Extraction of business model solution patterns

Few studies using quantitative business model patterns extraction have been published and no best practices have yet emerged. Amshoff *et al.* (2015) applied multidimensional scaling (MDS), Hunke *et al.* (2017) a *k*-medoids clustering algorithm, as well as Curtis (2021), and Holzmann *et al.* (2019) used the IBM SPSS Statistics TwoStep Cluster approach, verified with a latent class analysis (LCA) and an agglomerative hierarchical clustering (AHC) analysis. Further studies can be found, but with variables of scales of measurement other than binary or with a focus on clustering the companies, see e.g. (Camisón and Villar-López, 2010). As mentioned in the main article, the small size of the dataset precludes the use of data-greedy methods. The methods that can be applied to such a small set are mainly MDS and AHC (Suppes *et al.*, 1989, pp. 218-222; Legendre and Legendre, 2012; Desarbo *et al.*, 2008, p. 281). In some cases, the data fit one grouping representation (MDS or AHC) better than the other (Holman, 1972; Pruzansky *et al.*, 1982). For this study, a preliminary review of the results favoured AHC over MDS. In consequence, the present study is based on AHC analysis, while MDS is used for triangulation purpose. To complete the triangulation and ensure that the obtained partition is robust, we will also use co-occurrence network analysis (Borcard *et al.*, 2018, pp. 117-119), also termed community detection or graph clustering (Newman, 2018; Brandes *et al.*, 2008).

A full description of the AHC analysis method and the extraction of solution patterns, as well as the MDS and co-occurrence network analyses, are presented below. This allows any interested person to fully reproduce the analysis. The implementation in the statistics software IBM SPSS Statistics (Version 27) of the AHC and MDS analysis methods is presented at the end of this file (Section S7). The co-occurrence network analysis has been performed in R (version 4.2.0), see Section S7.

The steps taken for the analysis are:

S1. Choice of a proximity measure
S2. Choice of the clustering method
S3. Execution of the clustering method
S4. Selection and internal validation of the clusters
S5. Confirmation with multidimensional scaling (MDS)
S6. Confirmation with co-occurrence network analysis
S7. Implementation in SPSS and R

S1. Choice of a proximity measure

We are interested in measuring the level of co-occurrence or dependency between configuration options. For binary datasets, proximity measures are based on the number of matches, mismatches, or absence of matches between each pair of configuration options. A match between two configuration options means that a company is using both configuration options in its business model; a mismatch means that one company is using only one of the two configuration options; and an absence of matches means that a company is not using any of the configuration options. Many proximity measures for binary datasets have been developed. Choi *et al.* (2010) identified 76 binary proximity measures. Following Legendre and Legendre (2012, p. 269), the elements considered for the choice of a proximity measure were 1) the nature of the study, 2) the mathematical constraints of the methods of analysis (clustering, etc.), and 3) the computational aspects, that is, the simplicity of using the chosen proximity measure. This led us to choose the Sorensen-Dice index, as motivated below.

The Sorensen-Dice index is an asymmetrical proximity measure, that is, a proximity measure that ignores the absence of matches between the configuration options (Legendre and Legendre, 2012). Compared with other asymmetrical proximity measures, the Sorensen-Dice index favours matches above mismatches by giving them twice their weight (Legendre and Legendre, 2012, p. 276). Because we were more interested in the co-occurrences between configuration options than their lack thereof, this suited our study well.

The Sorensen-Dice index limits the choice of specific AHC (and MDS) methods, but this was not problematic for this study, as developed in the next section. The Sorensen-Dice index is available in all statistical software packages, which makes it simple to use in this study or similar studies. The Sorensen-Dice index, named after Sørensen (1948) and Dice (1945), has different names in the literature and is sometimes simply called the Dice index or the Sorensen index.

S2. Choice of a clustering method

The clustering method was mainly chosen by elimination. Several clustering methods for AHC are available, the most commonly used being the single linkage method, the complete linkage method, the unweighted pair group method with arithmetic mean (UPGMA method), the centroid linkage method, the median linkage method, and Ward's method. The centroid linkage method, the median linkage method, and Ward's method presumes the use of a Euclidean distance as a proximity measure (Anderberg, 1973, p. 141; Everitt *et al.*, 2011, p. 79; Romesburg, 1984). The Sorensen-Dice index is related to the Euclidean distance. Therefore, these methods were not considered in this study. Of the remaining clustering methods, the single linkage method often presents the problem of chaining: clusters tend to grow by the addition of single objects (Romesburg, 1984). As we were interested in finding solution patterns, that is, combinations of at least two configuration options, chaining was a potential issue. The complete linkage method is the least likely to create chaining and was thus an interesting option, although this linkage method tends to find compact clusters and is thus

sensitive to the shape of the clusters (Everitt *et al.*, 2011, p. 79). Finally, the UPGMA method lies between single and complete linkage in terms of chaining and is therefore often recommended (Romesburg, 1984, pp. 126-127). Therefore, we selected both the complete linkage method and the UPGMA method for clustering.

S3. Execution of the clustering

Configuration options that were not present in the companies' business models were removed from the analysis, as they could not contribute to the clustering (or MDS analysis). The AHC produced a complete dendrogram showing the complete hierarchy of the configuration options based on their similarities. The dendrograms obtained using the complete linkage and UPGMA methods were relatively similar, with the complete linkage method presenting less chaining, as expected. The obtained dendrograms are shown in Figure S1. The detailed execution of the cluster analysis using SPSS is given in Section S7.

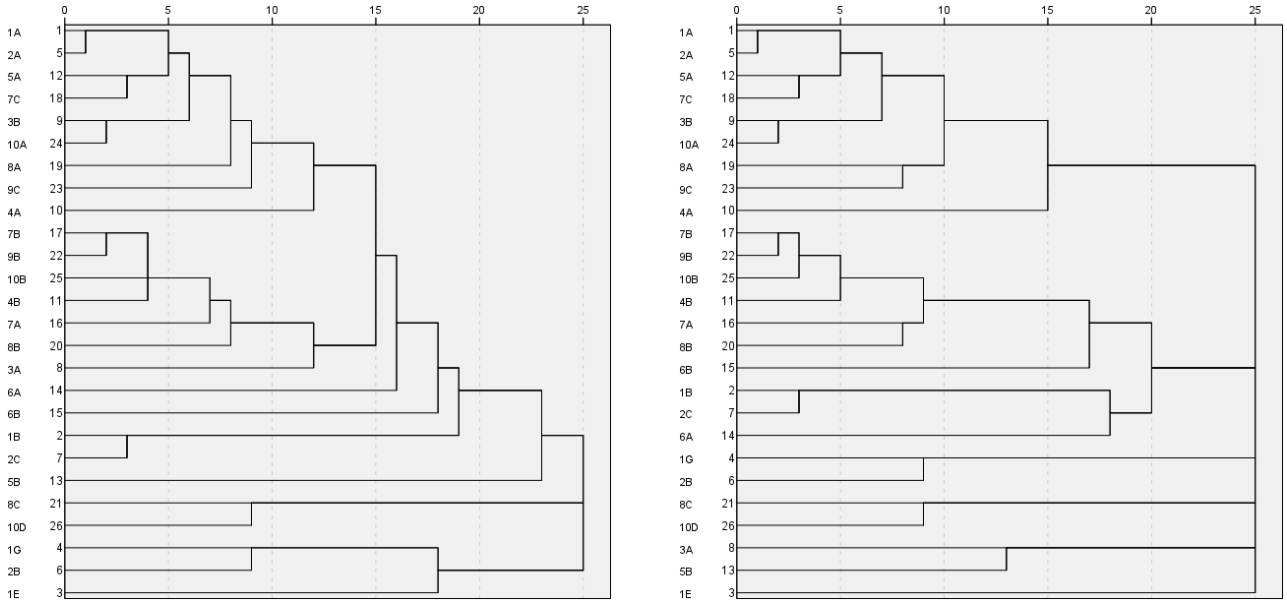

Figure S1 Left: Dendrogram of the configuration options using the UPGMA method. Right: Dendrogram of the configuration options using the complete linkage method.

S4. Selection of the clusters and internal cluster quality

Internal cluster quality indices were used to choose the most relevant clusters. These indices assess, for a given number of clusters, how well each cluster object fits a cluster (cohesiveness) and how well clusters are separated (separability). These indices also act as internal quality indicators (hence, their names), that is, internal validity indicators, because high cohesiveness and separability are desirable properties of a cluster partition (Vargha *et al.*, 2016).

These properties, cohesiveness and separability, can measured in different ways. Therefore several indices have been developed (27 such indices are listed in Desgraupes, 2017), which can provide different results (see e.g., Peeters *et al.*, 2011, pp. 2912, 2914). As recommended by Gordon (1999, pp. 63-65), several internal cluster quality indices were computed before deciding on the optimal number of clusters. The level of agreement between the selected indices was used to determine the optimal cluster partition. Only indices which did not assume that the proximity matrix values were Euclidean were considered. The chosen indices were the agglomeration coefficient (distance at which objects/clusters are grouped using the proximity measure, also called the stepsize criterion), the point-biserial correlation coefficient, the McClain-Rao index, the Gamma index, the C-index (see Milligan and Cooper, 1985) and the silhouette coefficient (see Rousseeuw, 1987). A synthetic description of these indices and their interpretation can be found in (Desgraupes, 2017) as well as (Milligan and Cooper, 1985 - the agglomeration coefficient is described under the stepsize criterion method, p. 164). The silhouette coefficient has the advantage of producing silhouette values for each configuration option, which can be used to further improve the cluster partition, if necessary (Rousseeuw, 1987; Kaufman and Rousseeuw, 1990). The indices were computed using Orlov (2022)'s macros (except for the agglomeration coefficient, computed by the SPSS Cluster function), see Section S7.

The computation of the chosen internal clustering indices yielded five possible partitions, as shown in Figure S2: partitions of four, nine, and eleven clusters using the UPGMA method (UA4, UA9, and UA11) and partitions of nine and ten clusters using the complete linkage method (Co9 and Co10). Silhouette analysis (see Figure S3) showed that Co9 and Co10 had two inhomogeneous clusters (Clusters 4 and 5, especially due to the negative value of configuration option 3A). The complete silhouette analysis (available in the SPSS files) showed that the next best cluster for 3A was Cluster 6 for Co9 and 7 for Co10. In this configuration, the partition of ten clusters becomes identical to the partition of ten clusters from the UPGMA method (which is not as good as UA4, UA9 and UA11), and the partition of nine clusters becomes identical to that of UA9. This eliminates definitely the cluster partitions obtained by the complete linkage method. UA4,

UA9, and UA11 were qualitatively compared with the heat map of the proximity matrix (containing the proximity measures computed with the Sorensen-Dice index). The heat map is shown in Figure S4. The heat map analysis led to the selection of UA9 for the remaining analysis.

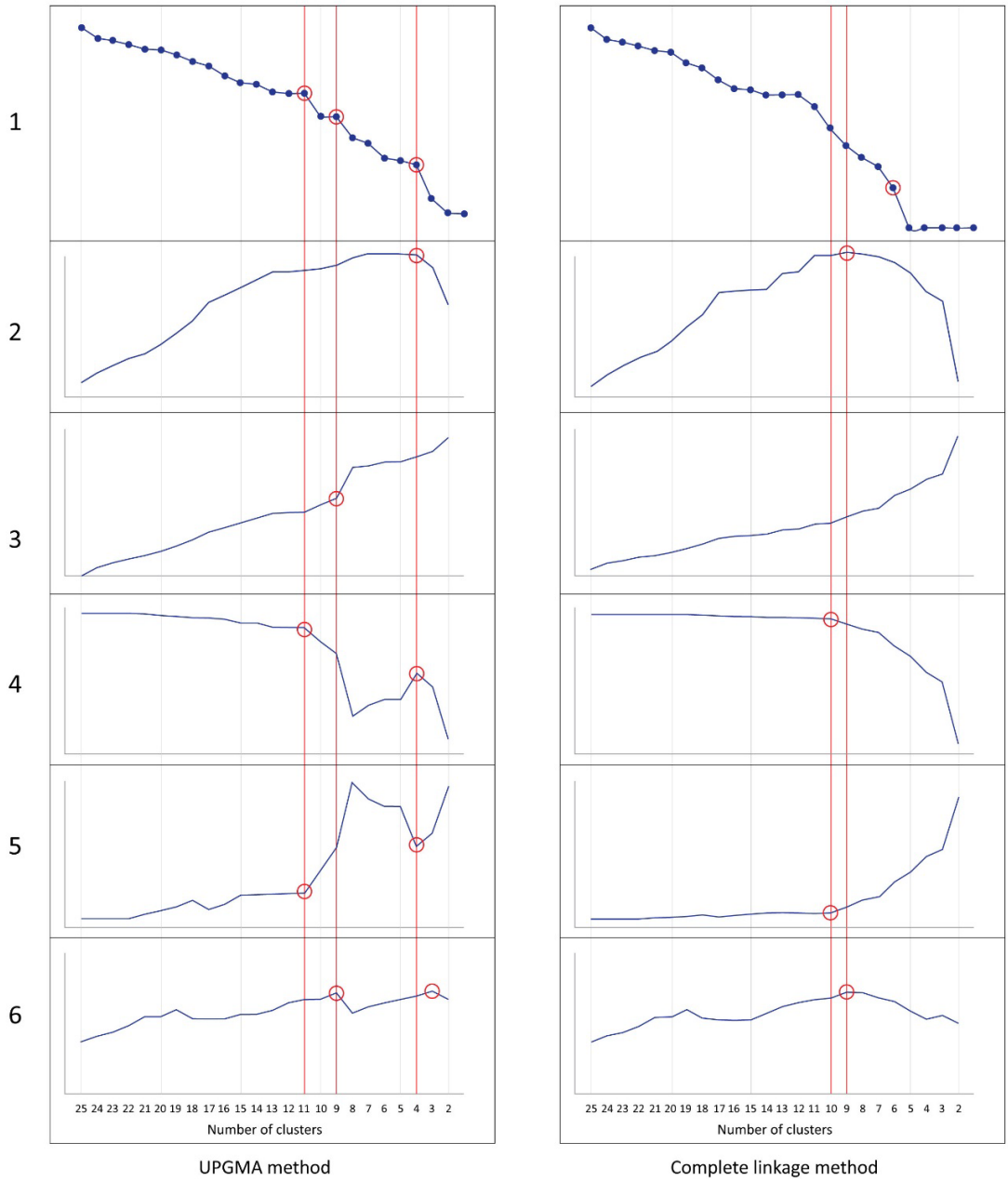

Figure S2 The indices for each partition (left: UPGMA method, right: Complete linkage method) where 1 is the agglomeration coefficient, 2 is the point-biserial correlation coefficient, 3 is the McClain-Rao index, 4 is the Gamma index, 5 is the C-index, and 6 is the silhouette value. The point-biserial correlation coefficient did not give exploitable results and is therefore not represented.

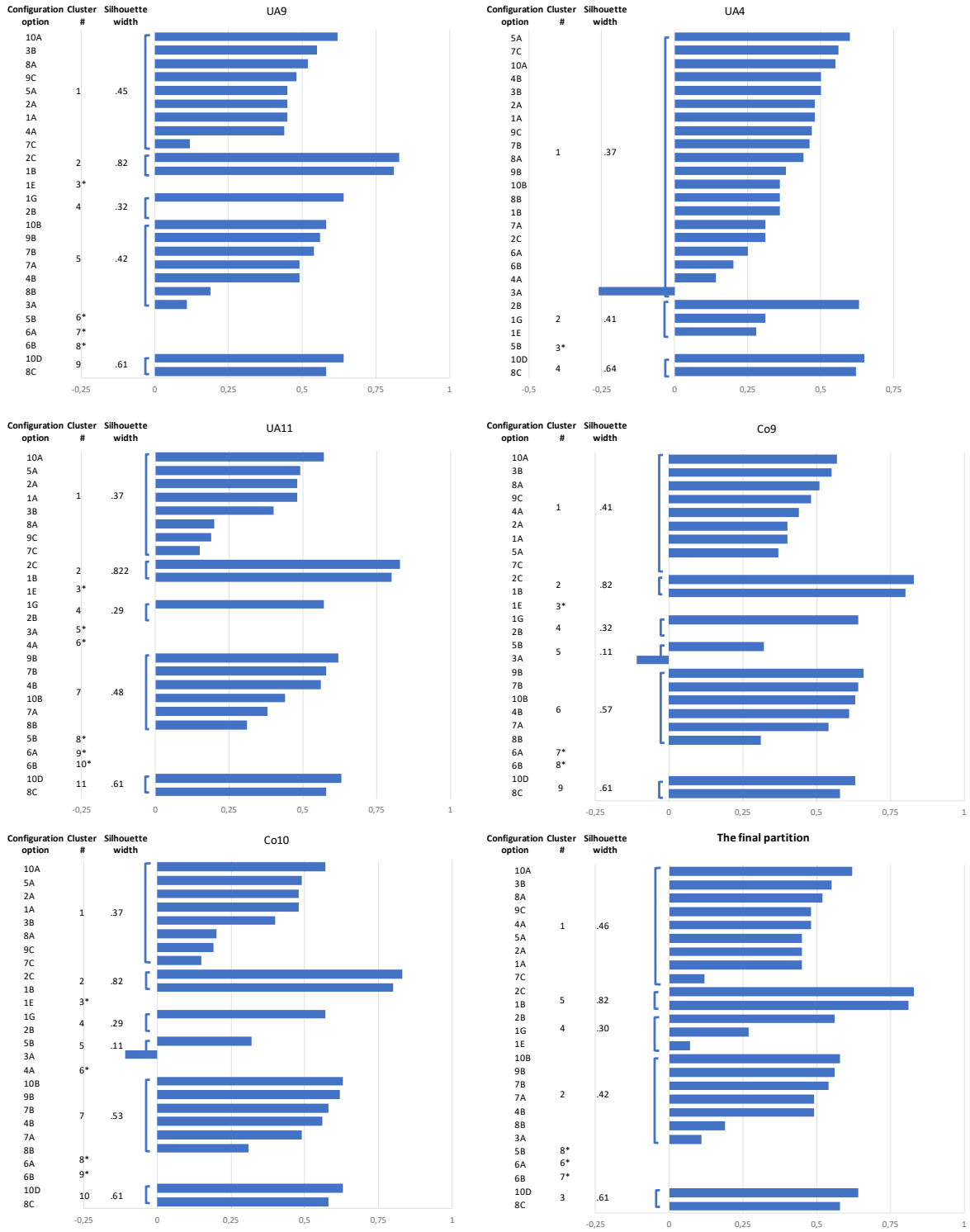

Figure S3 Silhouettes of UA4, UA9, UA11, Co9, Co10, and of the final partition. *The silhouette value of singleton clusters is indeterminate and is generally set to 0 (Kaufman and Rousseeuw, 1990, p. 85).

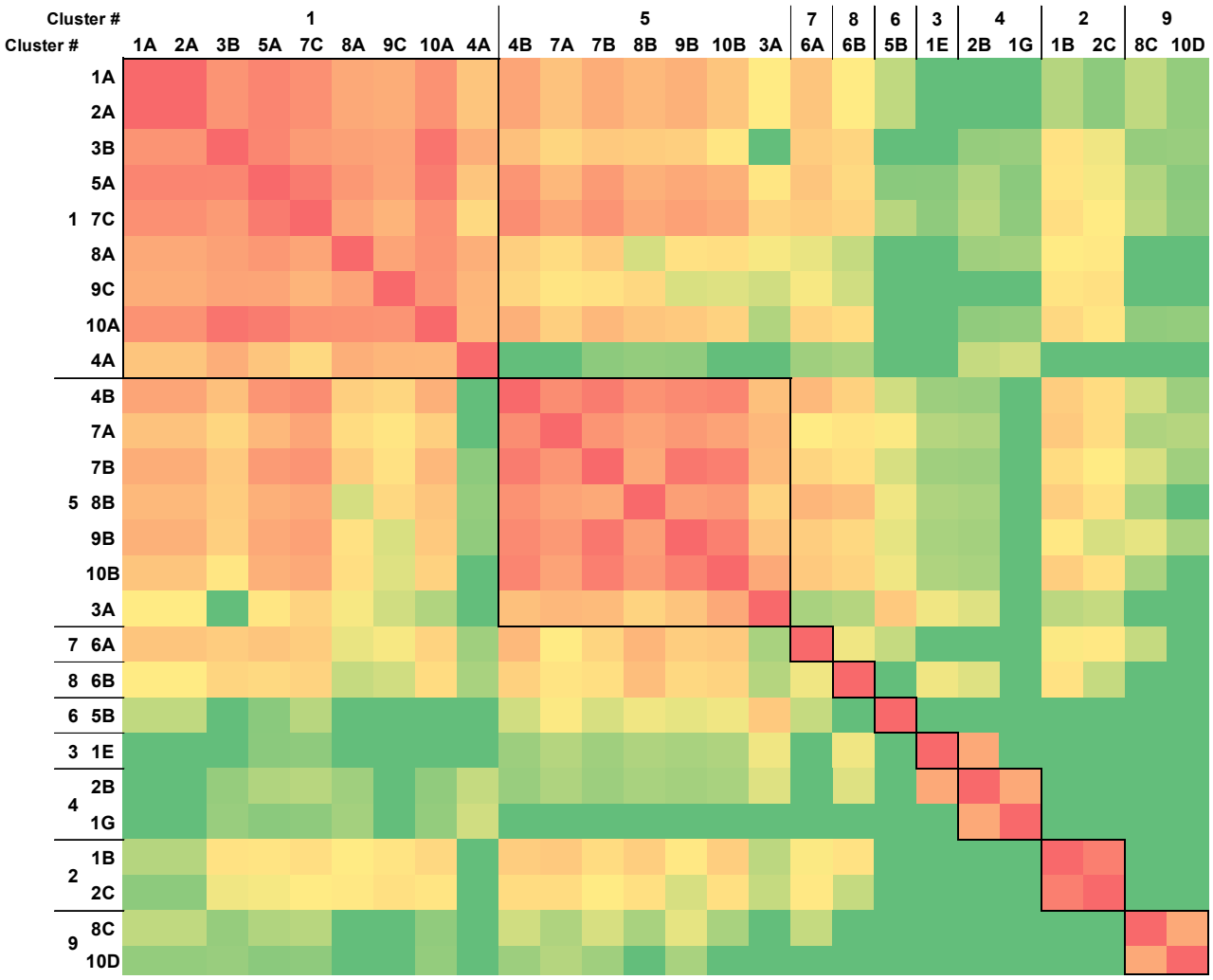

Figure S4 Heat map of the proximity measures computed with the Sorensen-Dice index. The heat map represents the partition of eight clusters obtained with UA9. In UA4, clusters 1, 5, 7 and 8 are grouped in one cluster. The heat map shows quite clearly that several configuration options do not fit together in this large cluster, especially 3A and 4A, compared to the UA9 clusters. In UA11, the two supplementary clusters were configuration options 3A and 4A (singleton clusters). The heat map shows quite clearly that 4A belongs to the first cluster (although with weaker proximity values than the other cluster members) and not to any other. Likewise, 3A fits well the second cluster. Therefore, UA9 is selected for the remaining analysis.

Finally, UA9 was investigated further. UA9 had a cluster with one poor silhouette width (Cluster 4) and four singleton clusters (1E, 5B, 6A, 6B). Further silhouette analyses and detailed scrutiny of the original led to the inclusion of configuration option 1E in Cluster 4 and to let 5B, 6A and 6B as singleton clusters. The silhouette of the final partition is presented in Figure S3. The average silhouette value of the partition is 0.42, the clusters are relatively homogenous, and the majority of the clusters are close to or above 0.5, indicating that a reasonable structure has been found (Kaufman and Rousseeuw, 1990, p. 88).

S5. Confirmation with MDS

A two-dimensional representation of the configuration options through MDS was also created. MDS works with distances, not proximity measures, therefore the Lance-Williams' non-metric measure, a distance that is the complement of the Sorensen-Dice measure in the case of binary variables (Anderberg, 1973, pp. 112-113), was used for this purpose.

The final partition of the configuration options could also be identified with the MDS mapping, see Figure S5, which provided confidence that this partition was relevant for the study. It can be noticed that, while the eight clusters can be *retrieved* from the MDS mapping, it is much more difficult to *identify* them directly from the mapping. This confirms that AHC analysis is better suited for this specific study than MDS.

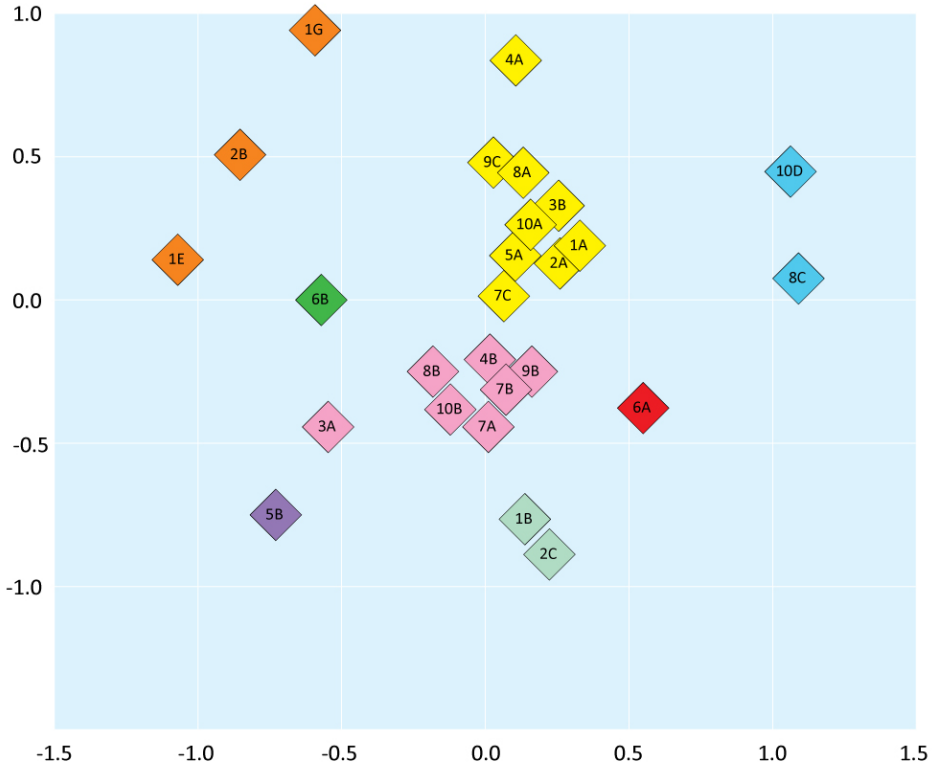

Figure S5 Identified solution patterns mapped onto the MDS chart.

S6. Confirmation with co-occurrence network analysis

The current problem can also be represented as a network or undirected graph where the vertices are the configuration options and edges are the proximity measures between the configuration options. Co-occurrence network analysis aims at clustering the graph into relevant clusters (also called modules or communities), usually based on a quality clustering index, *modularity* (Brandes *et al.*, 2008). While the AHC analysis creates clusters by aggregating the objects sequentially, the co-occurrence network analysis is optimizing modularity by considering the full graph. A current difficulty with the co-occurrence network analysis is that often a threshold needs to be determined, in order to limit the edges of the graph to the important proximity values (Borcard *et al.*, 2018, p. 118). Regarding business patterns extraction, there is no theory, rules, or principles (yet) to guide us in the establishment of a relevant threshold. Therefore, co-occurrence network analysis cannot be used so far directly for finding the best partition. However, it can be used for triangulation purpose, by incrementally varying the threshold from 0 (no edge suppressed) to 1 (all edges suppressed) and compare the extracted partitions to a known partition, here UA9. A well-known measure of comparison between partitions is the adjusted Rand index (ARI, Hubert and Arabie, 1985).

The process presented above was coded and performed with R (version 4.2.0). The partition obtained with this process that was closest to the known partition is represented Figure S6 (threshold: 0.56, modularity: 0.28). With ARI = 0.93, the partition is very close to the known one. The only difference with UA9 is that configuration option 6A is not a singleton cluster but is included in a larger cluster. The silhouette of this partition had already been studied during the AHC analysis and was inferior to that of the final partition. The partition found by the co-occurrence network analysis also confirms the relevance of including configuration option 1E in the {1G; 2B} cluster.

With very similar results, both the MDS and co-occurrence network analyses corroborates the findings from the AHC analysis.

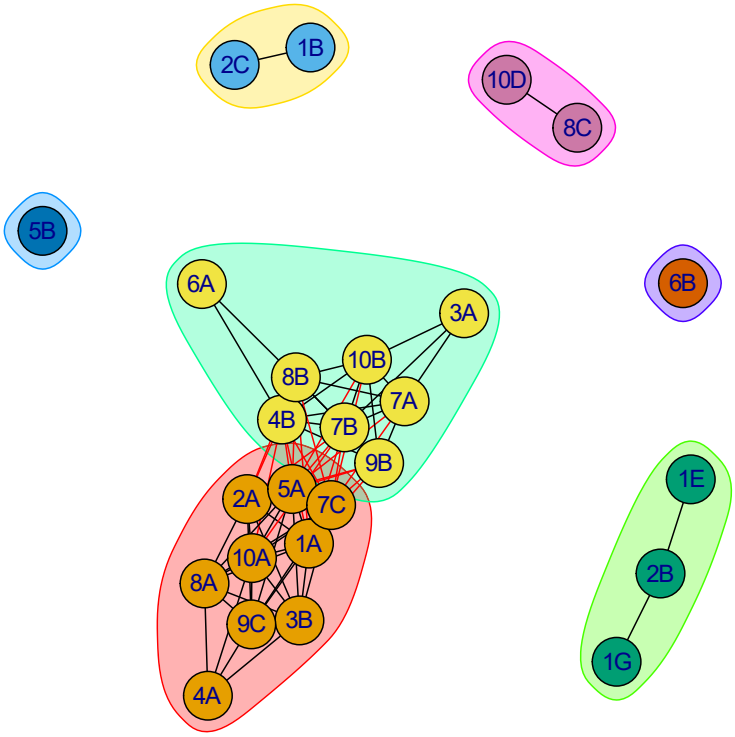

Figure S6 Partition obtained by co-occurrence network analysis.

S7. Implementation in SPSS and R

The HCA was performed in SPSS on the dataset presented in "SupplementaryMaterial_04_AHC-MDSAnalysis_input.sav" (it corresponds to the complete binary characteristic list formatted for SPSS) using the code in "SupplementaryMaterial_03_AHC-MDSAnalysis_code.sps" (Parts 1 and 2). The code computes the proximity matrix with the Sorensen-Dice index, as well as the dendrograms, cluster memberships and the agglomeration coefficient for the UPGMA and complete linkage methods.

The other indices used for the selection of the relevant were computed with Orlov (2022)'s macros (Part 3). These macros are available at https://www.spsstools.net/en/macros/KO-spssmacros/. The macro file should preferably be copied in the same folder as the supplied SPSS files.

The specific macros used for the selection of the most relevant partitions were:

- !KO_RPBCLU (Version 2, July 2018) for the point-biserial correlation coefficient and the McClain-Rao index,
- !KO_GAMMACLU (Version 2, January 2022) for the Gamma index,
- !KO_CINDEX (Version 1, September 2001) for the C-index and
- !KO_SILHOU (Version 3, July 2017) for the silhouette index.

In order to determine the most relevant partitions with Orlov (2022)'s internal clustering criteria macros, a new SPSS input file must be created containing the proximity matrix and cluster memberships from the UPGMA and complete linkage methods. This new input file is "SupplementaryMaterial_05_AHCAnalysis_ClusterSelection_input.sav". Part 3 of the instructions from "SupplementaryMaterial_03_AHC-MDSAnalysis_code.sps" can then be followed. The results are presented in an SPSS output file, except the silhouette indices for each cluster membership, which appended to the "SupplementaryMaterial_05_AHCAnalysis_ClusterSelection_input.sav" input file. These silhouette indices were subsequently analysed in Excel. The code for the computation of the silhouette indices of the final cluster is in Part 4 of the "SupplementaryMaterial_03_AHC-MDSAnalysis_code.sps" file.

The heatmap was created in Excel using the proximity matrix previously obtained.

The MDS analysis with the Multidimensional Scaling function of SPSS is also available at the end of "SupplementaryMaterial_03_AHC-MDSAnalysis_code.sps".

The input file used for the co-occurrence network analysis is "SupplementaryMaterial_01_CompleteBinaryCharacteristicList.csv". The R code can be found in the file "SupplementaryMaterial_06_CoOccurrenceNetworkAnalysis_code.R".

References

Amshoff, B., Dülme, C., Echterfeld, J. and Gausemeier, J. (2015), "Business model patterns for disruptive technologies", *International Journal of Innovation Management*, Vol. 19 No. 3, https://doi.org/10.1142/s1363919615400022.

Anderberg, M.R. (1973), *Cluster Analysis for Applications*, Academic Press, New York, NY, https://doi.org/10.1016/C2013-0-06161-0.

Borcard, D., Gillet, F. and Legendre, P. (2018), *Numerical Ecology with R*, 2nd ed., Springer International Publishing, Cham, Switzerland, https://doi.org/10.1007/978-3-319-71404-2.

Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z. and Wagner, D. (2008), "On modularity clustering", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20 No. 2, pp. 172-188, https://doi.org/10.1109/TKDE.2007.190689.

Camisón, C. and Villar-López, A. (2010), "Business models in Spanish industry: A taxonomy-based efficacy analysis", *M@n@gement*, Vol. 13 No. 4, pp. 298-317, https://doi.org/10.3917/mana.134.0298.

Choi, S.-S., Cha, S.-H. and Tappert, C.C. (2010), "A survey of binary similarity and distance measures", *Journal on Systemics, Cybernetics and Informatics*, Vol. 8 No. 1, pp. 43-48, available at: http://www.iiisci.org/journal/sci/FullText.asp?var=&id=GS315JG (accessed 16 February 2022).

Curtis, S.K. (2021), "Business model patterns in the sharing economy", *Sustainable Production and Consumption*, Vol. 27, pp. 1650-1671, https://doi.org/10.1016/j.spc.2021.04.009.

Desarbo, W.S., Grewal, R. and Scott, C.J. (2008), "A clusterwise bilinear multidimensional scaling methodology for simultaneous segmentation and positioning analyses", *Journal of Marketing Research*, Vol. 45 No. 3, pp. 280-292, https://doi.org/10.1509/jmkr.45.3.280.

Desgraupes, B. (2017), "An R package for computing clustering quality indices [Vignette]", available at: https://CRAN.R-project.org/package=clusterCrit (accessed 10 June 2022).

Dice, L.R. (1945), "Measures of the amount of ecologic association between species", *Ecology*, Vol. 26 No. 3, pp. 297-302, https://doi.org/10.2307/1932409.

Everitt, B.S., Landau, S., Leese, M. and Stahl, D. (2011), *Cluster Analysis*, 5th ed., John Wiley & Sons, Chichester, United Kingdom, https://doi.org/10.1002/9780470977811.

Gordon, A.D. (1999), *Classification*, 2nd ed., Chapman and Hall/CRC, New York, NY, https://doi.org/10.1201/9780367805302.

Holman, E.W. (1972), "The relation between hierarchical and euclidean models for psychological distances", *Psychometrika*, Vol. 37 No. 4, pp. 417-423, https://doi.org/10.1007/BF02291218.

Holzmann, P., Breitenecker, R.J. and Schwarz, E.J. (2019), "Business model patterns for 3D printer manufacturers", *Journal of Manufacturing Technology Management*, Vol. 31 No. 6, pp. 1281-1300, https://doi.org/10.1108/jmtm-09-2018-0313.

Hubert, L. and Arabie, P. (1985), "Comparing partitions", *Journal of Classification*, Vol. 2 No. 1, pp. 193-218, https://doi.org/10.1007/BF01908075.

Hunke, F., Schüritz, R. and Kuehl, N. (2017), "Towards a unified approach to identify business model patterns: A case of e-mobility services", in Za, S., Drăgoicea, M. and Cavallari, M. (Eds.), *Exploring Services Science. IESS 2017*, Springer International Publishing, Cham, Switzerland, pp. 182-196, https://doi.org/10.1007/978-3-319-56925-3_15.

Kaufman, L. and Rousseeuw, P.J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, Hoboken, NJ, https://doi.org/10.1002/9780470316801.

Legendre, P. and Legendre, L. (2012), *Numerical Ecology*, 3rd ed., Elsevier, Amsterdam, The Netherlands, available at: https://www.sciencedirect.com/bookseries/developments-in-environmental-modelling/vol/24/ (accessed 8 April 2022).

Milligan, G.W. and Cooper, M.C. (1985), "An examination of procedures for determining the number of clusters in a data set", *Psychometrika*, Vol. 50 No. 2, pp. 159-179, https://doi.org/10.1007/BF02294245.

Newman, M.E.J. (2018), *Networks*, 2nd ed., Oxford University Press, Oxford, United Kingdom, https://doi.org/10.1093/oso/9780198805090.001.0001.

Orlov, K. (2022), "Internal clustering criteria", *Kirill's SPSS Macros Page, Raynald's SPSS Tools*, available at: https://www.spsstools.net/en/macros/KO-spssmacros/ (accessed 16 March 2022).

Peeters, G., Giordano, B.L., Susini, P., Misdariis, N. and McAdams, S. (2011), "The Timbre Toolbox: Extracting audio descriptors from musical signals", *The Journal of the Acoustical Society of America*, Vol. 130 No. 5, pp. 2902-2916, https://doi.org/10.1121/1.3642604.

Pruzansky, S., Tversky, A. and Carroll, J.D. (1982), "Spatial versus tree representations of proximity data", *Psychometrika*, Vol. 47 No. 1, pp. 3-24, https://doi.org/10.1007/BF02293848.

Romesburg, H.C. (1984), *Cluster Analysis for Researchers*, Lifetime Learning, Belmont, CA.

Rousseeuw, P.J. (1987), "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", *Journal of Computational and Applied Mathematics*, Vol. 20, pp. 53-65, https://doi.org/10.1016/0377-0427(87)90125-7.

Suppes, P., Krantz, D.H., Luce, R.D. and Tversky, A. (1989), *Foundations of Measurement Volume 2: Geometrical, Threshold, and Probabilistic Representations*, Academic Press, San Diego, CA, https://doi.org/10.1016/C2009-0-21665-5.

Sørensen, T. (1948), "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons", *Biologiske Skrifter*, Vol. 5 No. 4, available at: https://www.royalacademy.dk/Publications/High/295_S%C3%B8rensen,%20Thorvald.pdf (accessed 27 April 2022).

Vargha, A., Bergman, L. and Takács, S. (2016), "Performing cluster analysis within a person-oriented context: Some methods for evaluating the quality of cluster solutions", *Journal for Person-Oriented Research*, Vol. 2 No. 1-2, pp. 78-86, https://doi.org/10.17505/jpor.2016.08.