tps00001&language=en
- GODI. Global Open Data index. Ranking position. http://2015.index.okfn.org/place/
- PM. Portal Maturity index. Index 0-100. https://www.europeandataportal.eu/es/dashboard#2018

The first two indicators refer to the obvious relationship between e-government and OGDP (Chan, 2013; Hansson, Belkacem, & Ekenberg, 2015), hence the inclusion of the position of the country in EGDI as a measure of government openness, transparency, democracy and will for open technology use. For a more specifically social indicator in this context, the CPI intends to measure the perceived effect of OGD initiatives as promoters of government transparency (Lourenço, 2013, 2015).

Next, according to Yang and Wu (2016), organizational capabilities and facilitating conditions, among other factors, are found relevant for OGD usefulness, thus the inclusion in the list of indicators of a country's GDP as a measure of a facilitator of investment and growth; however, no conclusive evidence was found of the positive relationship between GDP and public transparency and accountability, so this indicator should be considered with caution (Harrison & Sayogo, 2014). Mejabi, et al. (2015) also found evidence of the very high effect government expenditure figures (EXP) had for the publishing organizations, while being not so high for other stakeholders. However, other researchers state that public budget cuts have actually promoted the use of open data to make citizens understand and become involved in political decisions (Huijboom & van den Broek, 2011).

One facilitating condition proposed by Thorsby, et al. (2017) was the degree of innovation of the PSO managing the portal, which was here referred to as PAT, to offer a measure of the degree of innovation within a country. In addition, in their qualitative research, Mejabi, Azee, Adedoyin, and Oloyede (2015) affirmed that the impact of the knowledge of the stakeholders, in terms of how to use open data on the level of institutionalization of the national data initiative, is high. Therefore, EHT was introduced as an indicator of interested and capable external users, as well as IU, since human capital has been observed as an antecedent of governmental transparency, while considering also its role as a measure of the resources available to a country's government.

Finally, other technical variables were considered to reinforce the open data aspect of the analysis, namely the age of the portal (AGE) measured in years since its inception (Thorsby, et al., 2017), and two separate ways to establish the evolution of the open data policies in the analyzed countries, after Reale's (2014) comparative analysis: the GODI and the PM, which apply only to EU-28. The need for different measures stems from trying to consider all the dimensions of open data and their relationship to OGD.

3.2.2 Data preparation

After completing the data acquisition and collection, data cleaning checked the resulting dataset to identify and correct possible errors such as missing values, outlier values or different data formats (Chu, Ilyas, Krishnan, & Wang, 2016). This guarantees the highest degree of data reliability.

Then, *Score*, our OGDP indicator, is calculated for each portal and for each time period. The new metric OGD *Score* is defined as:

$$Score_i = \frac{{NumD_i}/{Max\{NumD_i\}} + {NumO_i}/{Max\{NumO_i\}} + {NumA_i}/{Max\{NumA_i\}}}{3}$$

where *NumDi* is the number of datasets per 100,000 inhabitants, *NumOi* is the number of organizations that collaborate in the portal per 100,000 inhabitants, and *NumAi* is the number of applications published in the portal per 100,000 inhabitants. In this way, *Score* is not an absolute measure, but a relative measure that is used to compare an OGDP against a set of OGDPs.

This calculation is performed in three steps:

1. Calculate the relative number of datasets/organizations/applications per 100,000 inhabitants.
2. Calculate the *ScoreD/ScoreO/ScoreA* of each portal as the relative number of datasets/organizations/applications divided by the maximum relative number of datasets/organizations/applications.
3. Calculate *Score* as the average of *ScoreD*, *ScoreO*, and *ScoreA*.

The components of the dataset present different dimensions and magnitudes; therefore, there is a need to normalize all components for effective comparison. The following equation is used to calculate n_{cpt}, the normalized value of the component *c* of the portal *p* in the time period *t*:

$$n_{cpt} = \frac{v_{cpt}}{Max_{ct}\{v_{cpt}\}}$$

where a larger value n_{cpt} represents a better performance.

Afterwards, the normalized values are categorized into six intervals:

$$cat_{cpt} = \begin{cases} 0 & n_{cpt} = 0 \\ 1 & 0 < n_{cpt} < 0.2 \\ 2 & 0.2 \leq n_{cpt} < 0.4 \\ 3 & 0.4 \leq n_{cpt} < 0.6 \\ 4 & 0.6 \leq n_{cpt} < 0.8 \\ 5 & 0.8 \leq n_{cpt} < 1 \end{cases}$$

At the end of this stage, the final dataset is ready to be used to perform the data processing.

3.2.3 Data processing

Once the final dataset is calculated, an iterative process is used to calculate all the combinations of *k* indicators from the set of *n* indicators. The objective is to test all possible combinations of indicators to find the best clustering solution. The number of combinations is equal to the binomial coefficient:

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1} = \frac{n!}{k!\,(n-k)!}$$

For each combination, the corresponding indicators are extracted from the final dataset to perform the clustering that divides data into groups with similar values. Clustering is useful when data are composed of multiple dimensions.

Clustering algorithms have been extensively studied in the past (Hartigan, 1975; Jain & Dubes, 1988), but as no clustering algorithm offers the best solution to any combination of initial requirements, their analysis is an area of study that is still continuing (Dehuri, Mohapatra, Ghosh, & Mall, 2006; Venkatkumar & Shardaben, 2016). In our method we have selected the density-based spatial clustering of applications with noise (DBSCAN), a clustering algorithm proposed in 1996 (Ester, Kriegel, Sander, & Xu, 1996). DBSCAN is designed to discover clusters of arbitrary shapes and requires only two input parameters: *epsilon*, the maximum distance between two points for them to be considered as being in the same neighborhood; and *minpoints*, the minimum number of points in a neighborhood for a point to be considered as a core point. The basic idea of DBSCAN is

that a neighborhood around a point of a given radius (*epsilon*) must contain at least a minimum number of points (*minpoints*).

Some studies show that DBSCAN outperforms other clustering algorithms such as K-means, SOM and CLARANS (Dehuri, Mohapatra, Ghosh, & Mall, 2006; Ester, Kriegel, Sander, & Xu, 1996). The main benefits of DBSCAN are: minimal requirements of domain knowledge to determine the input parameters, discovery of clusters with arbitrary shapes and good efficiency in dealing with large databases (Ester, Kriegel, Sander, & Xu, 1996).

Because clustering is an unsupervised pattern classification, and domain knowledge is not needed to perform the clustering, the correct partition of the input data is not available, and to determine how the proposed classification fits the input data, i.e. to validate the results obtained, is fundamental. The most commonly-used approaches for cluster validation are based on internal cluster validity indices (CVIs) (Arbelaitz, Gurrutxaga, Muguerza, Pérez, & Perona, 2013). Basically, the idea behind these indices is to measure the compactness and separation of the clusters.

Different studies have evaluated and compared internal CVIs (Arbelaitz, Gurrutxaga, Muguerza, Pérez, & Perona, 2013; Gurrutxaga, Muguerza, Arbelaitz, Pérez, & Martín, 2011). Some works have shown that there is no single CVI that outperforms the rest (Dimitriadou, Dolňicar, & Weingessel, 2002; Maulik & Bandyopadhyay, 2002; Milligan & Cooper, 1985). An extensive study (Arbelaitz, et al., 2013), which compared 30 CVIs in 720 synthetic and 20 real datasets, found that a group of indices such as Calinski–Harabasz, COP, Davies–Bouldin, and Silhouette, perform better than the other analyzed CVIs. This research uses the Davies–Bouldin index (DBI) because it presents the following advantages over other indices that allows it to be used to guide a cluster-seeking algorithm (Davies & Bouldin, 1979): as it does not depend on either the number of clusters analyzed nor the method of partitioning the data, it requires little user interaction, and there is no specification of parameters; it is computationally feasible for large datasets; and it provides meaningful results for data of arbitrary dimensionality. The minimization of the DBI "…appears to indicate natural partitions of data sets" (Davies & Bouldin, 1979, p. 224).

3.2.4 Data analysis

The objective of the final stage is to identify the best clustering solutions from all the calculated clustering results. Once the best clustering solutions are identified, they can be analyzed, visualized and interpreted.

For this research, the following guidelines were specified to draw the clusters:

- To maximize the number of portals (countries) grouped in each cluster, because the objective is to classify as many portals as possible.
- To maximize the number of clusters, because the objective is to make a fine classification of the portals.
- To minimize the DBI, because the objective is to find natural partitions of the portals.

4. Results and discussion

During stage 1 - data aggregation - the data acquisition process generated a dataset of 28 countries by 22 fields and 3 periods. The data collection generated a dataset of 28 countries by 10 fields and 3 periods. With this data we were able to calculate the parameter *Score.*

In Figure 2 it can be seen how *Score* has evolved over the three-year period: the darker the country, the larger the *Score*. The leading countries are consistently the same for the whole period of analysis: Austria, Finland, UK and France. Greece, Luxembourg and Estonia have progressed from nothing to the maximum in these three years, whereas countries like Netherlands, Germany and Spain have become less trendy, meaning that their OGDPs have become less advanced compared to those of their neighbors, and that their OGD policies have stagnated. Slovenia and Ireland have also seen a positive development in these three years. On the other side of the spectrum, there are some countries that *Score* 0, meaning no OGD at all. Therefore, those policies that endeavor to maintain and increase the level of sophistication of the portal and the number of datasets and collaborators, result in higher quality portals, as proposed by Veljković, et al. (2014), Sayogo, et al. (2014), or Máchová and Lnénicka (2017).

Fig. 2: Evolution of indicator *Score* 2015-2017

Source: own

During stage 3 - data processing - the sampling calculated all the combinations of k indicators from the set of 10 socioeconomic indicators (CPI, EGDI, EHT, EXP, GDP, IU, GODI, PAT, PM, and POP) and the two calculated parameters (Age and *Score*), for k from 2 to 6: 66, 220, 495, 792, and 924 combinations. Therefore, the total number of combinations was 2,497 for each period. The clustering was iteratively performed with *epsilon* taking values from 0.5 to 3.0, with steps of 0.1, and *minpoints* taking integer values from 2 to 7. As a result, the total number of clustering executions was 2,497 by 26 by 6 by 3, obtaining a total of 1,168,596 iterations.

Table 1 shows the results of the clustering algorithm after 5 iterations for the period 2015-2017. It can be seen that all of the analyzed technical and socioeconomic indicators are valid antecedents of OGD clusters in relation to our parameter *Score* with one exception: PAT. This would mean that innovation (measured as the number of patents) is not as effective as a possible differencing indicator for OGD development as opposed to the research of Thorsby, et al. (2017).

Take in Table 1

The first interesting outcome of Table 1 is that the number of clusters has decreased over time for all combinations of indicators, meaning that the stances have been polarized. Considering that the values for *epsilon* have been increasing, it is revealed that these (fewer) groups are more internally homogeneous and externally different. As it is, in 2017, the algorithm shows a Europe divided in two, regardless of the number of indicators used for the clustering. The second group is formed by France, Germany and the UK, all three countries of a similar size and considered to be the economic and historic core of the EU-28. This outcome is aligned with the geopolitical significance articulated by Reale (2014) and Yang, et al. (2015).

The more recurrent indicators are related to economic issues. For 2015, national public expenditure (EXP) seems to be the most frequent indicator, as it shows as a parameter for all 5 clustering iterations, although in 2016 it loses some of its relevance since it does not show in any combinations for the best clustering in the first four iterations, being back on the second iteration in 2017. The same pattern, although somewhat less marked, happens to GDP. We can infer, therefore, that financial antecedents as facilitators are related to the development of clusters of similarly-evolved national OGDPs, concurring with some of the findings of Mejabi, et al. (2015) and Yang and Wu (2016). Still, the variability of the economic indicators as main factors supports the need for caution as advised by Harrison and Sayogo (2014).

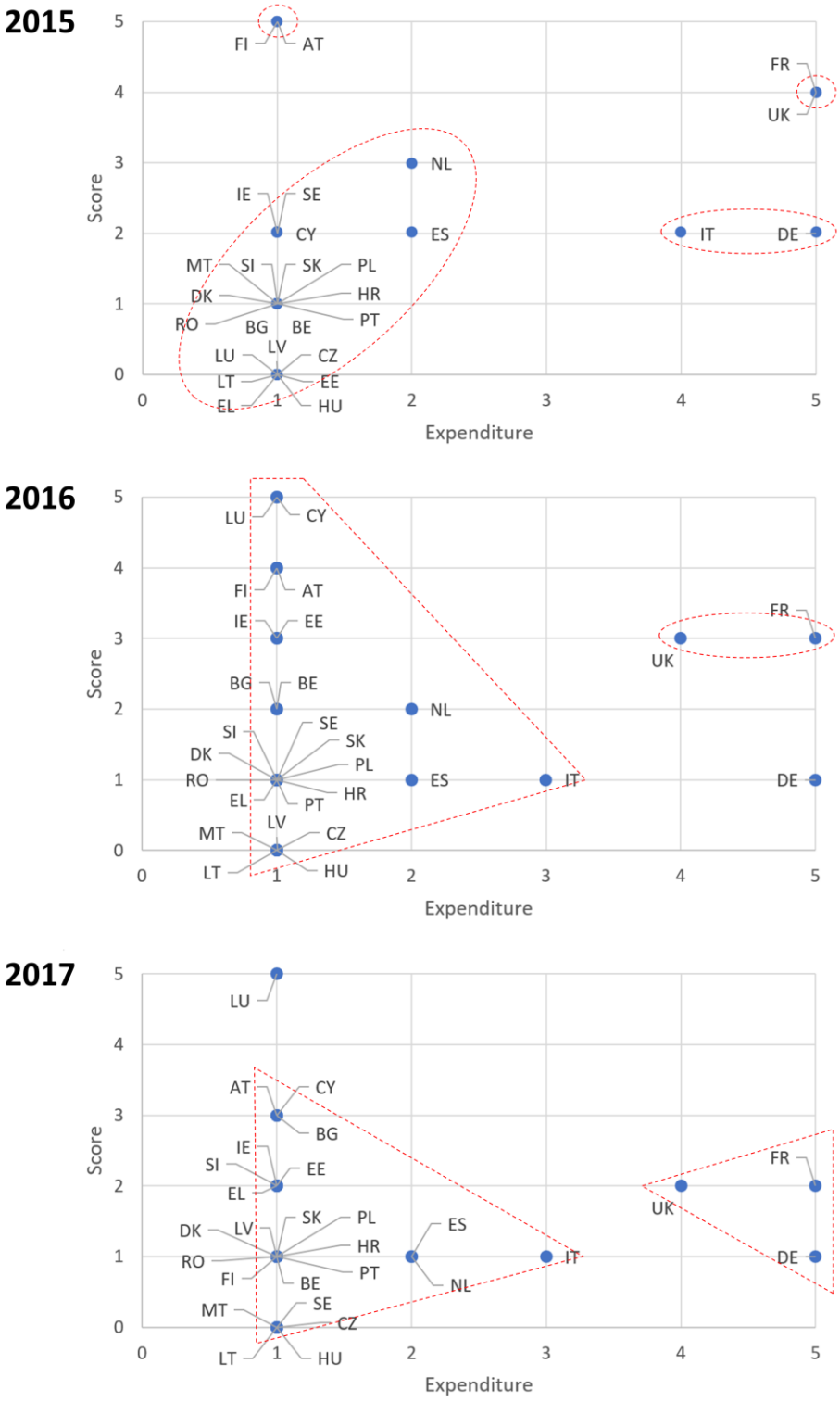
Among the sociopolitical indicators, EGDI and GODI seem to be the main differentiation factors. This is probably due to the European e-Government Action Plan 2016-2020 (https://ec.europa.eu/digital-single-market/en/european-egovernment-action-plan-2016-2020), which aims to scale up the development of truly valuable e-public services that could boost the

single European market. Therefore, the link between e-Government policies, open government and OGD is evidenced, as posited by Attard, et al. (2015), Janssen, et al. (2017) and Lourenço (2013). The second social indicator to consider is the population (POP) of the country, which appears as a differential parameter in all three years, concurring with the literature (Thorsby, et al., 2017), although always in combination with EXP. This again echoes the power of the financial facilitator on OGD development.

In order to better understand the relationship between *Score* and the antecedents, we need to look at the clusters' evolution. Looking at the most relevant indicators, the following relationships have been represented: EXP-*Score* (Figure 3), EGDI-*Score* (Figure 4), GODI-*Score* (Figure 5) and POP-*Score* (Figure 6). Changes in the clusters seem to reveal a polarization trend that splits most of EU-28 into two groups with opposite behaviors, although this partition may be in two ways. The clustering based on an economic factor, EXP (Figure 3) sets apart the leading economic countries of the EU, while the social indicator EDGI (Figure 4) results in a less homogenous but stable cluster of countries that balances a more homogeneous cluster housing the more-e-government advanced countries.

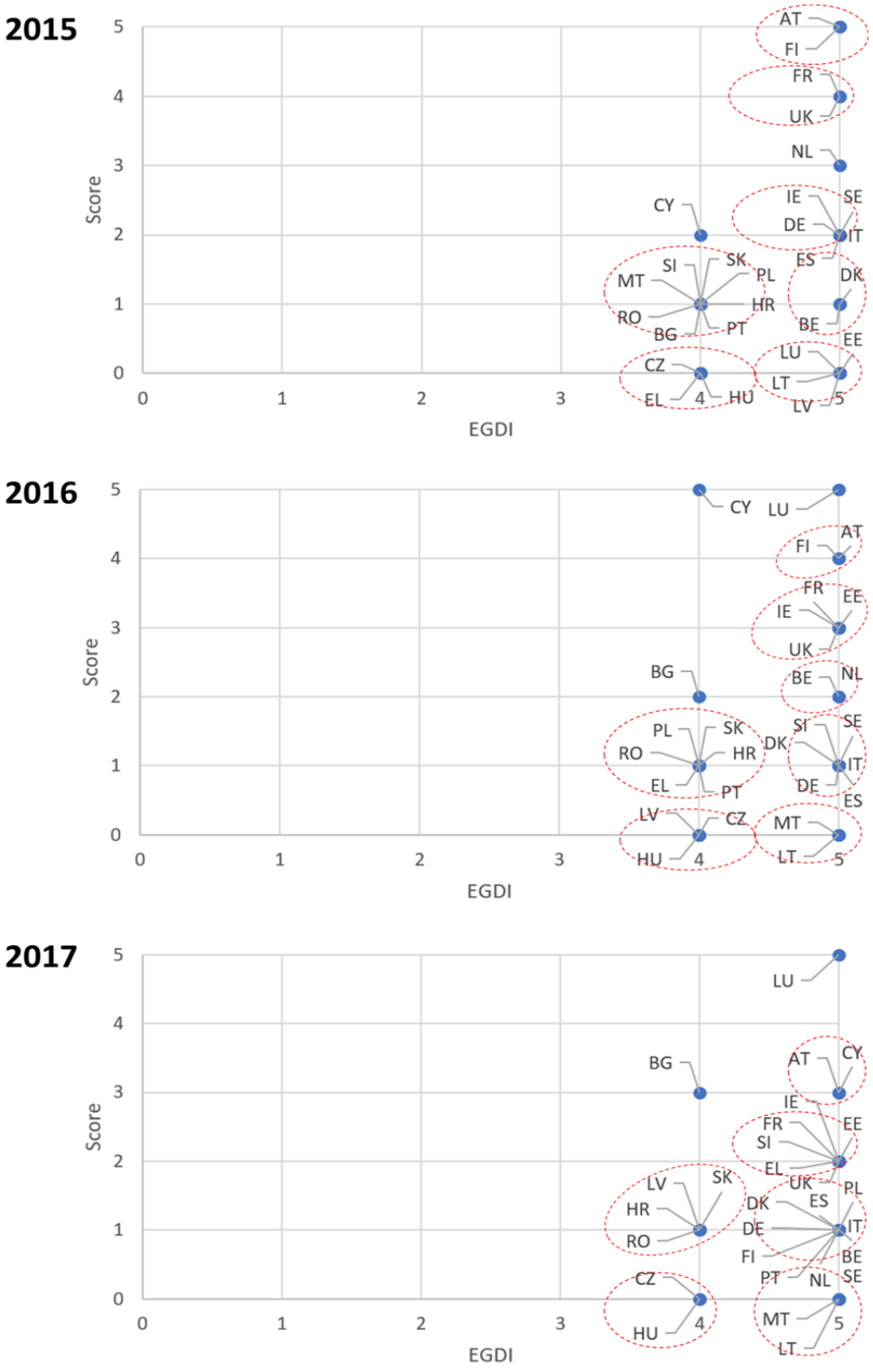
Regarding the other two parameters, the relationship GODI-*Score* (Figure 5) shows a two-cluster formation, but in this case the results could be misleading, since one of the clusters is made up of those countries that have a value of GODI=0, meaning that the original database does not offer information for those countries, and thus they cannot be properly considered by the clustering algorithm. Finally, the clustering based on the countries' population, POP (Figure 6) reveals a similar trend to that of EDGI's (Figure 4) although less defined, while stressing the independent behavior of the economic leading countries as shown in Figure 3.

Fig. 3: Cluster evolution for EXP – *Score* 2015-2017



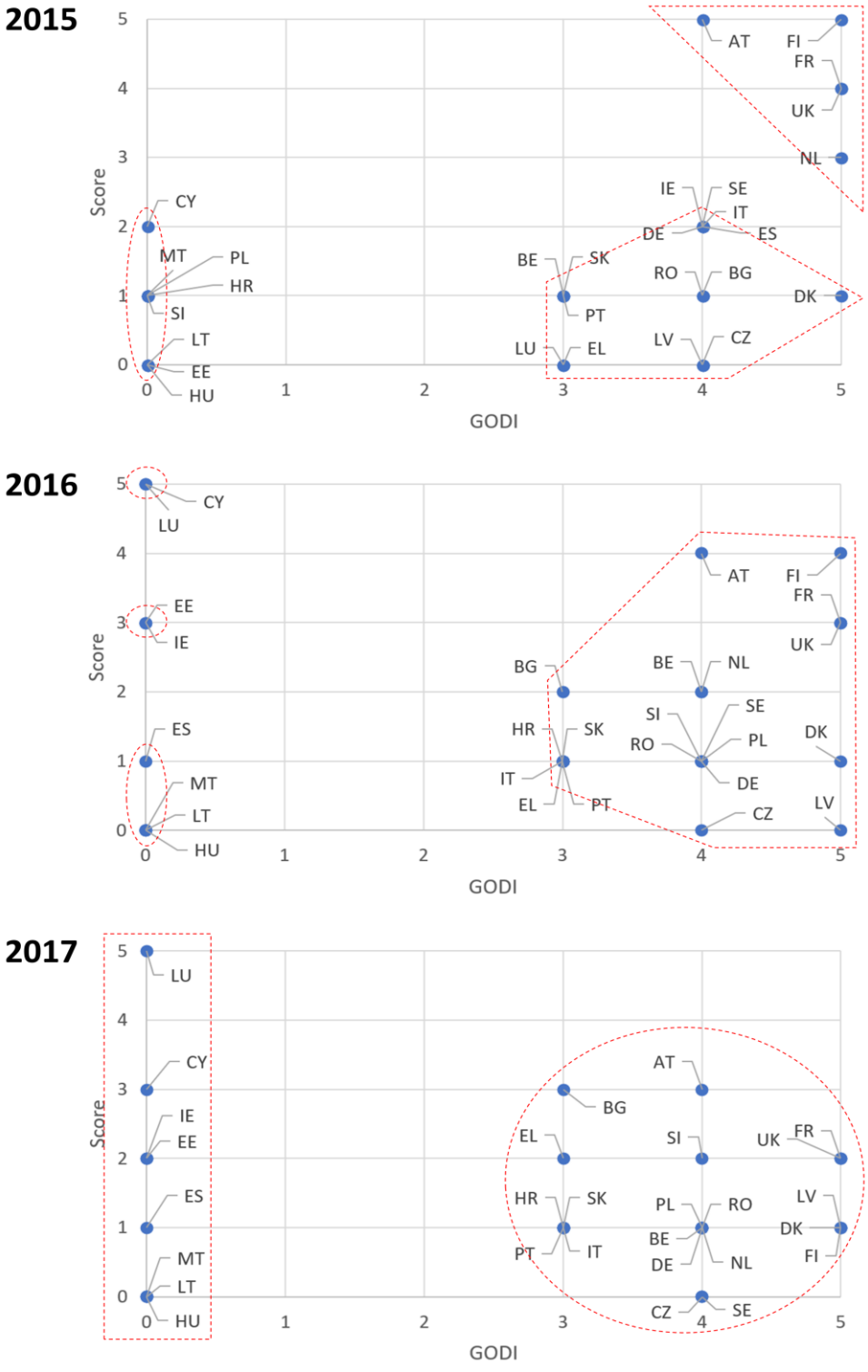
Source: own

Fig. 4: Cluster evolution for EGDI – *Score* 2015-2017



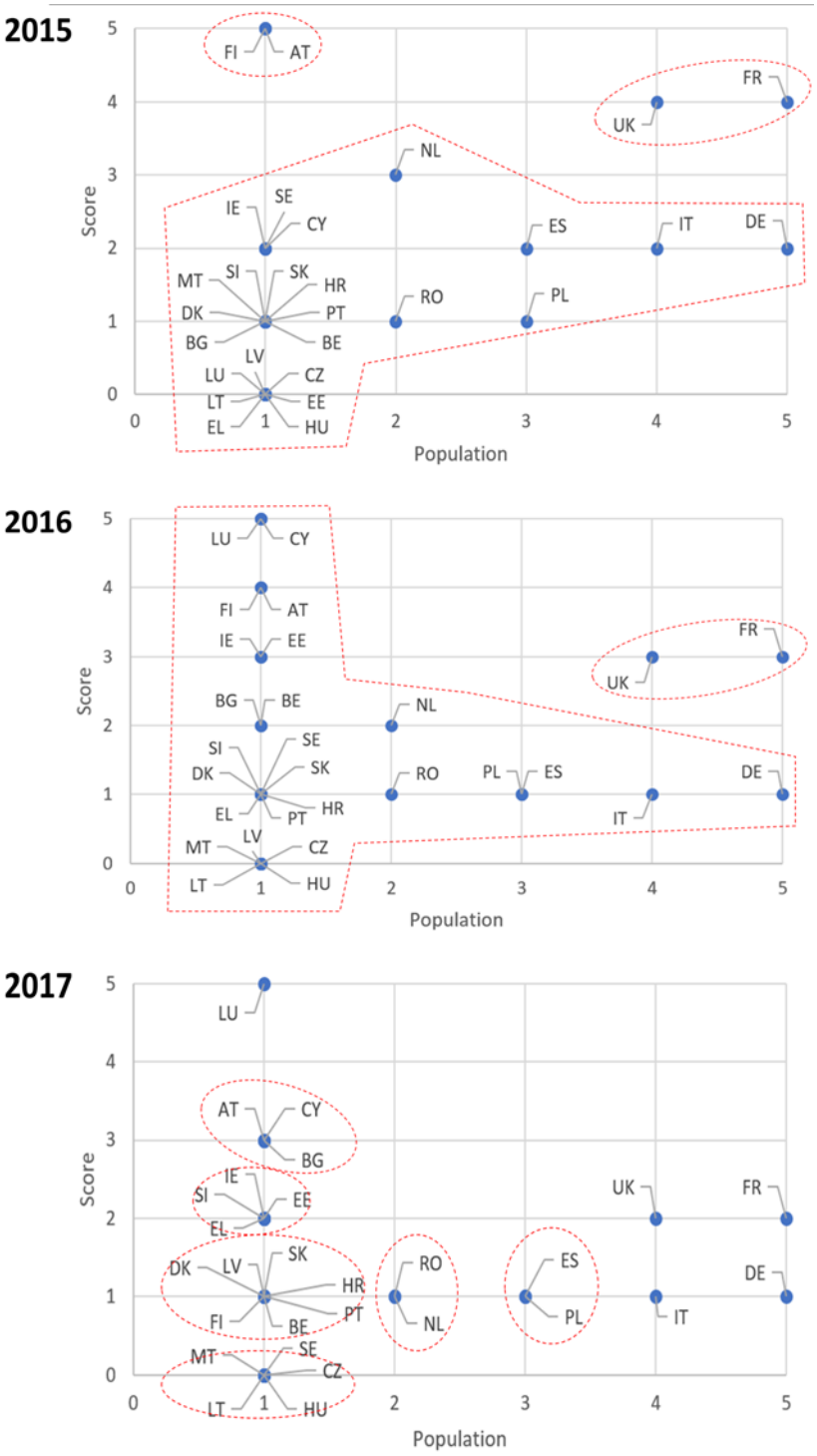
Source: own

Fig. 5. Cluster evolution for GODI – *Score* 2015-2017



Source: own

Fig. 6: Cluster evolution for POP – *Score* 2015-2017



Source: own

5. Conclusions: considerations, development and expectations

This research aimed to contribute to the existing OGD literature regarding assessment frameworks, supporting the development of a more comprehensive assessment tool. By using cluster analysis, the development of several technical indicators over a three-year span and their possible relationship to other socioeconomic factors, was revealed. Policy makers and other stakeholders might be able to use these results to reflect on how to add value to existing or future OGDPs and to forecast future scenarios.

The EU-28 countries have experimented with different paths in their OGD development since 2015. While there is a group of core, leading countries whose scores are very high, there are three other groups that have shown a meteoric growth, stagnation or simply no interest in national OGD. It would be in the interest of the latter two groups to analyze the behaviors and policies of the faster-evolving ones.

In particular, the results of this research leave a number of factors to be considered, such as acknowledging that money matters: in almost every clustering combination for 2015, 2016 and 2017, GDP, EXP or both can be found. Besides, the link between e-government and OGD is apparent, since EGDI is also a recurrent indicator in all three years, but especially after 2016.

As for considerations with regard to OGD development, and bearing in mind the reflections of Jetzek, et al. (2019), what can be gathered from this research is that the true measure of the quality of the OGDPs is the generation of value to the different stakeholders. Data, and information, are public (such as patents, which do not feature as a significant antecedent, thus it could be a signal of not offering true value to European OGD users) but knowledge is not: it is imperative to understand how these stakeholders create knowledge and benefit from it by re-using public sector data. Best practitioners, like Belgium and Luxembourg, should then be studied in depth by OGDP managers.

Finally, the expectations are that EU-28 countries are slowly homogenizing their OGD approaches, which is line with the EU harmonization process and the building of the single market, although individual countries are still free souls which can keep their idiosyncrasies. Even in those cases where the number of cluster does not decrease, they are getting closer. EU policies should reflect on why there are these discrepancies, and how they could contribute to speeding up the homogenization process.

This research has several limitations. First of all, the *Score* parameter is still a bit limited in nature. More important than the absolute number of datasets is the quality of data (following Charalabidis, et al., 2018, and Vetrò, et al., 2016), which we will consider in future works. In this line of thought, relying on external databases has made it more difficult to obtain consistent results

in terms of the clustering analysis. Indeed, we have not been able to carry out the study beyond December 2017 because, at the time of writing, there are socioeconomic statistics that have not yet been published. Likewise, the lack of data on the part of several countries for some of the indicators (e.g. GODI =0) must be considered when interpreting these results. We also plan to study the feasibility of cross impact models combined with cluster analysis to forecast the evolution of OGDPs, notwithstanding the continuation of the existing research so as to improve the longitudinal analysis with current and future data.

References

Aichholzer, G., & Burkert, H. (2004). *Public sector information in the digital age: Between markets, public management and citizens' rights.* Edward Elgar Publishing.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition, 46*(1), 243-256. doi:10.1016/j.patcog.2012.07.021

Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. *Government Information Quarterly, 32*(4), 399-418. doi:10.1016/j.giq.2015.07.006

Capgemini Consulting. (2015). *Creating value through Open Data.* Publications Office of the European Union.

Chan, C. M. (2013). From Open Data to Open Innovation Strategies: Creating E-Services Using Open Government Data. *Hawaii International Conference on System Sciences*, (pp. 1890-1899). doi:10.1109/HICSS.2013.236

Charalabidis, Y., Alexopoulos, C., & Loukis, E. (2016). A Taxonomy of Open Government Data Research Areas and Topics. *Journal of Organizational Computing and Electronic Commerce, 26*(1-2), 41-63. doi:10.1080/10919392.2015.1124720

Charalabidis, Y., Zuiderwijk, A., Alexopoulos, C., Janssen, M., Lampoltshammer, T., & Ferro, E. (2018). *The World of Open Data: Concepts, Methods, Tools and Experiences.* Springer. doi:10.1007/978-3-319-90850-2

Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data Cleaning: Overview and Emerging Challenges. *International Conference on Management of Data*, (pp. 2201-2206). doi:10.1145/2882903.2912574

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1*(2), 224-227. doi:10.1109/TPAMI.1979.4766909

Davies, T., & Frank, M. (2013). There's no such thing as raw data': exploring the socio-technical life of a government dataset. *Annual ACM Web Science Conference*, (pp. 75-78). doi:10.1145/2464464.2464472

Dehuri, S., Mohapatra, C., Ghosh, A., & Mall, R. (2006). A Comparative Study of Clustering Algorithms. *Information Technology Journal, 5*(3), 551-559. doi:10.3923/itj.2006.551.559

Dimitriadou, E., Dolňicar, S., & Weingessel, A. (2002). An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika, 67*(1), 137-159. doi:10.1007/BF02294713

dos Santos Brito, K., Silva Costa, M., Cardoso Garcia, V., & Romero de Lemos Meira, S. (2014). Experiences Integrating Heterogeneous Government Open Data Sources to Deliver Services and Promote Transparency in Brazil. *Annual Computer Software and Applications Conference*, (pp. 606-607). doi:10.1109/COMPSAC.2014.87

Edelmann, N., Höchtl, J., & Sachs, M. (2012). Collaboration for Open Innovation Processes in Public Administrations. In Y. Charalabidis, & S. Koussouris, *Empowering Open and Collaborative Governance* 21-37. Springer. doi:10.1007/978-3-642-27219-6_2

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *International Conference on Knowledge Discovery and Data Mining*, 226-231.

European Parliament and Council of the EU. (2003). *Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information.* Retrieved January 25, 2019, from http://data.europa.eu/eli/dir/2003/98/oj

European Parliament and Council of the EU. (2013). *Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information Text with EEA relevance.* Retrieved January 25, 2019, from http://data.europa.eu/eli/dir/2013/37/oj

Gomes, A., & Soares, D. (2014). Open government data initiatives in Europe: northern versus southern countries analysis. *8th International Conference on Theory and Practice of Electronic Governance*, 342-350. doi:10.1145/2691195.2691246

Government of Canada. (2014). *Directive on Open Government.* Retrieved January 25, 2019, from https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=28108

Gurrutxaga, I., Muguerza, J., Arbelaitz, O., Pérez, J. M., & Martín, J. I. (2011). Towards a standard methodology to evaluate internal cluster validity indices. *Pattern Recognition Letters, 32*(3), 505-515. doi:10.1016/j.patrec.2010.11.006

Hansson, K., Belkacem, K., & Ekenberg, L. (2015). Open Government and Democracy: A Research Review. *Social Science Computer Review, 33*(5), 540-555. doi:10.1177/0894439314560847

Harrison, T., & Sayogo, D. S. (2014). Transparency, participation and accountability practices in open government: A comparative study. *Government Information Quarterly, 31*(4), 513-525. doi:10.1016/j.giq.2014.08.002

Hartigan, J. (1975). *Clustering Algorithms.* Wiley.

Hossain, M., Dwidedi, Y., & Rana, N. (2016). State-of-the-art in open data research: Insights from existing literature and a research agenda. *Journal of Organizational Computing and Electronic Commerce, 26*(1-2), 14-40. doi:10.1080/10919392.2015.1124007

Huijboom, N., & van den Broek, T. (2011). Open data: An international comparison of strategies. *European Journal of EPractice, 12*(1), 1-13.

Jain, A., & Dubes, R. (1988). *Algorithms for Clustering Data.* Prentice-Hall, Inc.

Janssen, M., Matheus, R., Longo, J., & Weerakkody, V. (2017). Transparency-by-design as a foundation for open government. *Transforming Government: People, Process and Policy, 11*(1), 2-8. doi:10.1108/TG-02-2017-0015

Jetzek, T., Avital, M., & Bjorn-Andersen, N. (2019). The Sustainable Value of Open Government Data. *Journal of the Association for Information Systems, 20*(6), Art. No. 6. doi:10.17705/1jais.00549

Kalampokis, E., Tambouris, E., & Tarabanis, K. (2011). A classification scheme for open government data: Towards linking decentralised data. *International Journal of Web Engineering and Technology, 6*(3), 266–285. doi:10.1504/IJWET.2011.040725

Lourenço, R. P. (2013). Open Government Portals Assessment: A Transparency for Accountability Perspective. *International Conference on Electronic Government*, 62-74. doi:10.1007/978-3-642-40358-3_6

Lourenço, R. P. (2015). An analysis of open government portals: A perspective of transparency for accountability. *Government Information Quarterly, 32*(3), 323-332. doi:10.1016/j.giq.2015.05.006

Máchová, R., & Lnénicka, M. (2017). Evaluating the Quality of Open Data Portals on the National Level. *Journal of theoretical and applied electronic commerce research, 12*(1), 21-41. doi:10.4067/S0718-18762017000100003

Martin, S., Foulonneau, M., & Turki, S. (2013). 1-5 Stars: Metadata on the Openness Level of Open Data Sets in Europe. *Research Conference on Metadata and Semantic Research*, 234-245. doi:10.1007/978-3-319-03437-9_24

Maulik, U., & Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(12), 1650-1654. doi:10.1109/TPAMI.2002.1114856

Mejabi, O., Azee, A., Adedoyin, A., & Oloyede, M. (2015). Challenges to Open Data institutionalisation: Insights from stakeholder groups in Nigeria. *Open Data Research Symposium.* Ottawa.

Milligan, G., & Cooper, M. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*(2), 159-179. doi:10.1007/BF02294245

Office of the President. (2009). *Memorandum on Transparency and Open Government.* Retrieved January 25, 2019, from https://obamawhitehouse.archives.gov/open/documents/open-government-directive

Open Government Working Group. (2007). *8 Principles of Open Government Data*. Retrieved January 27, 2019, from opengovdata.org: https://public.resource.org/8_principles.html

Open Knowledge Foundation. (n.d.). *Open Data.* Retrieved January 15, 2019, from Open Data Handbook: http://opendatahandbook.org/glossary/en/terms/open-data/

Open Knowledge Foundation. (n.d.). *Open Definition 2.1*. Retrieved January 27, 2019, from The Open Definition: http://opendefinition.org/od/2.1/en/

Parks, W. (1957). The Open Government principle: Applying the right to know under the constitution. *The George Washington Law Review, 26*(1), 1-22.

Reale, G. (2014). Opportunities and Differences of Open Government Data Policies in Europe. *Athens Journal of Social Sciences, 1*(3), 195-206. doi:10.30958/ajss.1-3-3

Reiche, K. J., & Höfig, E. (2013). Implementation of Metadata Quality Metrics and Application on Public Government Data. *Annual Computer Software and Applications Conference Workshops*, 236-241. doi:10.1109/COMPSACW.2013.32

Sanad, A., Usman, T., Babur, H. M., Fariha, H., Kinza, A., & Irm, S. (2018). Dimensions of Open Government Data Web Portals: A Case of Asian Countries. *International Journal of Advanced Computer Science and Applications, 9*(6), 459-469. doi:10.14569/IJACSA.2018.090663

Sandoval-Almazan, R., & Gil-Garcia, J. R. (2014). Towards an Evaluation Model for Open Government: A Preliminary Proposal. *International Conference on Electronic Government*, 47-58. doi:10.1007/978-3-662-44426-9_4

Saxena, S. (2018). Drivers and barriers to re-use Open Government Data (OGD): a case study of open data initiative in Philippines. *Digital Policy, Regulation and Governance, 20*(4), 358-368. doi:10.1108/DPRG-08-2017-0045

Sayogo, D., Pardo, T., & Cook, M. (2014). A framework for benchmarking open government data efforts. *Hawaii International Conference on System Sciences*, 1896-1905. doi:10.1109/HICSS.2014.240

Sharon, D. S. (2010). Stewardship and usefulness: Policy principles for information-based transparency. *Government Information Quarterly, 27*(4), 377-383. doi:10.1016/j.giq.2010.07.001

Styrin, E., Luna-Reyes, L. F., & Harrison, T. M. (2017). Open data ecosystems: an international comparison. *Transforming Government: People, Process and Policy, 11*(1), 132-156. doi:10.1108/TG-01-2017-0006

Thorsby, J., Stowers, G. N., Wolslegel, K., & Tumbuan, E. (2017). Understanding the content and features of open data portals in American cities. *Government Information Quarterly, 34*(1), 53-61. doi:10.1016/j.giq.2016.07.001

van den Broek, T., & van Veenstra, A. F. (2018). Governance of big data collaborations: How to balance regulatory compliance and disruptive innovation. *Technological Forecasting and Social Change, 129*, 330-338. doi:10.1016/j.techfore.2017.09.040

Veljković, N., Bogdanović-Dinić, S., & Stoimenov, L. (2014). Benchmarking open government: An open data perspective. *Government Information Quarterly, 31*(2), 278-290. doi:10.1016/j.giq.2013.10.011

Venkatkumar, I. A., & Shardaben, S. J. (2016). Comparative study of data mining clustering algorithms. *International Conference on Data Science and Engineering*, 1-7. doi:10.1109/ICDSE.2016.7823946

Yang, T.-M., & Wu, Y.-J. (2016). Examining the socio-technical determinants influencing government agencies' open data publication: A sudy in Taiwan. *Government Information Quarterly, 33*(3), 378-392. doi:10.1016/j.giq.2016.05.003

Yang, T.-M., Lo, J., & Shiang, J. (2015). To open or not to open? determinants of open government data. *Journal of Information Systems, 41*(5), 596–612. doi:10.1177/0165551515586715

Yang, Z., & Kankanhalli, A. (2013). Innovation in Government Services: The Case of Open Data. *International Working Conference on Transfer and Diffusion of IT*, 644-651. doi:10.1007/978-3-642-38862-0_47

Younsi Dhabi, K., Lamharhar, H., & Chiadmi, D. (2018). Exploring dimensions influencing the usage of Open Government Data Portals. *12th International Conference on Intelligenent Systems: theories and applications*, 1-6. doi:10.1145/3289402.3289526

Zhang, Y., Hua, W., & Yuan, S. (2018). Mapping the scientific research on open data: A bibliometric review. *Learned Publishing, 31*(2), 95-106. doi:10.1002/leap.1110

Zuiderwijk, A., & Janssen, M. (2014a). Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly, 31*(1), 17-29. doi:10.1016/j.giq.2013.04.003

Zuiderwijk, A., & Janssen, M. (2014b). The negative effects of open government data - investigating the dark side of open data. *Annual International Conference on Digital Government Research*, 147-152. doi:10.1145/2612733.2612761

Funding acknowledgement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Table 1: All countries, indicators always combined with *Score*, lower DBI, per year

2015	Indicator	*Epsilon*	DBI	# of clusters	Clusters
2	EXP	1.1 – 1.9	0.263887356	4	1. AT FI 2. BE BG HR CY CZ DK EE EL HU IE LV LT LU MT NL PL PT RO SK SI ES SE 3. FR UK 4. DE IT
3	EGDI EXP	1.1 – 1.9	0.309189882	4	1. AT FI 2. BE BG HR CY CZ DK EE EL HU IE LV LT LU MT NL PL PT RO SK SI ES SE 3. FR UK 4. DE IT
4	EGDI EXP POP	2.0 – 2.2	0.36854063	3	1. AT FI 2. BE BG HR CY CZ DK EE EL HU IE LV LT LU MT NL PL PT RO SK SI ES SE 3. FR DE IT UK
5	EGDI EHT EXP GDP	2.3 – 2.4	0.411790456	3	1. AT FI 2. BE BG HR CY CZ DK EE EL HU IE IT LV LT LU MT NL PL PT RO SK SI ES SE 3. FR DE UK
6	EHT EXP GDP IUI POP	2.7 – 2.9	0.432498477	2	1. AT BE BG HR CY CZ DK EE FI EL HU IE IT LV LT LU MT NL PL PT RO SK SI ES SE 2. FR DE UK
2016	**Indicator**	**Epsilon**	**DBI**	**# of clusters**	**Clusters**
2	GODI	1.5 – 1.9	0.312322424	4	1. AT BE BG HR CZ DK FI FR DE EL IT LV NL PL PT RO SK SI SE UK 2. CY LU 3. EE IE 4. HU LT MT ES
3	CPI PM	1.5 – 2.2	0.381135965	2	1. AT BE BG HR CY CZ DK EE FI FR DE EL HU IE IT LT LU NL PL PT RO SK SI ES SE UK 2. LV MT
4	Age CPI PM	1.8 – 2.2	0.438483082	2	1. AT BE BG HR CY CZ DK EE FI FR DE EL HU IE IT LT LU NL PL PT RO SK SI ES SE UK 2. LV MT
5	EHT EXP GDP POP	2.3 – 2.8	0.495345307	2	1. AT BE BG HR CY CZ DK EE FI EL HU IE IT LV LT LU MT NL PL PT RO SK SI ES SE 2. FR DE UK
6	EHT EXP GDP IU POP	2.3 – 2.9	0.508936778	2	1. AT BE BG HR CY CZ DK EE FI EL HU IE IT LV LT LU MT NL PL PT RO SK SI ES SE 2. FR DE UK
2017	**Indicator**	**Epsilon**	**DBI**	**# of clusters**	**Clusters**
2	GODI	2.0 – 2.9	0.592144808	2	1. AT BE BG HR CZ DK FI FR DE EL IT LV NL PL PT RO SK SI SE UK 2. CY EE HU IE LT LU MT ES
3	EXP POP	2.0 – 2.2	0.423727856	2	1. AT BE BG HR CY CZ DK EE FI EL HU IE IT LV LT LU MT NL PL PT RO SK SI ES SE 2. FR DE UK
4	EXP GDP Population	2.0 – 2.4	0.384918969	2	1. AT BE BG HR CY CZ DK EE FI EL HU IE IT LV LT LU MT NL PL PT RO SK SI ES SE 2. FR DE UK
5	EGDI	2.0 – 2.4	0.401863723	2	1. AT BE BG HR CY CZ DK EE FI EL HU IE IT LV

	EXP GDP Population				LT LU MT NL PL PT RO SK SI ES SE 2. FR DE UK
6	EGDI EXP GDP IU Population	2.0 – 2.6	0.417361248	2	1. AT BE BG HR CY CZ DK EE FI EL HU IE IT LV LT LU MT NL PL PT RO SK SI ES SE 2. FR DE UK

Source: own