FG Federal Governance a graduate journal

Forum of Federations
THE GLOBAL NETWORK ON FEDERALISM

CLASSIFYING CASES IN FEDERAL STUDIES.

AN ILLUSTRATION OF WHY POLITICAL SCIENTISTS SHOULD DO MORE CLUSTER ANALYSIS

by Johanna Schnabel and Damien Wirths

Johanna Schnabel, Institut d'Etudes Politiques, Historiques et Internationales (IEPHI), Université de Lausanne, Switzerland
Email: JohannaMaria.Schnabel@unil.ch

Damien Wirths
Institut de Hautes Etudes en Administration Publique (IDHEAP), Université de Lausanne, Switzerland
Email: Damien.Wirths@unil.ch

Abstract: Typologies are widely used in research on federalism, e.g. to distinguish dual from cooperative or coming-together from holding-together federations. More general, ideal types, archetypes and categories are frequently used in political science research to define concepts and classify cases. As recently as in 2014, Filho et al. pointed out that Cluster Analysis is still hardly used when it comes to developing typologies in political science. Rather, political scientists rely on more intuitive methods or factor analysis. Our paper argues that Cluster Analysis is of great usefulness because it a) focuses on the relationship between cases and not variables and b) draws on empirical data when identifying the clusters. This paper proposes to apply this fruitful approach to the field of federalism to exemplify its major heuristic potential. Furthermore, we emphasize that testing the secondary validity is a crucial step. Our paper provides two original examples from comparative federal politics and public management that illustrate the strength of Cluster Analysis both in testing and generating hypotheses through the establishment of typologies. For both examples, the validity of the Cluster Analysis is tested by checking for correlations between the clusters and the distribution of power. Hence, the typologies established through Cluster Analysis not only define our respective dependent variables related to aspects of intergovernmental coordination within federations and the

Schnabel/Wirths, 2016, "Classifiying cases in federal studies. An illustration of why political scientists should do more cluster analysis", *Federal Governance*, vol. 13 no. 1, pp. 68-86.

normative density of evaluation clauses in the Swiss federation, but also offer strong insights in issues of regional autonomy.

Keywords: comparative politics, federalism, cluster analysis, classification, intergovernmental councils, evaluation clauses

Federal Governance
Forum of Federations
325 Dalhousie Street, Suite 700
Ottawa, Ontario K1N 7G2 Canada

ISSN 1923-6158

www.federalgovernance.ca

About Federal Governance

Federal Governance is an online graduate journal on theory and politics of federalism and multi-level governance. Its mandate is to engage the global federalism community and reach out to outstanding graduate students interested in federalism and multi-level governance. By providing a platform for graduate students to have early success in their careers, Federal Governance seeks to promote and sustain interest in federalism and multi-level governance research among graduate students. Allied with the Forum of Federations and founding partner, Institute of Intergovernmental Relations at Queen’s University; Federal Governance aims to contribute to a global dialogue on federalism.

Terms of Use

Federal Governance
Forum of Federations
325 Dalhousie Street, Suite 700
Ottawa, Ontario K1N 7G2 Canada

ISSN 1923-6158

www.federalgovernance.ca

1. Introduction[1]

In this paper, we are promoting Cluster Analysis as a method to produce robust typologies in political science in general and in comparative politics in particular. Specifically, we want to insist on validation of the classification as a crucial step when running a Cluster Analysis in political science. Drawing on original examples of classification, we provide guidance for scholars who want to use Cluster Analysis to construct classifications in comparative politics. We argue that researchers obtain robust classifications if they use the validation step of the method to refine initial classifications they obtain from clustering algorithms. Because we validate the classifications by testing the relationship between our typologies and external variables related to federalism, the focus of our paper is on comparative federal studies.

Indeed, classification is a core element in comparative politics in general. Because "typological work [is] a base for comparison" (von Beyme, 2011, p. 29) comparatists need either to use existing classifications (or typologies) or construct their own classifications to develop "empirically falsifiable explanatory theories" (von Beyme, 2011, p. 29). Cluster Analysis allows creating robust concepts that rely on empirical observations and mathematical algorithms. It is a case-based method of classification that regroups empirical observations into are groups of cases that are more similar to each other than to members of other clusters. The implication is twofold: homogeneity of cases within one cluster and heterogeneity between the clusters (Uprichard, 2009; Wiedenbeck & Züll, 2010).

Although there is a vast literature on Cluster Analysis due to the fact that the method is frequently used in natural sciences, little literature deals with the use of this method in the social sciences. Nevertheless, classifications are widely used in Political Science, in particular in Comparative Politics, because they are useful tools to define concepts or construct categorical variables (Collier, Laporte, & Seawright, 2008). Our aim is to contribute to filling this gap by providing political scientists with more systematic tools to classify cases. In contrast to most textbooks on Cluster Analysis, which frequently neglect this paramount step, we want to insist on validation when running a Cluster Analysis in political science. We want to emphasize the need to test primary and, in particular, secondary validity in order to refine the initial classification. Primary validity consists of significance tests of the independent variables that have served to create the cluster. Secondary validity consists of significance tests between cluster outcomes and variables the classification is expected to explain. This step seeks to confirm that the classification has certain features that one wants it to have. This means that the variable one chooses to test secondary validity is related to the dependent or independent variables of the hypothesis the classification will be used to test. In the examples used in this paper, these external variables relate to aspects of federalism that we seek to explain. With Cluster Analysis primarily being a case-based method, "the interpretation and construction of clusters need to be case driven – 'case driven' in the sense that prior theoretical and empirical knowledge about the case need be incorporated for any adequate construction and interpretation of the clusters" (Uprichard, 2009, p. 139).

[1] We would like to thank Prof. Eugene Horber for his valuable help regarding the methodology and Prof. Philippe Blanchard and Prof. Jean-François Savard for their helpful comments.

Another particularity of Cluster Analysis is that outcomes (i.e. the classification) are not necessarily symmetric. Asymmetry means that they do not always cover all theoretical possible combinations, but types are defined by the actual existence of combinations, meaning that a theoretical combination might not be represented by the clustering outcome.

Our paper is structured as follow: We will first contextualize Cluster Analysis within classifications in political science. Then, we will illustrate the different steps of a robust Cluster Analysis – variable selection, clustering method, validation – insisting on the crucial role of validation of results of Cluster Analyses in political science. Finally, we will present two original examples of classification in comparative politics in order to illustrate the different steps of a robust Cluster Analysis.

2. Classification in Political Science

Classifications are widely used in political science because they are useful tools to define concepts or construct categorical variables. In general, classifications consist of several elements: a general concept, row variables, column variables and, as combinations of these, different types. Often but not necessarily typologies are represented as cross-tabulations where cells represent types. Each type is defined by a certain combination of row and column variables. These different combinations constitute different values of a categorical variable. Moreover, they represent categories or sub-concepts of an overarching concept. In some cases, classifications even establish a hierarchy between these categories (Collier et al., 2008). Furthermore, when assigning cases to the different types identified, one obtains information about similarity and dissimilarity of the objects under study (Romesburg, 1984, p. 2). Given that similarity and dissimilarity are the basis of comparison, classifications are frequently used in comparative politics.

Cluster Analysis corresponds to the two purposes of classification, i.e. concept formation and construction of categorical variables[2]. It is indeed a tool to identify clusters, to develop concepts, but also to reduce data, test hypothesized types and identify homogenous subgroups (Uprichard, 2009). In order for a classification to make sense, it has to fulfill both conditions of mutual exclusiveness and collective exhaustiveness (Collier et al., 2008). These two conditions ultimately guide validation, a step that is crucial when using Cluster Analysis in political science.

As Elman (2005) points out scholars use typologies and classifications frequently but rarely reflect on the way they are constructed. Consequently, research in social sciences still lacks manuals on how to develop typologies and classifications in a systematic and replicable way. At the basis of systematic reflection about the construction of typologies and classifications lies an important distinction, namely the one between *typologies* that are based on theory and *classifications* that are based on empirical observations. Cluster Analysis illustrates the latter while both descriptive and explanatory typologies[3]

[2] In addition to this, scholars use classifications for the purpose of selecting cases. Assigning cases to groups, Cluster Analysis can assist researchers in selecting most similar cases (several or all cases of one cluster) or most different cases (cases of different clusters) in earlier stages of the research process.

[3] Indeed, Elman (2005, p. 296) defines explanatory typologies as classifications that are “based on an explicitly stated theory”.

(Elman, 2005) represent a way to construct typologies in which theoretical knowledge guides the identification of types. In a first step, scholars refer to theory when choosing dimensions and their attributes to be represented in the classification. The second step consists of the assignment of cases to these previously established types. This means that theory-based typologies cover all theoretically possible types. This means that the whole set of parameters of all dimensions are represented in their different unique combinations. This implies that the typology might contain empty cells if one type does not match cases of the real world such as illustrated in Table 1. In empirically based classifications such as Cluster Analysis, however, types are defined by structures in real-world data. Because Cluster Analysis is a method "to organize data into homogenous groups" (Kettenring, 2006, p. 3), classification is based on the data instead of theory and, thus, there are no empty cells. But as Table 2 shows, empirical methods such as Cluster Analysis can have classifications as an outcome that do not represent all theoretically possible combinations because one dimension can be crucial for the definition of one or several clusters but not all. Table 1 and Table 2 illustrate the differences between descriptive/explanatory typologies and classifications produced by cluster analysis using a very simple hypothetical example for the purpose of illustration.

		ELECTORAL SYSTEM	
		PROPORTIONAL VOTE	MAJORITY VOTE
PARTY SYSTEM	TWO-PARTIES SYSTEM	– *empty cell* –	Country A, Country D
	MULTI-PARTIES SYSTEM	Country C, Country E, Country F	Country B

Table 1: Hypothetical example of a descriptive or explanatory typology.

	CLUSTER 1 = Country 1, Country 5	CLUSTER 2 = Country 3, Country 4, Country 6	CLUSTER 3 = Country 2
ELECTORAL SYSTEM	proportional vote	majority vote	
PARTY SYSTEM	multi-parties system	two-parties system	
POLITICAL SYSTEM I	democracy	democracy	autocracy
POLITICAL SYSTEM II	parliamentary	presidential	presidential
FEDERAL-UNITARY	federation		unitary

Table 2: Hypothetical Outcome of a Cluster Analysis

Another aspect of systematic reflection about methods of classification is the distinction between case-oriented and variable-oriented approaches. Even though Cluster Analysis can classify variables, it is more frequently used to classify cases[4]. Hence, we look at Cluster Analysis as a case-oriented approach. This is because the creation of clusters is

[4] We thank Philippe Blanchard for his input on this aspect, see also for instance Uprichard (2009).

based on relationships between the cases instead of the relationship between variables. Consequently, Cluster Analysis is distinct from approaches such as Correspondence Analysis, Discriminant Analysis or Principal Component Analysis, which are mostly used to develop classifications by measuring the similarity of variables instead of cases (Uprichard, 2009). Among these methods of multivariate analysis, Cluster Analysis is the only one that identifies clusters when one has no or very little knowledge about the structure of the data (i.e. if and which clusters can be identified). As Kettenring (2006) points out, other methods of multivariate analysis that classify cases or variables are closer to that end of a continuum between 'no knowledge of clusters' and 'well-known clusters' where cases or variables are assigned to well-known clusters. Moreover, Cluster Analysis not being a probabilistic method, it does not need variables used to fulfill statistical assumptions such as normal distribution.

Whereas Cluster Analysis clearly is an empirically-based, case-oriented method, it can be both inductive and deductive, or structure seeking and structure imposing (Uprichard, 2009, p. 140). Cluster Analysis is always inductive in that "previously *unknown* clusters emerge" (Uprichard, 2009, p. 133). This makes Cluster Analysis different from other methods of classification where cases are assigned to previously established categories (types). Instead, Cluster Analysis both identifies and then defines clusters, and assigns cases to these clusters. Cluster Analysis can also be deductive when variables are chosen and numbers of clusters are determined according to pre-existing theory. The question relevant for all methods of classification now is: why can we actually expect the existence of types? Hence, the null hypothesis is that no types (clusters) exist in the data. To develop an opposite assumption (existence of types (clusters)), scholars can rely on their own intuition and knowledge about cases, or refer to theory and, in particular, existing typologies when wanting to refine them. Additionally, one can even impose differences.

3. The validation issue

The validation of a clustering solution is the confirmation whether the null hypothesis (non-existence of types) has to be rejected. Computing a Cluster Analysis on a data set can be represented as three fundamental steps. These are (1) selection of variables, (2) clustering method and determination of the number of clusters (3) validation of results and interpretation (see below). Since the ultimate goal of clustering is to provide researchers or practitioners with meaningful insights, the application of this method to political science leads us to strengthen the validation steps of analysis as an answer to numerous textbooks dealing with numerical taxonomy or data mining. In the following section, we will illustrate how to ensure "the practical significance of results" (Filho and al., 2014) that we will call here the *secondary* validity of the classification (Romesburg, 1984), which completes the *primary* validity of the classification (related to the relevant distribution of cases among clusters and the significance of internal variables). Secondary validity confirms the explanatory power of the classification for example by testing the relationship between the classification and an external variable related to the phenomenon one seeks to explain. Conditions shaping this validity can be considered as "expectations" in the sense of Blatter and Haverland (2014), an umbrella term including *propositions* (causals connections that characterize a paradigm or theory) and *predictions* (concrete observations that we can expect in the empirical world). After testing primary and secondary validity, each cluster has to be interpreted to show the average value of each significant variable. To put it differently, interpretation consists mainly in defining clusters by looking at cluster centers. In this paper, we will illustrate the validation issue by reference to two examples where the explanatory power of the typologies depends on *expectations* related to federalism. This means that in this case, federalism theory is the key feature of validation. Thus, we refer to theories of federalism in order to validate the outcome of the cluster analyses.

4. Methodology

Following Moses, Rihoux, and Kittel (2005), we consider methodology as a toolbox containing tools (methods). In order to produce classification in social science, we propose the following framework which represents instructions to use cluster methods in order to achieve a research objective. These steps are the same for social sciences and natural sciences. But the validation step is more important in social sciences compared to natural sciences.

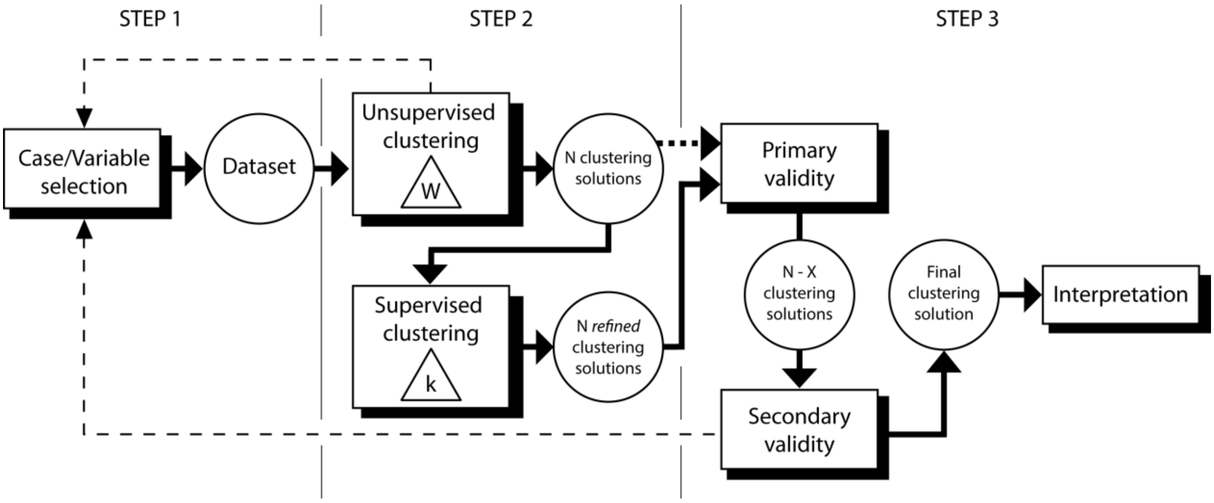

Figure 1: The Three Steps of Cluster Analysis in Political Science

First, scholars select variables from a data set. As highlighted in the literature, Cluster Analysis differs from other statistical techniques in the sense that the data set has not to satisfy mathematical conditions such as normal distribution (Filho and al., 2014). The main issue is to select which variables are used to calculate the similarity between cases (which will mostly depend on data type). Second, the clustering method (also called “algorithm”) has to be determined. Hierarchical methods (also called “unsupervised” in the recent literature) are appropriate when one has expectations about the number of clusters. Hierarchical methods either subsequently agglomerate groups (clusters) up to a single cluster or divide a single cluster into sub-groups until each group consists of only one case. Fusion points indicate numbers of clusters that make sense (initial classification(s)). Non-hierarchical methods (or “supervised methods”) are used when the number of clusters we want to obtain is specified either because one wants to impose a number of clusters or because theory provides a number of clusters. Then, one or more clustering solutions (classifications) can be computed and researchers need to test their validity in order to refine the initial classification (or select the most robust one among several clustering solutions). The main difference between Cluster Analysis in natural and social sciences is that in social sciences such as political science this last step is of paramount importance. This difference is mainly due to the huge number of concurrent paradigms in humanities leading to more complexity and contingency in the validation and interpretation of results. Studying power relations, political science has also to deal with an enormous number of factors and variables depending on ontological and epistemological assumptions of the researcher In natural sciences the meaning of a result (like a classification) is related to explicit assumptions (well-established paradigms, natural laws) guiding the research design. These limit the room for interpretation compared to social sciences (Dépelteau, 2000).

4.1. Selecting Variables for Cluster Analysis

The selection of variables has a major impact on the result of any Cluster Analysis because clusters can be very different depending on the variables chosen. In general, variables can be chosen based on theoretical assumptions, more intuitively based on case-knowledge, or a combination of both (Wiedenbeck & Züll, 2010, p. 528). The selection of variables depends on how much theory is available to provide guidance. When scholars can rely on comprehensive theory they can choose core concepts and hypotheses of the theory to guide the selection of variables. The task is more difficult when it comes to selecting variables when little theory or a handful of theoretical assumptions only are available. Here, scholars cannot rely on a previously existing theory to choose their variables. Consequently, they need to select the variables on which they run the Cluster Analysis either intuitively or collect as many variables as possible. Because the latter alternative can result in a very high number of variables it might be useful to run a Principle-Component Analysis (PCA), Correspondence Analysis or Factor Analysis (FA) first to reduce the number of variables. Both methods summarize several correlated variables into new variables, namely uncorrelated (principal) components. The Cluster Analysis is then run on these new variables, which makes it easier to define properties of the clusters once the clusters are identified.

4.2. Clustering Methods

After selecting the variables, one needs to choose the right clustering method (algorithm). As a first step, one needs to decide whether one wants to obtain a pre-defined number of cases or whether the definition of the right number of cases should be part of the Cluster Analysis. If the earlier is the case, scholars can run supervised methods directly, which assign cases to clusters in order to define their nature. If the number of clusters is yet to be determined one has to rely on unsupervised methods before applying supervised methods. Unsupervised methods produce visual representations of how data are merged into clusters and enable scholars to identify a range of possible cluster solutions. They apply specific clustering algorithms that determine the degree of similarity between cases. Most of them are well known from Sneath and Sokal (1973) but the literature has developed several refinements, especially in the field of computer and big data.

We propose a two-stage sequence combining the two different kinds of clustering methods: first unsupervised and second supervised (step 2 in fig. 1). The former identify relevant numbers of possible clusters whereas the latter allocate the cases according to these preselected numbers. With unsupervised methods, an object remains in a cluster once it is assigned (and then a hierarchy can be established to visualize clusters and sub-clusters with a dendrogram). With a supervised method, cluster affiliations change during the process with several iterations until the best partitioning corresponding to the number of clusters defined by the researcher is attained.

Although papers and user manuals often propose a two-step sequence of analysis[5], it is important to mention that this sequence has to fulfill a strong methodological requirement: to be coherent, clustering algorithms of both steps have to be similar. It makes no sense to determine the number of clusters (first step) with an algorithm based on a distance measure (like the euclidian distance) to finally allocate cases maximizing the within-cluster similarity (like k-means) in the second step. Hence, the best solution is to use methods with equivalent algorithms. Since Ward's (1963) hierarchical minimum variance method and k-means both minimize the sum of within-group variances (Lapointe & Legendre, 1994), they can be successfully combined[6]. This is why we suggest a two-step sequence using first Ward's Method and, second, k-means. *N (possible) clustering solutions* will be produced by the unsupervised method. Supervised clustering using k-means has then to be computed for each of them to obtain *N refined clustering solutions* where within-cluster variations are minimized.

The unsupervised method derives a dendrogram from the data matrix using Ward's (1963) minimum variance clustering method. Each node of the tree represents a fusion point between two cases or groups of cases until every case is assigned to at least one cluster. The longer the distance between two fusion points the bigger the difference between groups/cases. Hence, focusing on groups with the biggest differences, several solutions of clustering are possible most of the time. For instance, in their classification of Scotch whiskeys, Lapointe and Legendre (1994) could define two, three, six or 12 subgroups of whiskeys. If no clear partitioning appears in the dendrogram the hypothesis of cluster existence in the data was wrong. Hence, the researcher has to modify the dataset or the variable selection (dotted line from “unsupervised clustering” in Figure 1). From a scientific point of view, the choice of a “partition level” (number of clusters) only depends on the explanatory power that the researcher wants his classification to have. This explanatory power will be determined during validation (step 3).

Finally, note that the proposed procedure loses the hierarchy highlighted during the unsupervised clustering. This means that information on subgroups is lost. The supervised method is not imperative if one seeks information on subgroups as well. In that case, one can go directly to the third step (dotted line from “N clustering solution” in Figure 1).

[5] For instance Burns and Burns (2008)

[6] Other algorithms fulfill the condition of equivalency equally well. However, these algorithms are often only refinements of either Ward's method or k means and do not provide significantly different results (Romesburg 1984).

4.3. Validation of Results

Concretely, the validation consists in variance analysis with the *N cluster solutions* provided by the k-means clustering (which is a categorical variable). The tests will be either ANOVA when the dataset consists of quantitative variables, or Chi-square if they are qualitative. The latter offers the possibility to test the strength of the relationship with an association measure like Cramer’s V. As explained above, two kinds of validations are required.

First, a primary validity has to be tested. This means that one has to test whether all the variables that have served to create the clusters (internal variables) are relevant to discriminate cases among groups. At the end of this first step, the researcher can either dismiss the X solutions in which fewer internal variables are significantly related to the cluster solution, or exclude those variables from the dataset that are not significant. The higher the number of significant variables is the more robust is the cluster solution. The goal of this validation is to determine the distinctiveness of clusters (and on which aspects of the cases). Another indicator of primary validity is the size of clusters. Cluster solutions in which one or more clusters consist of one or two cases only can be eliminated – unless one is interested in outliers or deviant cases. Although most examples of Cluster Analysis stop the analysis here, we have highlighted above that social science needs to test the secondary validity as well.

Hence, the explanatory power of the typology has also to be demonstrated and secondary validity has to be tested between the satisfying solutions (*N-X*) and variables that have not served to create the clusters but that are relevant for the explanatory power of the classification. These external variables have to be relevant for propositions formulated in the research design. This is because we want the classification to have explanatory power regarding a specific research question. Yet, since the classification does not test the hypothesis formulated in the research design itself but only defines a concept of a variable of the research design, one should not run the test of secondary validity on the dependent variable of the research design. What is more, the outcome of the classification could be used to define the dependent variable of the research design itself. In this case, one should not choose the independent variable(s) of the research design to test the explanatory power of the classification for the same reasons. Rather, we suggest selecting a context variable that is related to the dependent or independent variables of the research design. When no solution survives this test, the researcher can go back to dataset and variables selection to reformulate his research design (dotted line from “Secondary validity” in Figure 1). Whether no solution survives, all the same, the null hypothesis (non-existence of types) has to be considered as true. Finally, one can define the characteristics of each cluster during an interpretation phase. Observing the centroids of the groups (the average value of each variable present in the cluster), one can define each significant internal variable and describe the clusters.

5. Classifying Evaluation Clauses in Swiss Laws

Since cantons enjoy autonomy and have their own legal context, the classification of evaluation clauses investigates whether the phrasing of evaluation clauses differs between cantons and levels of government. Evaluation clauses are legal basis requiring a mandatory evaluation of the impact (effectiveness) of a public policy. They represent a crucial form of procedural institutionalization of policy evaluation. Despite their widespread use in developed countries, no standards have yet been established in their phrasing. In the absence of clarification of this research object, questions concerning origins and effects of mandatory evaluation remain impossible to answer. Past studies dealing with this topic systematically neglect the legal feature of the evaluation process and focus on organizational elements instead. In order to be able to compare evaluation clauses both in time and between cantons or levels of government, a classification of evaluation clauses has been developed. This crucial step is part of a broader research project aiming at identifying different types of legal bases of the obligation to evaluate public policies, explore its causes and to examine its implementation[7].

5.1. Selection of Variables

The classification of evaluation clauses in Swiss laws is an example of a topic where strong paradigms are well established that can guide the variable selection. Hence, irrespectively of the legal context, there are necessary elements that should be clarified in every evaluation clause in every canton and on both levels of government in order to avoid ambiguities during the evaluation procedure. These elements represent an identifiable structure in the clause phrasing that we consider here are as the expected normative density i.e. the degree of detail of an act (OFJ, 2007). Recently a new ‘unité de doctrine’[8] has been suggested which allows for formulating evaluation clauses according to eight dimensions (six mandatory and two optional ones) related to the implementation of the evaluation. These items are consistent with the literature on managing the design of policy evaluation (Shaw, Greene, & Mark, 2006, p. 367) and with aspects of evaluation plans to be developed (Rossi, Lipsey, & Freeman, 2003, p. 33).

[7] This study is part of the SynEval research project, funded by the Sinergia program of the Swiss National Science Foundation. SynEval analyses the relationship between different attributes of political systems and the practice and institutionalisation of policy evaluation. Therefore, SynEval addresses the fundamental questions of how policy evaluation in Switzerland is influenced by the Swiss political system, and how policy evaluation in turn influences the Swiss political system. These questions are answered with an innovative and fruitful research track, as attributes of policy evaluation are linked with policy, polity, and politics in a comprehensive approach. More information about SynEval: http://syneval.ch/index.php/en/

[8] It is recommended that fundamental aspects of the evaluation process have to be specified in every future evaluation clause and that past clauses that do not satisfy the criteria have to be modified when the related laws are debated again

CHARACTERISTICS OF CLAUSES	
FORMAL FEATURES	SUBSTANTIAL FEATURES
- Target group of the evaluation results: Which authority has to receive the evaluation results? - Evaluation period: When the evaluation has to be realized? - Authority in charge to present the report: Which authority has to present the evaluation results? - Form of the final product: How do the evaluation results have to be presented? - Authority in charge to do the evaluation: Which authority has to implement the evaluation? (optional)	- Criteria to evaluate: Under which criteria does the object have to be investigated? - Evaluation object: Which aspect has to be examined? - Evaluation goals: What are the goals of the evaluation? (optional)

Table 3: Constitutive Elements of Evaluation Clauses

Theoretically, these eight elements related to the phrasing of evaluation clauses can be divided into two categories. On the one hand, items can belong to a formal dimension of the clause (*how* does the clause have to be implemented). This category is related to the nature of the evaluator-stakeholder relationship and refers to the methods and procedures used to do the evaluation. On the other hand, they can be related to another category referring to a more substantial dimension (*why* is evaluation needed) dealing with the evaluation goals and the questions asked by the policymaker regarding the implementation that the evaluation has to answer.

5.2. Clustering Method and Initial Classification

Regarding evaluation clauses, 319 evaluation clauses (cases) have been classified. The Cluster Analysis was run on the eight variables listed above that were recoded with binary attributes: the presence of a characteristic (code 1) or the absence (code 0). To compute the substantial and formal dimension of clauses, for each observation, we summed the attributes (code 1 or 0) related to both features (substantial, formal) and each of the two dimensions was given the same importance since they have been weighted by the inverse of the number of characteristics in their type (in order to obtain two scaled variables from 0 to 1). Since we had no prior assumptions about the number of groups that should emerge and because we are not interested in sub-groups, the two-stage sequence of analysis consisting of unsupervised Ward's methods and supervised k-means was run. The dendrogram (see appendix) indicates that three (N) clustering solutions were plausible. The graph shows that the clustering becomes too confused with more than five groups and that a four-groups solution doesn't make sense. Therefore, the three possible solutions were imposed on three distinct k-means procedures.

5.3. Validation of Results: Primary Validity

Regarding the primary validity of the classification of evaluation clauses, the significance of the relationship between the eight internal variables and the three cluster solutions was tested with chi-squared significance and Cramér's V tests[9] (Table 4) in order to determinate to what extent they allow to discriminate the clusters (the stronger the relationship, the more the variable is discriminant).

	TWO CLUSTERS	THREE CLUSTERS	FIVE CLUSTERS
TARGET GROUP OF THE EVALUATION RESULT	,846**	,697**	,830**
EVALUATION PERIOD	,423**	,688**	,666**
AUTHORITY IN CHARGE OF PRESENTING THE REPORT	,369**	,283**	421**
AUTHORITY IN CHARGE OF DOING THE EVALUATION	,287**	,288**	,465**
FORM OF THE FINAL PRODUCT	,791**	,619**	,771**
EVALUATION GOALS	,334**	,513**	,622**
CRITERIA OF EVALUATION	,014	,480**	,630**
EVALUATION OBJECT	,151**	,367**	,532**

Table 4: Cramér's V tests of the relationship between internal variables and cluster outcomes (indicating a significance level of < 0.05)**

It turned out that the results of this procedure are invariably highly significant for all cluster solutions but the two-clusters one. This means that this is not a sufficient meaningful distinction and that the two-clusters solution can be dismissed. In contrast, the three-clusters and five-clusters solution both produce outcomes that are significantly different (distinctiveness) in all dimensions. Thus, the secondary validity of these two solutions has to be tested in order to determine the final cluster solution.

[9] Chi-squared significance tests were used because both the internal variables and the cluster outcome variable are categorical variables.

5.4. Validation of Results: Secondary Validity

Clustering solutions with three or five groups satisfy the *primary* validity conditions. Nevertheless, to give explanatory power to the classification, it has to have certain features that are expected in the literature (secondary validity). Switzerland being a federation, these features are related to federalism. This *consistency* test of our typology with literature is twofold. First, comparative studies have suggested that the emergence of evaluation (and more specifically his institutionalization) could be historically explained by three factors: the political constellation, the fiscal situation and the constitutional features (Derlien & Rist, 2002). Consequently, our classification would have to represent these differences. Since, in Switzerland, cantons differ on these dimensions, one can expect differences between cantons regarding the normative density of clauses. Second, several Swiss authors have highlighted that, even though the lower levels of government have important participatory powers in constitutional revision and law making, most of the evaluation activities have been concentrated at the federal level (Bussmann, 2008; Spinatsch, 2002). These authors suggest that the small size of the cantons raises a question of a critical mass for evaluation capacities – and the evaluation clauses at the cantonal level are expected to have a weaker normative density than at the federal level. By the mean of these theoretical statements, we assume that evaluation clauses on both levels of government follow patterns related to their normative density. We should be able to distinguish differences between clauses more or less focused on these two aspects (formal and substantial). Consequently, valid clusters have to satisfy at least the following properties: there needs to be a significant relationship[10] between the cluster solution and a variable 'level of government' and between the cluster solution and a variable 'canton'.

Since the level of government (cantonal or federal) was only significant with the three-cluster solution[11] we had to select this classification and the final outcome (final clustering solution) of the classification of evaluation clauses is that there are three groups in the dataset that have a sufficient level of explanatory power related to the research design linking evaluation clauses to federalism. The meaning of each cluster can be described by its centroids (the average value of each variable present in the cluster). It has been shown above that Cluster Analysis can produce classifications in which not all elements are equally defining for each type. This is, indeed, the case of the classification of evaluation clauses. Two clusters contain only three elements in average; evaluation clauses in these two clusters can thus be defined as weak clauses. In one of these groups, clauses are mainly focused on the timeframe of the evaluation (*WTF* for weak clauses time focused). The second group consists of clauses that focus on evaluation criteria (*WCF* for weak clauses criteria focused). The third group contains an average of six elements. therefore, these clauses are strong. Seeking to shed light on the relationship between federalism and evaluation clauses, a further analysis has been run in order to find out whether (a) certain cluster(s) is more present on either one level of government. Descriptive statistics show that strong clauses are most of the time found in federal laws and that cantonal laws tend to have weaker evaluation clauses (both WTF and WCF), which gives cantons more room for maneuver compared to federal laws.

[10] Tested by the means of Chi-square tests and Cramér's V.

[11] Two-cluster and five-cluster solutions were not significant with $p < 0.05$.

6. Classifying Intergovernmental Councils in Federal States

The second example is part of a study on Intergovernmental Relations (IGR), i.e. systems of Intergovernmental Councils (IGC), in federal systems[12]. Because comparative and systematic research on IGR is still underdeveloped this study aims at classifying IGC in eight federations. IGC such as the Conference of Cantonal Directors in Switzerland, the Council of the Federation in Canada or the Conference of Cultural Ministers in Germany are institutions that provide an arena in which members of governments (of the federal and the subnational level or the latter only) interact to coordinate their policies. No such classification exists yet and most studies on IGR focus on a lower number of cases. In contrast to the classification of evaluation clauses, this classification seeks not to explain differences or similarities between constituent units such as cantons but between federations. The classification seeks to explain differences among federations in terms of commitment to coordination IGC create.

6.1. Selection of Variables

The classification of Intergovernmental Councils (IGC) illustrates a situation in which very little theory is available to guide the selection of variables. Thus, most variables were chosen in an inductive way by looking at statutes and similar documents on the functioning of such councils. But because some previous insights could be taken from previous studies certain variables were chosen in a deductive way. Table 5 illustrates. Another mechanism was used to further channel the selection of variables, namely the construction of an overall concept defining the explanatory character of the classification. Because the classification is expected to serve a particular purpose, namely to identify how and to which extent IGC differ (within and, particularly between federations) in the degree of *commitment to coordination* they create, this theoretical construct provided further guidance for choosing the variables. Commitment to coordination is a new concept developed for the purpose of a research project on Intergovernmental Relations for which the classification of IGC was established. This concept also combines deduction and induction. It deductively draws on theory of federalism and coordination as well as on institutionalist arguments to group the variables chosen into four dimensions. It inductively draws on empirical observations taken from statutes and similar documents. Institutionalization is inspired by to Bolleyer's (2009) study, coordination by research on policy coordination (e.g., Braun, 2008; Peters, 2004) and salience draws on a concept developed by Trench (2006). Variables were grouped into three dimensions without previous PCA or FA but based on theoretical assumptions as well as empirical insights. Hence, this example illustrates iteration between deduction and induction, theory and data.

[12] This study on "Intergovernmental Relations as a Federal Safeguard" is a PhD project conducted at the University of Lausanne, supervised by Dietmar Braun. It looks at IGR from a Rational Choice point of view, defining Intergovernmental Relations as a federal safeguard, i.e. an incentive mechanism preventing opportunistic behavior in a federal context where governments seek to challenge the distribution of power to pursuit their own interests (Bednar, 2009).

COMMITMENT		
INSTITUTIONALIZATION	COORDINATION	SALIENCE
- Decision-making(*) - Chair - Secretariat(*) - Number of Committees(*) - Executive Committee - Members of committees and working groups - Definition of Functions(*) - Document of Establishment - Integration(*)	- Level of Coordination (**) - Bindingness of Outcomes	- Salience of Policy Areas (***) - Circular resolutions - Representation by staff (***)

Table 5: Variables and Dimensions chosen for the Purpose of Classifying Intergovernmental Councils. Variables marked with one asterisk are taken from Bolleyer's (2009) study on IGR and those marked with three asterisks are inspired by Trench's (2006) concept of salience. The level of coordination is a variable that draws on research by Peters (2004) and Braun (2008).

6.2. Clustering Method and Initial Classification

In the case of intergovernmental councils, 192 IGC have been classified. Ward's method produced a dendrogram that pointed to the existence of a range of three different cluster solutions: two clusters, three clusters and four clusters of intergovernmental councils. Consequently, these different cluster solutions were imposed on k-means clustering. An initial classification based on a two-clusters solution, for example, shows that these two clusters differ on several variables: decision-making rules, number of committees, members of commissions and working groups, document of establishment, level of coordination, bindingness of outcomes, circular resolutions, and representation by staff. That clusters do not differ in terms of regularity of meetings, chair, secretariat, existence of an executive committee, definition of functions, integration and policy issues is due to the empirical nature of the classification. Hence, this example of initial classification of IGC is an asymmetric one (see above). However, because the dendrogram suggests a range of three different cluster solutions, the initial classifications had to be refined through tests of primary and secondary validity.

6.3. Validation of Results: Primary Validity

In terms of size and distinctiveness of types of intergovernmental councils, the primary validity of all three cluster solutions can be confirmed. There is no cluster with only a very little number of cases. As for the relationship between internal variables and the final cluster solution, it turned out that it most significant in a four-cluster solution (Table 6).

	TWO CLUSTERS	THREE CLUSTERS	FOUR CLUSTERS
DECISION-MAKING	,915**	,651**	,580**
REGULARITY OF MEETINGS	,258**	,177**	,264**
CHAIR	,443**	,441**	,536**
SECRETARIAT	,166**	,320**	,307**
NUMBER OF COMMITTEES	0,001**	,106	,191**
EXECUTIVE COMMITTEE	,133	,214**	,137
MEMBERS OF COMMISSIONS AND WORKING GROUPS	,476**	,471**	,522**
DEFINITION OF FUNCTIONS	,254**	,505**	,618**
DOCUMENT OF ESTABLISHMENT	,121	,153	,225**
INTEGRATION	,040	,171**	,304**
LEVEL OF COORDINATION	,146	,359**	,396**
BINDINGNESS OF OUTCOMES	,364**	,509**	,491**
POLICY ISSUES	,101	,179**	,214**
CIRCULAR RESOLUTIONS	,251**	,142	,223**
REPRESENTATION BY STAFF	,072	,136	,134

Table 6: Significance texts between internal variables and cluster outcomes (measure used: Cramér's V, ** indicating a significance level of < 0.05)

Thus, both the two-cluster solution and the three-clusters solution were rejected at this stage. The *refined clustering outcome* (N-X) consists of the four-clusters solution only and excludes the variable "representation of staff" because it is not significant for the cluster solution. After refinement of the initial classification, the new outcome of the analysis shows that clusters differ on the following variables: decision-making rules, regularity of meetings, chair, number of committees, members of commissions and working groups, definition of function, document of establishment, level of coordination, bindingness of outcomes, policy issues, and circular resolutions. The classification is still asymmetric because clusters do not differ in terms of secretariat arrangements, the existence of an executive committee and inter-council relations (cf. appendix). Because there is still no cluster of insufficient size, the primary validity of this classification can be confirmed.

6.4. Validation of Results: Secondary Validity

In the case of the classification of Intergovernmental Councils, the classification is supposed to define a dependent variable measuring differences between federations. Hence, it is necessary to test whether the classification makes sense regarding an external variable (not used to generate the cluster) that is related to by not identical with the dependent variable of the study. In this example, the external variable “federation” assigns Intergovernmental Councils to the federal state they are part of. This variable is of interest because the research project the classification was established for aims at shedding light on differences between federations. A chi-square test confirms the hypothesis on the existence of a relationship between the cluster variable and the external variable “federation”: Cramér’s V indicates that this relation is not only extremely significant but also rather strong (0,535)[13]. Consequently, it makes sense to use the classification to measure and explain differences between federations. Because it passes the text of secondary validity, we can conclude that the classification meets its objective. Therefore, the final classification we obtain consists of four categories of intergovernmental councils (see appendix). If we look at intergovernmental councils as institutions, we can use the classification to define an independent variable “strength of the institution”. In this case, two clusters (clusters 1 and 4) consist of strong institutions, councils in cluster 2 score intermediate on institutional strength and cluster 3 contains weak institutions.

7. Conclusion

Most textbooks on Cluster Analysis consider validation as a separate step of the selection of the number of clusters focused on primary validity. The main reason being that outside social science, the aim of many fields using Cluster Analysis is “only” to gather a huge quantity of data into groups (structure seeking). Often, Biology, Medicine, Computer science, and marketing do not need to deal with the meaning of the clusters that emerged from their data mining. Consequently, we emphasize that, in recent years, the main methodological focus has been put into improving algorithm performance (processing larger and larger data sets), and not on the validation criteria. In this paper, suggest a two-step sequence of Ward’s method and k-means clustering within a three-step Cluster Analysis procedure. This procedure emphasizes that the validation step is of a paramount importance in research designs applying Cluster Analysis to social science that have to demonstrate the practical significance of the classifications. In this sense, Cluster Analysis is more structure imposing in social science than in other disciplines. Validation becomes an integral part of the process and leads the researcher to keep or dismiss solutions produced by several methodological possibilities (supervised or unsupervised methods).

After the variable selection and the clustering method, Cluster Analysis can produce several classifications accordingly. For instance, in the evaluation clauses’ example, three classifications were suggested by a single method. In their paper, Filho and his

[13] Because the test of primary validity showed that most independent variables are significant in the four-clusters solution, it is not necessary to test the secondary validity for other cluster solutions.

colleagues have shown that, testing Robert Dahl's typology (1976) and comparing several ways to classify countries according to polyarchy dimensions, "depending on the method used, the classification will be different" (Filho and al. 2014: 2413). Although they considered that the choice "rests on the searchers' ability to connect theoretical expectations and empirical classification", we have argued here that this *ability* (shaped by the ontological and epistemological assumptions supporting the research design) can be strengthened by an appropriate use of the validation step. The practical significance of our paper is that we have shown that testing primary and secondary validity helps to dismiss irrelevant classification: either because the clusters are not enough distinct and internal variables are not sufficiently significant (primary validity) or because the explanatory power is too weak (secondary validity). Both the classification of evaluation clauses and intergovernmental councils illustrate this process of refinement. In both cases, unsupervised clustering has produced a dendrogram suggesting a range of three possible cluster solutions (initial clustering solutions). The primary validity of classifications of intergovernmental councils suggested that two solutions could be dismissed and that one internal variable should be excluded. The secondary validity of the remaining cluster solution (N-X clustering solution) could be confirmed because the classification is able to explain differences between federations, as it is expected to do. Hence, the final clustering solution consists of five distinct clusters. After testing for primary validity, one cluster solution could be dismissed in the case of evaluation clauses (N refined clustering solutions). Tests of secondary validity showed that another clustering solution could be dismissed and that the final classification consists of three distinct clusters. This classification is able to explain differences in the normative density of evaluation clauses both between levels of government and among cantons.

When no solution survives to the validation (no final output), a first and easy conclusion can be is that there is no group in the dataset. Hence, the classification hypothesis, (based on the assumption that data could be grouped into similar clusters) would be wrong and we would have to admit that a classification of the concerned research object is irrelevant. In this case, the researcher can go back to step 1 and change his data set by selecting different variables or cases (Figure 1). Modifying the selection of variables or cases is significant from a theoretical point of view since it means that the theory and/or the expectation supporting the classification were empirically falsified and, in this sense, Cluster Analysis can provide a strong contribution in theory-building.

References

[1] Bednar, J. (2009). *The Robust Federation. Principles of Design*. Cambridge etc.: Cambridge University Press.

[2] Bolleyer, N. (2009). *Intergovernmental Cooperation*. Oxford: Oxford University Press.

[3] Braun, D. (2008). Organising the Political Coordination of Knowledge and Innovation Policies. *Science and Public Policy*, *35*(4), 227–239.

[4] Burns, R. P., & Burns, R. (2008). *Business research methods and statistics using SPSS*: Sage.

[5] Bussmann, W. (2008). The emergence of evaluation in Switzerland. *Evaluation, 14*(4), 499-506.

[6] Collier, D., Laporte, J., & Seawright, J. (2008). Typologies: Forming Concepts and Creating Categorical Variables. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford Handbook of Political Methodology*. Oxford: Oxford University Press, 152–173.

[7] Dahl, R. (1976). *Polyarchy: Participation and Opposition*. New Haven, CT, and London: Yale University Press.

[8] Dépelteau, R. (2000). *La démarche d'une recherche en sciences humaines: De la question de départ à la communication des résultats*. Brussels: De Boeck.

[9] Derlien, H.-U., & Rist, R. C. (2002). Policy evaluation in international comparison. In N. B. a. L. Transaction Publishers (Ed.), *International Atlas of Evaluation* (pp. 439-455).

[10] Elman, C. (2005). Explanatory Typologies in Qualitative Studies of International Politics. *International Organization*, *59*(02), 293–326.

[11] FOJ. (2007). Evaluations rétrospectives *Guide de législation* (3e ed., pp. 163-174). Bern.

[12] Kettenring, J. R. (2006). The Practice of Cluster Analysis. *Journal of Classification*, *23*, 3–30.

[13] Lapointe, F.-J., & Legendre, P. (1994). A classification of pure malt Scotch whiskies. *Applied Statistics*, *43*(1), 237-257.

[14] Moses, J., Rihoux, B., & Kittel, B. (2005). Mapping political methodology: reflections on a European perspective. *European Political Science, 4*(1), 55-68.

[15] Peters, B. G. (2004). The Search for Coordination and Coherence in Public Policy: Return to the Center? Working Paper.

[16] Romesburg, H. C. (1984). *Cluster Analysis for Researchers*. Morrisville, North Carolina: Lulu Press.

[17] Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2003). *Evaluation: A systematic approach*: Sage publications.

[18] Trench, A. (2006). Intergovernmental Relations: In Search of a Theory. In S. L. Greer (Ed.), *Territory, Democracy and Justice. Regionalism and Federalism in Western Democracies*. Houndmills, Basingstoke, Hampshire, New York: Palgrave Macmillan, 224–256.

[19] Shaw, I., Greene, J. C., & Mark, M. M. (2006). *The Sage handbook of evaluation*: Sage.

[20] Spinatsch, M. (2002). Evaluation in Switzerland, Moving toward a Decentralized System. *International Atlas of Evaluation, New Brunswick and London*, 375-391.

[21] Uprichard, E. (2009). Introducing Cluster Analysis. What Can It Teach Us about the Case? In D. Byrne & C. C. Ragin (Eds.), *The SAGE Handbook of Case-Based Methods*. London: SAGE, 132–147.

[22] Von Beyme, K. (2011). The evolution of comparative politics. In D. Caramani (Ed.), *Comparative Politics*. Oxford: Oxford University Press, 23-36.

[23] Wiedenbeck, M., & Züll, C. (2010). Clusteranalyse. In C. Wolf & H. Best (Eds.), *Handbuch der sozialwissenschaftlichen Datenanalyse*. Wiesbaden: VS Verlag für Sozialwissenschaften, 525–552.

[24] Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association, 58*(301), 236-244.