Assessing Qualitative Similarities Between Financial Reporting Frameworks Using Visualization and Rules: COREP vs. Pillar 3

Wenmei Yang[1] *and Adriano S. Koshiyama*[2*]

[1]*Large commercial bank supervision department,*
China Banking Supervision Commission
No. Jia 15, Financial Street, Beijing, China, 100800
e-mail: yangwenmei@cbrc.gov.cn
Tel: +86 10 66279524

[2]*Department of Computer Science, University College London*
Gower Street, London WC1E 6BT
United Kingdom - Tel: +44 (0) 20 7679 2000
e-mail: a.koshiyama@cs.ucl.ac.uk

Summary

Financial institutions are struggling with larger volume, more specific and greater frequency of regulatory reporting after the global financial crisis in 2008, especially those that need to report to multiple jurisdictions. To help to improve reporting efficiency, this paper aims to assess the existence of similarities between templates related to credit and counter party credit risk of COREP and Pillar 3 regulatory reporting frameworks by applying Correspondence Analysis and Association Rules Mining. Our results suggest a high degree of overlap between these reporting frameworks, more prominently the three business functions as Front office, Finance and Risk. These patterns can be used as guidance for financial institutions to reshape their reporting architecture.

Keywords: Regulatory Reporting Framework; Reporting Architecture; Correspondence Analysis; Association Rules Mining

1. Introduction

After the 2008 Global Financial Crisis, financial regulators, international institutions and governments around the world proposed stricter reporting to more effectively monitor and mitigate risks. New regulations, such as Basel III and Dodd Frank Act and Markets in Financial Instruments Directive 2 (MiFID2) have, still and will drive the major changes to regulatory reporting (Degryse, 2009; Walker, 2011; Hortin, 2016; Covi, 2016). Relevant amount of research has been devoted on evaluating these regulatory frameworks (Maximilian, 2012; Viral and Ryan, 2016; Ryan, 2017), checking if they enhance financial stability and the economic impacts of this disclosure. The overall picture is that the volume, complexity and pace of regulatory reporting for financial institutions are growing significantly.

For example, financial institutions under the supervision of U.S. Federal Reserve must submit multiple files including call reports, stress testing reports, in addition to increasing requirements on data granularity and submission frequency (FSOC, 2014). This trend is indicated by the document of Basel Committee on Banking Supervision (BCBS) named Principles for Effective Risk Data Aggregation and Risk Reporting (BCBS 239) (EBA, 2016; BCBS, 2016). Hence, financial institutions must fill and submit reports more frequently and with greater level of detail. When the financial institution is operating globally, the volume and complexity of reporting requirements increase dramatically (Covi, 2016; Leuz and Wysocki, 2016).

In this sense, financial institutions are forced to improve their data quality and integration across business functions and product lines, given the short implementation time frames as well as the uncertainty in rule making (BCBS, 2013; Ernst and Young, 2012; BCBS, 2015; Viral and Ryan, 2016). Such challenges tend to increase especially if the companies deal with each regulation separately (across different departments, business lines and geographies, i.e. a fragmented response), instead of addressing common challenges across different reporting frameworks together (i.e. a harmonized response). It is easy to perceive that there are similar requirements from various regulatory frameworks, but very limited research has been devoted to building tools to quantify and check eventual overlaps. With the knowledge of these connections or affinities among the regulations, financial institutions would be able to optimize their business processes, technology platform and data infrastructure.

In order to measure associations between a set of variables, the financial literature is populated of research that applies the Principal Component Analysis (Scheinkman and Litterman, 1991; Fontana and Scheicher, 2016; Poynter et al., 2015), mostly because the researchers are dealing with quantitative/continuous types of variables (such as returns, prices and yields). However, in the case of regulatory frameworks, most of the data that can be extracted are textual and categorical, being more conventionally displayed as contingency tables rather than in a time series. In this case, similar approaches such as Correspondence Analysis (Beh and Lombardo, 2014) and Association Rules Mining (Agrawal et al., 1993) are preferred techniques, being both popularly used in other areas, such as psychology, biology, marketing, etc. (Greenacre and Pardo, 2006; Költringer and Dickinger, 2015; Higuera-Mendieta et al., 2016).

Therefore, this work quantifies the similarity between elements of two different regulatory reporting frameworks: the most up to date Common Reporting (COREP) issued by the European Bank Authority and Pillar 3 (EBA, 2017; BCBS, 2017). With this information at hand, the financial institutions can improve regulatory reporting efficiency and timeliness as well as reducing compliance costs. By applying the Correspondence Analysis and Association Rules Mining techniques, we aim to assess these relationships from different perspectives and levels of aggregation as well as intend to show that our findings are robust across methodologies. Our work also related to some previous contributions in the realm of data visualization and parsing in finance and business (Goel and Gangolly, 2012; Kleinknecht and Ng, 2015; Jaffe, 2015; Fisher et al., 2016).

Hence, in terms of contributions, the visualized patterns displayed in this paper can be used as a guidance for financial institutions to reshape their reporting architecture. For example, the templates which are more closed regarding their data source were clustered by our methods, i.e. three business functions as Front office, Finance and Risk, were assembled. This result can be used to clear reporting responsibilities and locate data in different IT systems. Secondly, the correspondence among templates and a particular data item or group of data items were uncovered. So, given a data item which is interested in, the templates that are positively or negatively correlated with particular data item could be figured out at a glance. All this can be used to streamline reporting, optimizing the reporting architecture, and assess points of failure along the process.

In this sense, this work is organized as follows: next section presents a background review, outlining some fundamentals on the COREP and Pillar 3 frameworks and describing the two techniques used in this work: Correspondence Analysis and Association Rules Mining. The third section exhibits the modelling strategy, showing how the data was fetched from the regulatory reporting frameworks, how they were transformed and manipulated, and the procedure pursued to apply both techniques appropriately. Then, we move to the results and analysis section, starting from the Correspondence Analysis and moving to Association Rules Mining. In both cases, we have begun by reporting the results from a high-level of aggregation and then moving to analyzes at the template level. After suggesting the main overlaps between COREP and Pillar 3, section 5 closes this work with some final remarks.

2. Background Review

2.1. Common Reporting (COREP)

The European Banking Authority (EBA) published a standardized reporting framework to address the Capital Requirements Regulation and Directive (CRR/CRD) reporting. It applies to all credit institutions and investment firms operating in European Economic Area, and almost 30 European countries have adopted this reporting framework. Since the first publication in 2006, the EBA has updated COREP several times to the newest version of DPM 2.7 (Data Point Model, see Table 1) in 2017, which contains 111 templates and covers capital adequacy and group solvency, credit and counter party credit risk, market risk, liquidity risk and operational risk, etc. The DPM is a structured representation of the data, identifying all the business concepts and its relations, as well as validation rules (EBA, 2012). To increase transparency in regulatory reporting, COREP has a relatively high level of data requirements in terms of volume and granularity (more than 12,000 data points in all). In this paper, we refer to data point as the data to be filled in templates.

On a case by case basis, the COREP templates will have to be completed and delivered monthly, quarterly or annually to the EBA in the XBRL format (Extendable Business Reporting Language), and it is expected that most national regulatory authorities will pass that requirement to financial institutions. Although COREP reports are highly standardized regarding content and delivery format, it can pose significant challenges for many financial institutions given the time available to complete the construction of the internal frameworks required to produce these reports.

2.2. Pillar 3 of Basel Accords

The Basel Accords refer to the sets of guidelines and recommendations for banking regulations issued by the BCBS. Although the Committee does not have the authority to enforce recommendations, a significant number of national regulators tend to implement the Committee's policies through national laws and regulations. The third instalment of the Basel Accords (Basel III) was developed as a response to the deficiencies of financial regulations indicated by the financial crisis in 2008. Basel III is based upon three essential aspects, which are called the "3 pillars". These three pillars are Capital adequacy requirements, Supervisory review and Market discipline. The Pillar 3, i.e. market discipline, refers to requirements for the public disclosure of regulatory information by financial institutions and aims to improve the transparency of financial markets and to promote comparability of banks' risk profiles within and across countries.

The first set of Pillar 3 disclosure requirements were issued in 2004 and were amended in 2009. The BCBS issued a revised standard of Pillar 3 in January 2015, followed by a consultative document in 2016 and the latest standard document in 2017 which sets out new proposals of disclosure requirement (see Table 2). According to the specifications document, the disclosure scope includes details for capturing the financial institution's risk profile, risk management, capital adequacy, liquidity, remuneration practices, among others.

The Pillar 3 report must be published concurrently with financial reports for the corresponding period (BCBS, 2017). There are 63 templates in all, and 16 of them require qualitative information disclosure. The format of some templates is flexible and can be adjusted by jurisdiction or financial institutions. Regarding quantitative data disclosure, a financial institution needs to disclosure at least 4000 data points under the requirements of the 2017 standard document. Compared with COREP, the data granularity of Pillar 3 is relatively low, and this is in accordance with their oriented objects. The data in COREP is only provided to the regulator, but the information in Pillar 3 report is publicly disclosed.

This paper does not cover all the templates in COREP and Pillar 3. It will focus on templates related to credit and counter party credit risk (CR and CCR). The quantitative information report templates are also not included in the scope of this paper. In terms of assessing CR, there are 12 templates in COREP and 25 templates in Pillar 3 related, and the total data points are 4007 and 1833, respectively. For some templates which report detailed information, such as C14.00 in COREP or CCR4 in Pillar 3, it is impossible to calculate the exact number of total data points, so we use some columns or sub-total data points instead.

2.3. Correspondence Analysis (CA)

CA is a technique of data analysis for approximating a high-dimensional tabular data into a low-dimension representation that can be displayed graphically (Greenacre, 2007). The basic idea of CA is like principal component analysis (PCA), with the main difference is that PCA applies to continuous data, but CA applies to categorical data. CA allows the user to display or summarizes information in a low-dimensional graphical form, usually two-dimensional, to inspect their correspondences or associations at a category level.

Suppose we have a sample of items, and we can use two categorical variables X and Y that contains I and J categories respectively, to describe the characteristics of each item.

$$X = \{X_1, X_2, \dots, X_I\}, \qquad Y = \{ Y_1, Y_2, \dots, Y_J\}$$

Now a $I \times J$ data matrix N (a.k.a contingency table), can provide a summary of notation. The element n_{ij} of N denotes the joint frequency of items which are classified into i-th row category and j-th column category at the same time.The i rows consisting of j elements can be thought as i points in a j-dimensional space. Similarly, the j columns consisting of i elements can be thought as j points in a i-dimensional space.

It is not easy to represent those points graphically for I and J larger than 3. Then, to obtain a reasonable view of the relationship between variable X and Y, we need to find a low-dimensional approximation on the premise of not losing too much information. Although the loss of information is inevitable, since the information contained in high dimensions cannot be simply ignored, our main objective is to minimize it. The fundamental mathematical technique used in CA to reduce dimensionality is the singular value decomposition (SVD), which is briefly outlined step by step in Appendix I.

In CA, to quantify the overall association between the categorical variables X and Y, a measure named χ^2 distance is introduced. The $\chi^2$is calculated as:

$$\chi^2 = \sum \frac{(observed - expected\,)^2}{expected} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \frac{(n_{i.}n_{.j})}{n})^2}{(\frac{n_{i.}n_{.j}}{n})} \qquad (1)$$

where $n_{i.}, n_{.j}$ represents the number of items in the i-th row and j-th column respectively, and n is the grand total. The expected count is produced under the assumption that X and Y are independent of each other, that is, the frequencies are produced by chance. The larger the value of χ^2, the less valid is the assumption of independence. A hypothesis test called Pearson χ^2-test can be used to assess the statistical significance (i.e., to compute a p-value) on the dependency of X and Y.

By applying the SVD decomposition over the contingency table, we get a matrix *S*. Some useful metrics can be obtained from it, such as the total inertia, inertia of a row or column. The total inertia of a data matrix is the amount of information contained in the table. It can be calculated as the sum of squares of the matrix S or the sum of squares of the singular values or eigenvalues.

$$total\ inertia = trace(SS^T) = \sum_{k=1}^{K} \alpha_k^2 = \sum_{k=1}^{K} \lambda_k^2 \quad (2)$$

The square root of total inertia can be interpreted as the correlation coefficient. As a rule of thumb, any value of the *trace* > 0.2 indicates a significant correlation between row variable and column variable (Bendixen, 2003). The inertia of a row (or column) shows the amount of information it contains. For a given row, its inertia is calculated as the row mass multiplied by the squared distance between the row and the average row profile.

2. 4. Association Rules Mining (ARM)

The methodology for mining frequent items and association rules was originated and gradually developed in processing data in large databases. Piatetsky-Shapiro (1991) firstly proposed the concept of “strong rules” and analyzed strong rules that were uncovered through some measures of interestingness. Agrawal et al. (1993) deepened this concept and expanded this problem to mining association rules from transaction data.

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of *n* binary attributes called *items* and let $D = \{t_1, t_2, \ldots, t_n\}$ be a set of transactions called the *database*. Each transaction in D has a unique transaction ID and contains a subset of the items in I. An *itemset* is a subset of *I*, and a k-item set contains k items. A *rule* is defined as an implication of the form: $X \Rightarrow Y$, where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. The itemset *X* and *Y* are called *antecedent* (left hand side or LHS) and *consequent* (right hand side or RHS) of the rule, and *X* and *Y* are mutually exclusive, i.e. they do not contain same items.

The $support$ for an item set or association rule $X \Rightarrow Y$ measures how frequently it occurs in the database *D* (Lantz, 2015). It can be calculated as the proportion of transactions in *D* that contains both *X* and *Y*.

$$Support(X \Rightarrow Y) = \frac{\#\, transactions\ containing\ X\ \cap Y}{\#\, transactions} \tag{3}$$

The accuracy or "predictive power" of an association rule can be measured by $confidence$, which is determined by the support of the item set containing both *X* and *Y* divided by the support of item set containing only *X* (Lantz, 2015).

$$Confidence(X \Rightarrow Y) = \frac{support(X, Y)}{support(X)} = \frac{\#\, transactions\ containing\ X\ \cap Y}{\#\, transactions\ containing\ X} \tag{4}$$

It can be noted that confidence can tell us the probability that the presences of *X* will result in the presence of *Y*. The rules that have either high support or high confidence are of the interested of analysts. Moreover, those rules that meet or surpass a certain threshold of support and confidence are so-called *strong rules.*

Another important measure is $lift$, as defined by:

$$Lift(X \Rightarrow Y) = \frac{support\ (\ X \cup Y\ )}{support(X)support\ (Y)} \tag{5}$$

Lift is also a measure of performance at predicting case, and it can be interpreted as the deviation of the support of $X \Rightarrow Y$ from the support expected under independence given the supports of *X* and *Y* (Hahsler and Chelluboina, 2011). An association rule $X \Rightarrow Y$ is strong if lift value is large (i.e. $lift \gg 1$).

The principal problem for association rules mining algorithm is that as the number of items grows, the number of potential rules will increase exponentially. If there are k unique elements in the database, there are 2^kpossible association rules, which make the searching work to be hopeless. A smarter rule learning algorithm called Apriori Algorithm limits the research scope for rules to a manageable size. The generating process can be summarized into two steps (Larose and Larose, 2014):

- Step 1: Dig out all frequent itemsets, i.e. find all itemsets with frequency$\geq \varphi$. The frequency of an itemset can be calculated as the number of transactions that contain this itemset.
- Step 2: Generate association rules from the frequent itemsets, by the minimum support σ and minimum confidence δ constraints.

To reduce the search scope of association rules, the Apriori algorithm makes use of a priori belief called *Apriori property*. The property states that all subsets of a frequent itemset must also be frequent, or put in another way, if an itemset is not common, then adding another item to it will not make it more frequent (Lantz, 2015).

3. Experimental Framework and Modeling Strategy

3.1. Correspondence Analysis (CA)

In this section, we explain how we are using CA to conduct similarity and affinity analysis for COREP and Pillar 3 templates (we show two typical cases of templates belonging to these frameworks in Appendix II). From the previous section, we know that categorical data is needed. But the information available for us are only the template formats and legislative packages such as Capital Requirements Directive IV (CRD IV), which is issued by EBA and comprised by the Regulation (EU) No. 575/2013 and the Directive 2013/36/UE. Fortunately, the blank templates contain a wealth of information that can be transformed into data. The main idea is to categorize all the data points of templates into different kinds, then the number of data points of each kind can be calculated. Finally, a contingency table can be obtained, in which the row and column variables are categories and templates, and the values are numbers of data points.

Being more specific, to extract and transform the information contained in the regulatory templates into proper data that satisfy the format requirements of CA, we made a three-step processing to the templates. Firstly, recognize the most likely source of each data point in our templates. According to the work distribution and business process in a financial institution, the required regulatory data can be generated by different departments. By analyzing the data requirements of COREP and Pillar 3, we concluded that all the data can be acquired from three kinds of departments: Front Office (FR), Risk Department(R) and Finance Department (FI). For example, customer number, credit risk exposure and net income can be obtained from front office, risk department and finance department, respectively.

Then, we calculated the number of data points of each source for every template. Finally, the numbers obtained were structured in the form of a contingency table (Table 3), in which the templates were regarded as column variables. The 13 columns named C7, C8, etc., are templates or groups of templates from COREP and Pillar 3; the values are the number of the data points sourced from FR, R and FI.

Also, to make a more thorough analysis, a detailed classification was developed. Given the similarity of definition, the data points have been summarized into several items, which are defined individually as a “theme”. Each theme includes one kind of data, which has the possibility of being calculated by the same department, process or from the same type of business transaction. After eliminating the repeated themes, the data points in our templates were categorized into 12 unique themes (Table 4). For example, the data theme with short name “EXP2” includes all the data points reporting credit and counter party credit risk exposures. The data theme ‘VAP1’ includes all the data points related to value adjustment and provisions, such as guarantees, credit derivatives, etc. The numbers at the end of each short name represent its data source, as 1 pose for FI, 2 for FR and 3 for R.

By calculating the number of data points which belong to data themes, we can construct another contingency table with 12 categories (Table 5). This table contains the same total number of data points as Table 4, yet they are classified more detailed as different data themes for nuanced research. In Table 5, we have 12 row variables which refer to data themes and 13 column variables which refer to templates or group of templates. Therefore, we focus our analyzes into this 12×13 matrix data.

It's worth to mention that the outliers can cause serious consequences to CA. Therefore, some special handling was made during the data transformation process to reduce the number of variables and control outliers. Some of the templates were grouped given similarity, and their data points were summed up. For instance, the variable "C9" represent a group of three templates including C 09.01, C 09.02 and C 09.03, while the variable "SEC" represent a group of four templates including SEC1, SEC2, SEC3 and SEC4 (Table 6). Also, certain templates, like C 14.00, require transaction-level reporting and financial institutions should list all the transactions one by one. This raises a problem with our data point counting since the number of rows in such templates is unknown. Another example is CR6 of Pillar 3, which require disclosing risk exposure and breakdown by portfolio. In both occasions, the number of data points was multiplied by the number of portfolios under internal rating-based approach.

Then the two contingency tables are analysed by CA. As stated before, our goal is to uncover the relationship between different groups of templates as well as if a correspondence exists between templates and departments in financial institution, or templates and data themes. The strength of association was evaluated through correlation coefficient, and a Chi-squared test, with the correlated patterns being visualized by symmetric and asymmetric bi-plot.

3.2. Association Rules Mining (ARM)

To conduct association rules mining, the information contained in templates needed to be transformed into binary attribute data. Firstly, the headers and first column of templates were categorized into “measures” and “dimensions”. For example, the measure “risk weighted exposure” and dimension “risk class - credit risk” define a specific data element “risk weighted exposure of credit risk” in the template. For the 37 templates concerned a list of 140 data elements, i.e. combinations of measures and dimensions, were generated after excluding the repeated ones with each other.

Then, based on the data elements, we generated binary attribute data for each template and obtained a binary incidence matrix with 37 rows (number of templates covered) and 140 columns (number of data elements). The matrix entries represent the presence 1 or absence 0 of data element in particular template (omitted here the data matrix). Although we lose some information on this process, we believe that this loss is offset by CA as well as the analytical capacity that can be obtained from ARM.

4. Results and Analysis

4.1. Correspondence Analysis

4.1.1. Analysis of Workload in Regulatory Reporting

We first apply CA to the contingency data in Table 3. It is a special case of the global picture because it only has three dimensions in categories. Then, by using only two CA components, it is possible to explain 100% of the data variability – only in cases where the features have no relationship at all, we would be able to explain around 66-70%. In our case the dimensions have a clear relationship: Front Office (FR), Risk Department (R) and Finance department (FI). For example, customer number, credit risk exposure and net income can be obtained from front office, risk department and finance department, respectively. This result suggest that one of the dimensions can be replicated by linearly combining the other two: we may be able to combine information from the FR and FI to replicate the same numbers of R.

The summary of the CA result is listed in Table 7. The table 7 (top) contains the variances and the percentage of variances retained by each dimension; Table 7 (middle) contains the coordinates, the contribution and the cos2 (quality of representation in [0, 1]) of the first 10 active column variables on the dimensions 1 and 2; and Table 7 (bottom) contains the coordinates, the contribution and the cos2 of the first 3 active row variables on the dimensions 1 and 2.

The trace, i.e. inertia, of this contingency (the sum of the eigenvalues) is 0.32, and the correlation coefficient is 0.565, which indicating a strong dependency between row and column variables. Besides, the chi-square = 2537.724, p = 0, which also means a strong link. Note that, the chi-square statistics = trace * n, where n is the grand total of the table (total frequency).

Then since the correspondence analysis has calculated the distance values across the rows and columns of the contingency table, it is understandable to plot them. There are two dimensions in a biplot of CA, and each dimension explains a certain percentage of the data variation or inertia (Figure 1). Theoretically, it is possible to plot any two dimensions in one plot, but in practice, most of the reported CA displays the first two principles because a combination of the first two dimensions captures the largest percentage of the variation and offers the most accurate and interpretable visualization.

The left picture of Figure 1 outlines a symmetric biplot in which both rows (red triangles) and columns (blue points) are plotted in the same space using the principal coordinates. In this plot, the distance between any row points or column points, which are approximate Chi-squared distance, gives a measure of their similarity (or dissimilarity). Column points with the similar profile are closed, C9 and C11 are the most different ones (looking from the perspective of the first component), whilst there are many similar profiles with short distance such as C12 and C13, SEC and C14, etc. (the ones clustered together). Being closer in this case mean that some of the templates are often used in a certain department (e.g., FI).

It is worth to notice that in the symmetric plot, the inter-distance between rows and columns cannot be interpreted. For example, we cannot conclude that C9 and FI have strong association directly from their closeness in the symmetric plot. This is because the symmetric plot is just an overlay of two separate maps and the distance between row points and column points is not defined or intended in this map. In this sense, to illustrate the associations between templates (columns) and departments (rows), the so-called asymmetric plot can be used.

The right picture of Figure 1 exhibits an asymmetric bi-plot in which the columns data points are in principal coordinates and the rows points are in standard coordinates. To interpret the distance between rows and columns, an analyst should connect a column point with the origin by an arrow: if the angle between row arrow and column row is acute, then there is a strong association between corresponding row and column. In this way, some pairs of row and column with strong association can be figured out in the asymmetric plot, such as FI and C9, R and C8, FR and CR1, and so on. Take FI and C9 as example, the association can be interpreted as that compare with other departments, there are more data points in C9 were generated from finance department (FI).

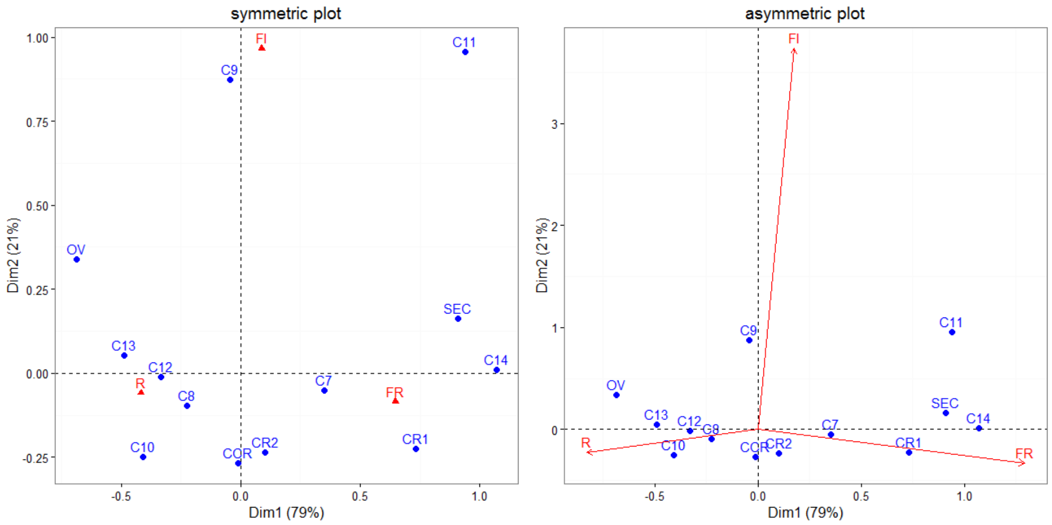


Figure 1.Scatter plot of CA result to 3-dimensional data in table 3 (In the symmetric plot, the distance between any points gives a measure of their similarity. In the asymmetric plot, the correlation between templates and departments can be found.)

In this section we focused on an aggregated perspective of our results, aiming to give an intro over the type of analysis and discussions that will be performed thoroughly in the Full Templates results section.

4.1.2. Full Templates Results

In this section, the 12×13 contingency data in Table 5 is analyzing. The research emphasis shifts to correspondence between templates and data themes, i.e. the distribution pattern of data items across templates. To start with, the strength of association between rows and columns was evaluated through the Chi-squared test – since we are dealing with a contingency table, this is the natural statistical test to evaluate potential association between rows and columns. In this case, the Chi-squared statistic equals to 10003, and the p-value is 0, which reveal the existence of a significant dependence; in other words the templates are not randomly distributed across data themes, but they have a structure that can be exploited to reduce some of the replications across data themes,

As CA is a method for dimensionality reduction, a target dimensionality need to be selected. This is not a problem when we only have a small number of rows or columns, like in our previous section, but it needs to be solved in this section. There is no clear-cut rule guiding the analyst's choice in CA (Lorenzo-Seva, 2011) and different approaches have been proposed, each one having its pros and cons. We adopted three methods, which are the scree plot, a cut-off threshold for total inertia retained and the average rule.

Firstly, a scree plot (Figure 2, left part) helps to identify how many of the components are needed, in which dimensions are plotted in order of the decreasing amount of explained inertia. The optimal dimensionality can be identified at which the scree plot shows a bend, which should be 4 or 5. The second approach is setting a cut-off threshold at an arbitrary level and keeping as many dimensions as necessary to account for the majority of the total inertia. Here we set the threshold to 80% (Figure 2, right part). Thus, the first four dimensions would be enough to achieve the satisfying level of total inertia explained. Furthermore, according to the average rule, all the dimensions that can explain more than the average inertia (expressed regarding percentages) should be considered as important and kept. Our data contains 12 rows and 13 columns, if the data were random, the expected value of inertia for each axis would be 1/12=8.33% in terms of columns and 1/11=9.1% in terms of rows. In left part of Figure 2, the level of 9.1% is plotted as a reference line, and the first five dimensions have eigenvalue above this threshold. In conclusion, our results suggest that a 4-dimensional solution seems appropriate.

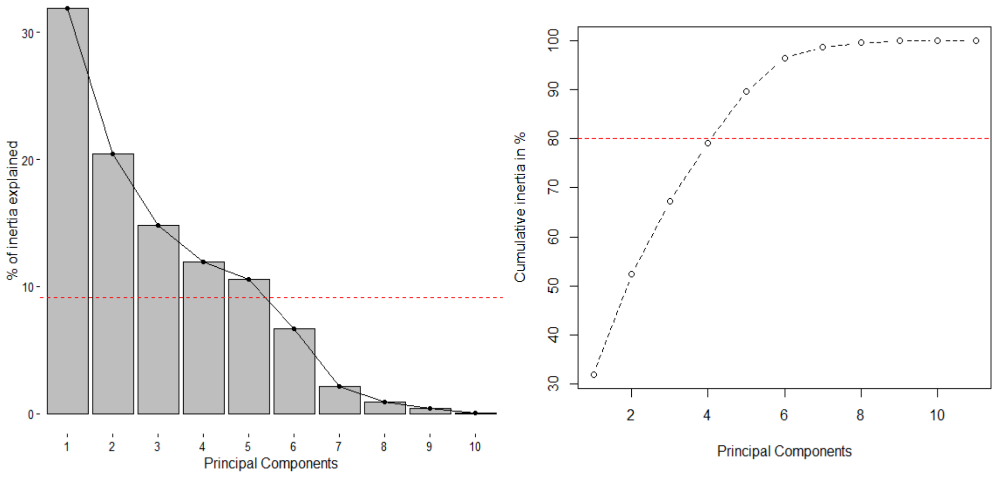


Figure 2.Inertia and cumulative inertia of principal components (The plots imply that the target dimensionality could be 4)

To understand the similarity of the templates by the proportion of the high-level data items present in each template, we need to plot in the relative position of column points in the space defined by the rows. Since we have determined the number of dimensions retained for interpretation, the next step is to assess which rows, i.e. themes, are determining those dimensions. This can be accomplished by inspecting the bar plot (Figure 3) which displays the contributions of themes to the definition of the first four dimensions. Moreover, the average contribution served as a threshold is indicated by a reference line, so the contributions above the average level can be considered as important for the definition of that dimension (Greenacre, 2007).

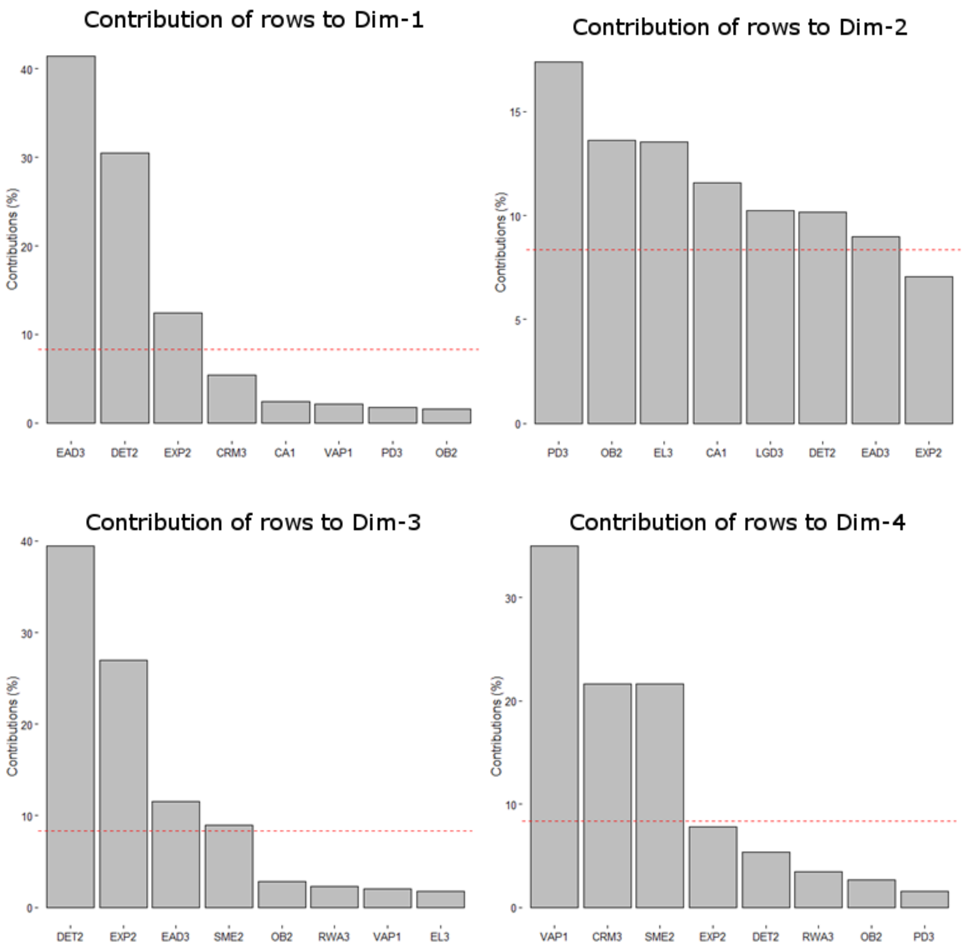
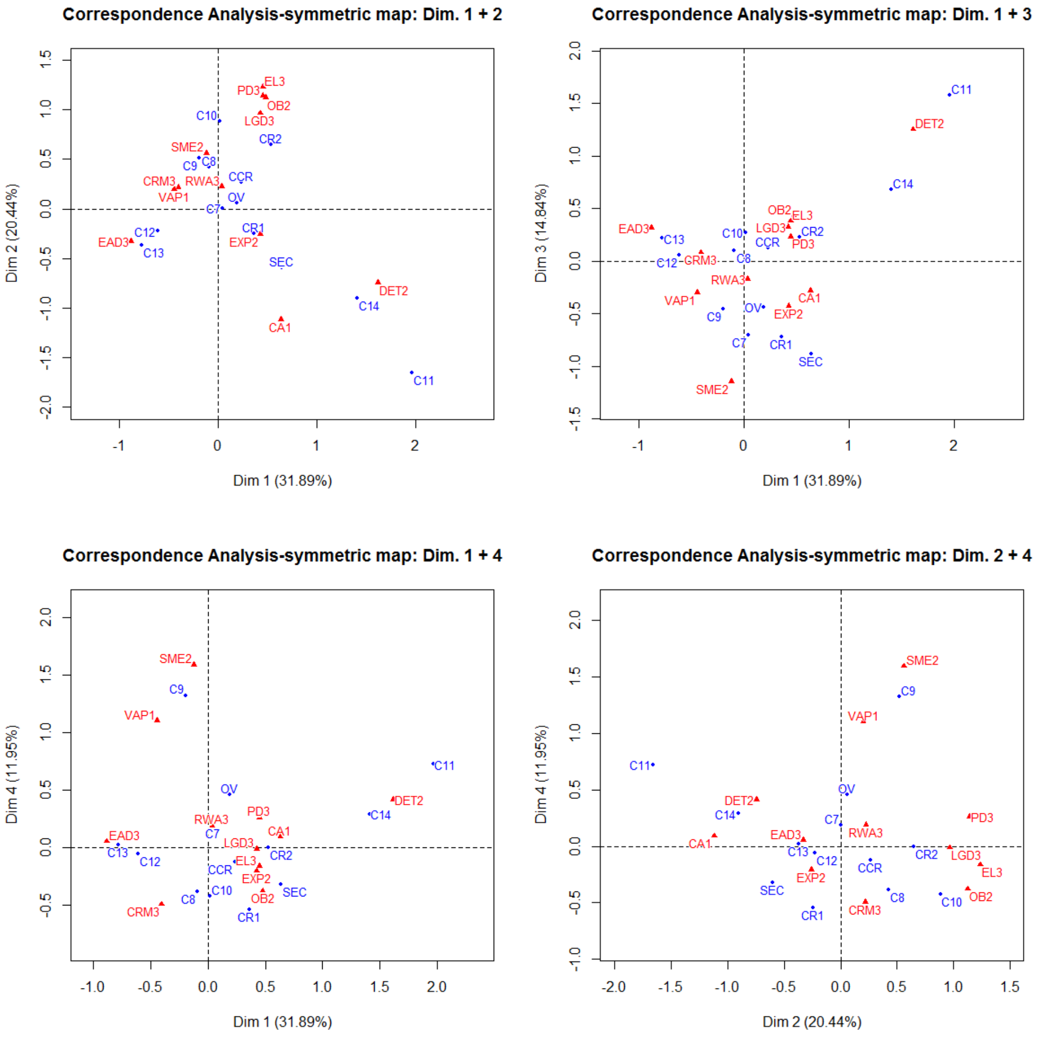


Figure 3. Bar plot of contributions of data with a threshold line (The plots display the contributions of data themes to the first 4 dimension's definition, or in another word we can find which data themes are mainly explaining the dimensions).

From the bar plots (Figure 3), EAD3, DET2 and EXP2 have substantial contributions to the first dimension, and they also have major roles in the definition of the third dimension. Furthermore, PD3 and VAP1 have a large contribution only to the second and fourth dimension, respectively. Given the distribution of the high-level data items present in each template, the similarity of the templates can be illustrated by the symmetric maps depicted in Figure 4.



Correspondence Analysis-symmetric map: Dim. 1 + 2

Correspondence Analysis-symmetric map: Dim. 1 + 3

Correspondence Analysis-symmetric map: Dim. 1 + 4

Correspondence Analysis-symmetric map: Dim. 2 + 4

Figure 4. Scatter plot of CA result to 12-dimensional data in table 5 (The distance between any templates gives a measure of their similarity on the basis of the proportion of themes present in each template. And the location of templates in each dimension indicates their correlation with themes).

Now these maps can be interpreted to find similarities between templates. Take the first plot, which presents the first and second dimension, as an example, in the space mainly defined by EAD3, DET2 and EXP2, C7 is most close to the average profile, and other templates like OV and CR1 are also similar to average or C7. Also, C11 is the most different template, and some pairs of templates, such as C12 and C13, C8 and C9, have a strong association with each other. If interested in other high-level data items, the analyst can proceed to observe the other three plots. For example, if PD3 and VAP1 are take into consideration, we can refer to the fourth plot. It can be seen that the association patterns of templates change in the plot, at least C8 are C9 no longer have a strong correlation.

It also can be assessed that the first dimension is determined by the opposition between EAD3 (negative pole) and DET2 and EXP2 (positive pole), and it is similar in the third dimension. The second dimension is determined by a larger number of themes, and each theme has a relatively smaller contribution. Regarding the greatest four items, PD3, EL3 and OB2 are in the positive pole and CA1 is in negative pole. The fourth dimension is determined by the opposition between VAP1, SME2 (positive pole) and CRM3 (negative pole).

Then it is possible to interpret the position of the templates relative to the dimensions regarding the different influence of each dimension on the templates. Still, take the first plot as an example, the more they lie on the right (the positive side of the first dimension) the more they will be "associated" with DET2 and EXP2 or, put another way, the more DET2 and EXP2 will make a high proportion of their data points. Moreover, the more the templates lie in the upper part the plot, the more they will be correlated to PD3, EL3 and OB2, while CA1 will make a higher proportion of the data points of the templates lying in the lower part of the plot.

To indicate the groups of templates that are similar regarding the distribution of high-level data items, a cluster analysis can be conducted and hence the templates can be isolated into groups in a scatter plot. The algorithm of cluster analysis is as follows (Greenacre M., 2007): Rows (or columns) are progressively aggregated in a way in which every successive merging produces the smallest change in the table's inertia, and this process goes on until the table is reduced to just one row "consisting of the marginal columns of the original table". Here the mathematical details of cluster analysis are beyond the scope of this paper, so the solution provided by the 'FactoMineR' packages of R software were used. Figure 5 plots different combinations of dimensions. The plots tell us that the templates belonging to the same cluster: C12, C13 to cluster 1; C7, C8, C9, C10, OV, CR1, CR2, CCR, SEC to cluster 2; C11, C14 to cluster 3 are those with more similar profiles. In practice, each cluster represents the templates that are jointly more/less frequent across the different departments; hence, simplifications can be made in the reporting architecture to streamline the data generation and auditing processes.

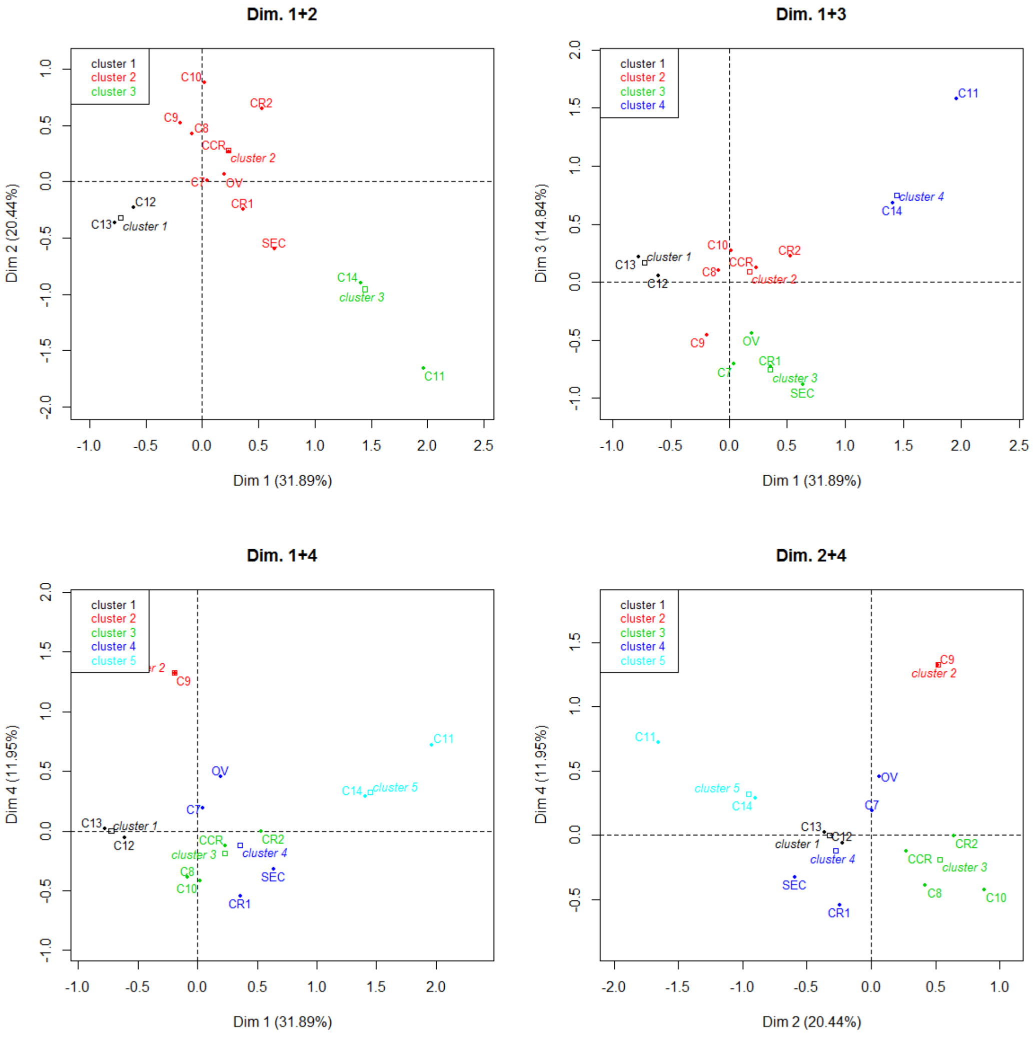


Figure 5. Cluster analysis to templates (The templates belonging to the same cluster are with more similar profiles in terms of particular data themes.)

4.2. Mining and visualization of association rules

To start with, an item frequency plot (Figure 6) was used to find the most frequent templates, i.e. which with high support value. The highest support value among these templates is less than 0.3.

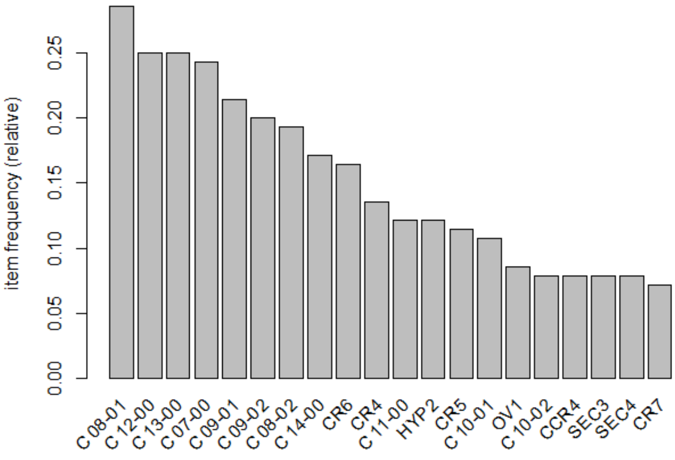


Figure 6. Item frequency for credit risk templates

To uncover association rules among templates, we applied the Apriori method (See Appendix). We set the minimum support σ and minimum confidence δ to be 0.05 and 0.5 respectively, as a result a total of 201 rules were generated. Among them, 53 rules have a rule length (Antecedent and Consequent, a.k.a., lhs and rhs respectively) of 2, 108 rules have a rule length of 3 and 40 rules have a rule length of 4 (Table 8).

To visualize these association rules, the most straightforward method is to use a scatter plot with two measures on the axes (Hahsler and Chelluboina, 2011). In the top plot of Figure 7, the support and confidence were used as X and Y axis respectively, and the lift measure is represented by the color (grey level) of the points. The color key can be found on the right side of the plot. In the second plot, we use lift in the Y axis and confidence as the dot color.

Bayardo Jr. and Agrawal (1999) argue that these rules reside on the support/confidence border are the most interesting ones, which can be easily found in the first plot. Rules with high lift level usually have a relatively low support in the second plot. We conclude that the support value of most of the rules ranges from 0.05 to 1, and the lift values are around 1 to 10. Hence, a support value larger than 0.1 was considered as a high level, and rules with lift values greater than 10 are high lift rules.

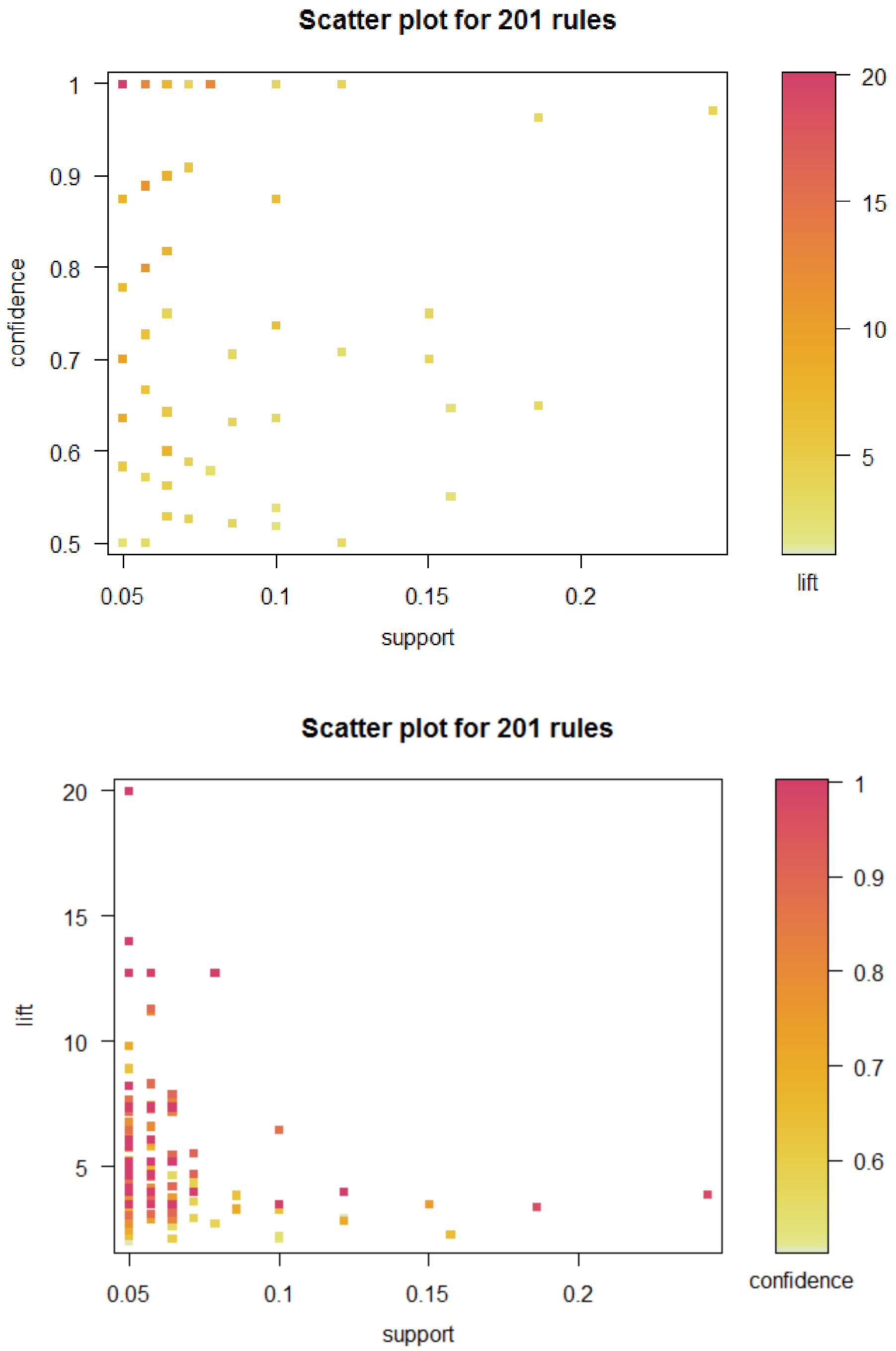


Figure 7.Scatter plot for association rules.

To make the representation of rules clearer and understand the patterns of rules, another graph-based visualization technique provided by the ‘aruleViz’ package in R software was adopted (Figure 8). To avoid cluttering of rules, only a small set of rules were chosen. In this plot, the templates and rules are represented as vertices connecting by directed edges, and the interest measures, here are also support and lift, are displayed by the size and color of the labels on the edges. From Figure 8, we can conclude that the templates can be separated into three independent groups, and templates in each group are more similar with each other..

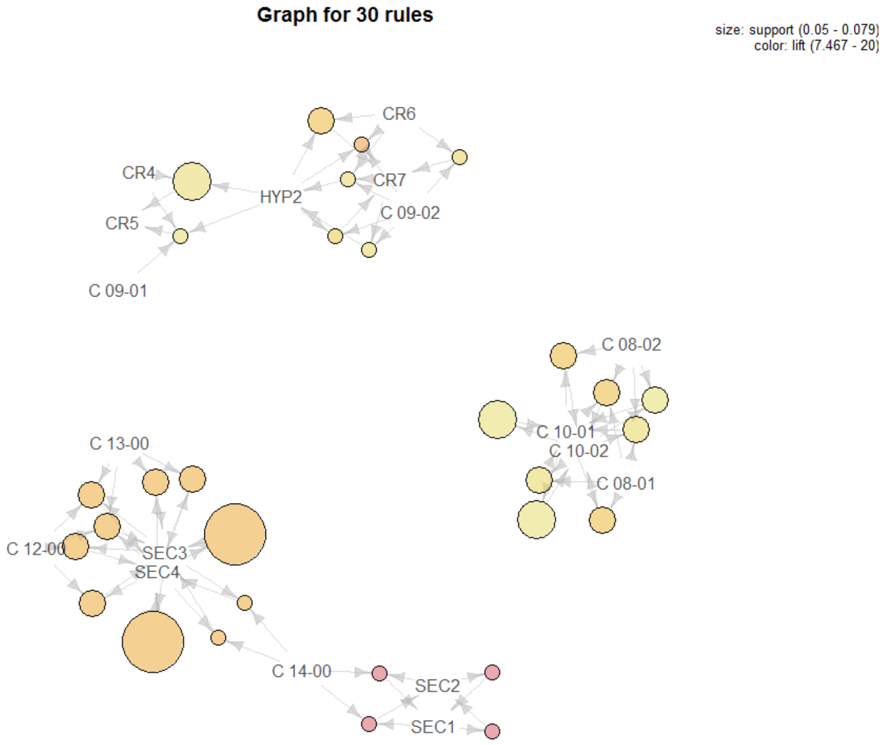


Figure 8.Graph-based visualisation for 30 rules.

5.2.1. High lift and high support rules

High lift implies strong association, so these templates which are highly connected with each other can be obtained by mining high lift rules. We explored the rules with lift larger than 10, thus 19 rules were obtained, and then listed and sorted by lift in Table 9. From Table 9, we can find the most high-lift rules are mainly related to the templates which require data of securitization transactions, such as C 14.00 and C 13.00 in COREP, and SEC1 and SEC2 in Pillar 3. This result can be interpreted as that the templates about securitization transactions in COREP and Pillar 3 have the highest degree of similarity.

In addition, the patterns of high support rules are also worth to be analyzed. This is because strong support means that the template takes a large proportion of the dataset, i.e. most of the data elements. As mentioned previously, these rules with support values greater than 0.1 were selected as top support rules. It can be seen from Table 10 that the high support rules are all about the templates in COREP. This result is highly associated with the characteristic of data granularity of COREP and Pillar 3.

5. Conclusions

Many financial institutions are adopting a 'fragmented response' approach to regulatory reporting, i.e. response to different regulations across different departments, business lines and geographies, which is a low efficiency and high cost way. By verifying the existence of associations between different regulatory reporting frameworks, this paper has demonstrated the reasonableness and superiority of a "harmonized response". The current reporting process could be optimized by creating or designating a specialized team whose responsibility is to centralized process regulatory reporting affairs. It can be inferred that the more templates an institution need to submit, the more linkages can be found among templates and the more efficient the harmonized approach could be.

This paper has demonstrated that there are statistically significant correlations between the templates of COREP and Pillar 3. Hence, the templates are connected regarding data source with the existence of repeated or highly correlated data themes. These significant correlations can be also verified by the high level of total inertia of contingency tables. All this has been explored and visualized with the aid of Association Rules Mining and Correspondence Analysis techniques.

In the long run, the information system of an institution could be improved accordingly. But there is a premise that the connections or affinities among the regulations are identified correctly. Furthermore, as the regulatory requirements are changing constantly, a once-and-for-all solution or system does not exist. In all, this paper adopted Correspondence Analysis and Association Rules Mining to analyses and visualize the relationships between regulations and has proved their applicability and effectiveness. It can be a new start point for financial institutions who would like to develop a truly global, efficient and scalable reporting architecture. The future winners will be institutions that look beyond the basic compliance. A real opportunity presented is to transform the reporting function and make it more responsive to regulatory changes, while simultaneously improving operational efficiencies and reducing compliance costs.

References

Acharya, V. V., & Ryan, S. G. (2016).Banks' financial reporting and financial system stability. *Journal of Accounting Research*, *54*(2), 277-340.

Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *Acmsigmoidrecord*(Vol. 22, No. 2, pp. 207-216). ACM.

Bank for International Settlements 2015.Basel Committee on Banking Supervision (BCBS).Revised Pillar 3 disclosure requirements.ISBN 978-92-9131-546-8 (online). Available at: http://www.bis.org/bcbs/publ/d309.pdf

BayardoJr, R. J., &Agrawal, R. (1999, August).Mining the most interesting rules.In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 145-154).ACM.

BCBS.(2013). Principles for Effective Risk Data Aggregation and Risk Reporting. Available at: http://www.bis.org/publ/bcbs239.pdf

BCBS. (2015). Pillar 3 disclosure requirements – consolidated and enhanced framework (Consultative Document). Available at: http://www.bis.org/bcbs/publ/d309.pdf

BCBS.(2016). Revised Pillar 3 disclosure requirements. Available at: http://www.bis.org/bcbs/publ/d356.pdf

BCBS. (2017). Standards Pillar 3 disclosure requirements – consolidated and enhanced framework. Available at: http://www.bis.org/bcbs/publ/d400.pdf

Beh, E. J., & Lombardo, R. (2014).Correspondence analysis: theory, practice and new strategies. *John Wiley & Sons*.

Covi, G. (2017). The emerging regulatory landscape: a new normal. *Journal of Banking Regulation*, *18*(3), 233-255.

Degryse, H. (2009). Competition between financial markets in Europe: what can be expected from MiFID?. *Financial Markets and Portfolio Management*, *23*(1), 93-103.

EBA (2012). Consultation on data point model related to Implementing Technical Standards on supervisory reporting. Eba.europa.eu. Available at: http://www.eba.europa.eu/-/consultation-on-data-point-model-related-to-implementing-technical-standards-on-supervisory-reporti-1

EBA (2016). European Banking Authority: Data Point Model and Taxonomies - European Banking Authority. Eba.europa.eu. Available at: https://www.eba.europa.eu/regulation-and-policy/supervisory-reporting/implementing-technical-standard-on-supervisory-reporting-data-point-model-

EBA (2017). European Banking Authority: Reporting framework 2.7. Available at: https://www.eba.europa.eu/risk-analysis-and-data/reporting-frameworks/reporting-framework-2.7

Ernst & Young. (2012). Setting the Pace of Change: Bank Regulatory Reporting Survey. Available at: http://www.ey.com/Publication/vwLUAssets/Federal_Reserve:_bank_regulatory_reporting_survey/$FILE/ErnstYoungFRBsurveyreportfinal.pdf

Fisher, I.E., Garnsey, M.R. and Hughes, M.E., 2016. Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. Intelligent Systems in Accounting, Finance and Management, 23(3), pp.157-214.

Fontana, A., &Scheicher, M. (2016).An analysis of euro area sovereign CDS and their relation with government bonds. *Journal of Banking & Finance*, *62*, 126-140.

FSOC (2014).Financial Stability Oversight Council Annual Report - Washington DC. Available at: https://www.treasury.gov/initiatives/fsoc/Documents/FSOC%202014%20Annual%20Report.pdf

Goel, S. and Gangolly, J., 2012. Beyond the numbers: Mining the annual reports for hidden cues indicative of financial statement fraud. Intelligent Systems in Accounting, Finance and Management, 19(2), pp.75-89.

Greenacre, M. (2007).Correspondence Analysis in Practice.*Chapman & Hall/CRC*.

Greenacre, M., &Pardo, R. (2006). Subset correspondence analysis: visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociological methods & research*, *35*(2), 193-218.

Hahsler, M., &Chelluboina, S. (2011).Visualizing association rules: Introduction to the R-extension package arulesViz. *R project module*, 223-238.

Hall, M. J. (2012). Basel II: panacea or a missed opportunity?. *PSL QuarterlyReview*, *57*(230).

Higuera-Mendieta, D. R., Cortés-Corrales, S., Quintero, J.,& González-Uribe, C. (2016). KAP Surveys and Dengue Control in Colombia: Disentangling the Effect of Sociodemographic Factors Using Multiple Correspondence Analysis. *PLoS neglected tropical diseases*, *10*(9), e0005016.

Hortin, N. (2016). The FRTB: Do not underestimate the standardised approach. *Journal of Securities Operations & Custody*, *8*(3), 197-200.

Jaffe, K. "Visualizing the Invisible Hand of Markets: Simulating Complex Dynamic Economic Interactions," Intelligent Systems in Accounting, Finance and Management, vol. 22, no. 2, pp. 115– 132, 2015.

Kleinknecht, M.; Ng, W. L. (2015): "Minimising Basel III Capital Requirements with Unconditional Coverage Constraint", Intelligent Systems in Accounting, Finance and Management, 22(2), pp. 1-19

Költringer, C., &Dickinger, A. (2015). Analyzing destination branding and image from online sources: A web content mining approach. *Journal of Business Research*, *68*(9), 1836-1843.
Lantz, B. (2015). *Machine learning with R*. Packt Publishing Ltd.

Larose, D. T. (2014). *Discovering knowledge in data: an introduction to data mining*.John Wiley & Sons.

Leuz, C., &Wysocki, P. D. (2016). The economics of disclosure and financial reporting regulation: Evidence and suggestions for future research. *Journal of Accounting Research*, *54*(2), 525-622.

Litterman, R. B., &Scheinkman, J. (1991). Common factors affecting bond returns. *The Journal of Fixed Income*, *1*(1), 54-61.

Lorenzo-Seva, U. (2011). Horn's parallel analysis for selecting the number of dimensions in Correspondence Analysis. Methodology, 7(3), 96-102.

Piatetsky-Shapiro, G. (1991). Discovery, analysis and presentation of strong rules. *Knowledge discovery in databases*, 229-248.

Poynter, J. G., Winder, J. P., & Tai, T. (2015).An analysis of co-movements in industrial sector indices over the last 30 years. *Review of Quantitative Finance and Accounting*, *44*(1), 69-88.

Ryan, S. G. (2017). Is Banks' Current Regulatory Capital Adequacy the Mechanism by which their Accounting Requirements Affect Financial Stability?. *Annual Review of Financial Economics*, *9*(1).

Samitas, A., &Polyzos, S. (2015). To Basel or not to Basel?Banking crises and contagion. *Journal of Financial Regulation and Compliance*, *23*(3), 298-318.

Walker, G. A. (2011). Basel III market and regulatory compromise.*Journal of Banking regulation,* 12(2), 95.

Appendix I: Singular Value Decomposition for CA

Consider the$I \times J$data matrix$N$, and the total sum of all values in the matrix is defined as grand total$n = \sum_i \sum_j n_{ij} = 1^T N\, 1$, where 1^Trepresent a $1 \times I$ vector of ones and 1 represent $J \times 1$ vector of ones to match the row and column lengths of$N$.First, compute the correspondence matrix P by normalizing the data matrix *N*:

$$P = \frac{1}{n} N$$

It can be interpreted as a probability matrix, in which an element P_{ij} denotes the probability that the corresponding element N_{ij} appears.Then, we can get a series of metrics, including:

$$Row\ and\ column\ masses: r_i = \sum_{j=1}^{J} p_{ij}\, c_i = \sum_{i=1}^{I} p_{ij}$$

$$Diagonal\ matrices\ of\ row\ and\ column\ masses: D_r = diag(r) D_c = diag(c)$$

It can be noted that the total sum of elements of the three quantities= $\{p_{ij}\}$, $r = \{r_i\}$ and $c = \{c_j\}$ are all equal to 1. The row mass (column mass) is the total frequency of a given row (column).Now apply SVD to P:

- Step 1: Calculate the matrix S of standardized residuals

$$S = D_r^{-\frac{1}{2}}(P - rc^T)D_c^{-\frac{1}{2}}$$

- Step 2: Calculate the SVD of S

$$S = UD_\alpha V^T \text{ Where } U^TU = V^TV = I$$

where D_α is the diagonal matrix of singular values ordered as $\alpha_1 \geq \alpha_2 \geq \ldots$

- Step 3: Standard coordinates Φ of rows, Γ of columns

$$\Phi = D_r^{-\frac{1}{2}}U$$

$$\Gamma = D_c^{-\frac{1}{2}}V$$

- Step 4: Principal coordinates F of rows, $G$of columns

$$F = D_r^{-\frac{1}{2}}UD_\alpha = \Phi D_\alpha$$

$$G = D_c^{-\frac{1}{2}}VD_\alpha = \Gamma D_\alpha$$

- Step 5: Principal inertias λ_k

$$\lambda_k = \alpha_k^2, \ k = 1,2,\ldots,K \text{ Where } K = \min\{I-1, J-1\}$$

We can note that the problem of finding low-dimensional best-fitting subspace is actually a low-rank approximation problem in mathematics. The only adaption is that the approximation under SVD are by weighted least squares, since the weights of rows and columns have been incorporated by their masses. The singular values in $D_\alpha$are the square roots of the eigenvalues of matrices.

According toEckart-Young theorem, if we construct another $I \times J$matrix $S_{(m)}$ from the first $m$columns of $U$and V and first m singular values in D_α as:

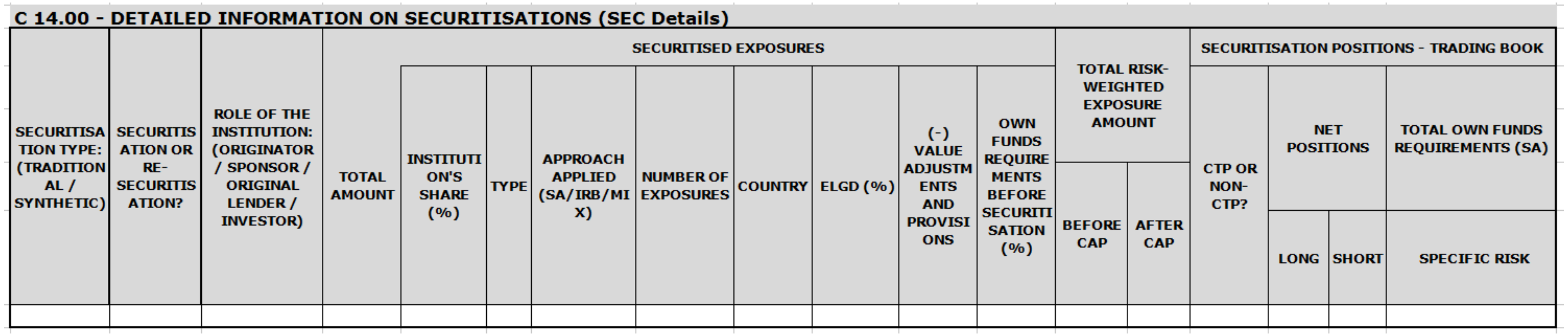
$$S_{(m)} = U_{(m)}D_{\alpha(m)}V_{(m)}^T$$

Then$S_{(m)}$is the least-squares rank $m$approximation of$S$, and the number of dimension has been reduced to m. The percentage of data variability that the approximation matrix can explain is determined by the sum of the m singular valuesor principal inertias(Greenacre, 2007).

Appendix II: Snapshot of example templates

1. Template C14 from COREP

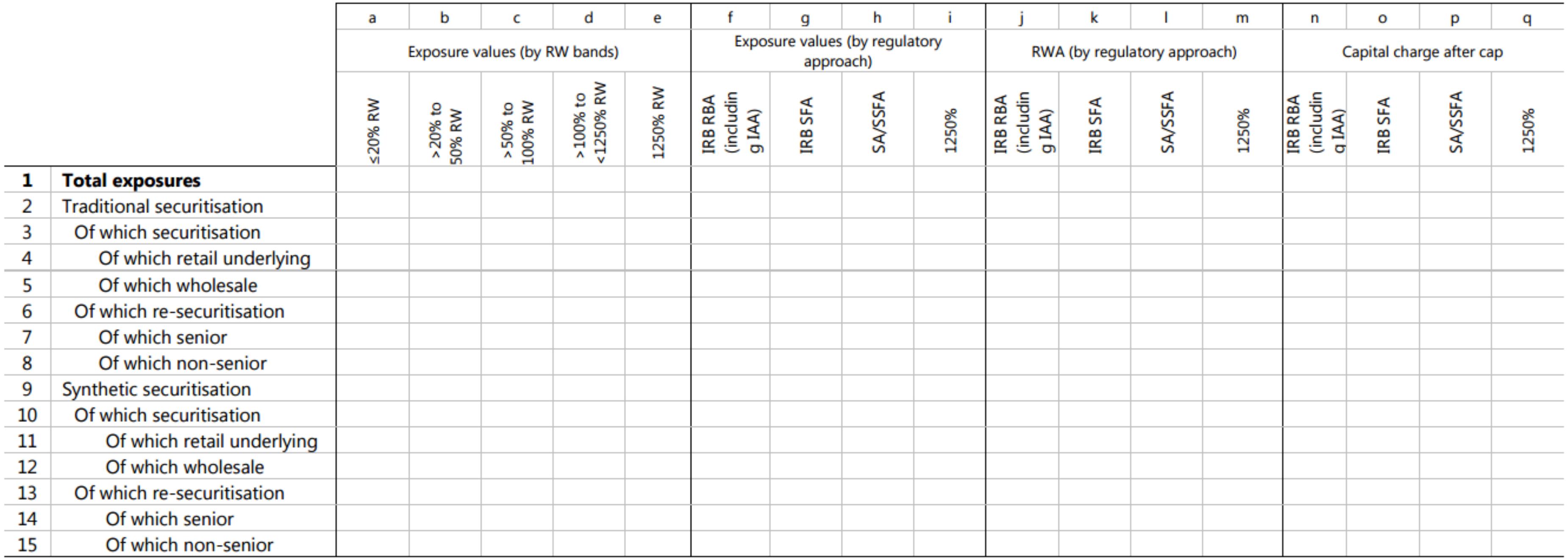
Due to limited space, just part of the template was presented here.



C 14.00 - DETAILED INFORMATION ON SECURITISATIONS (SEC Details)

SECURITISATION TYPE: (TRADITIONAL / SYNTHETIC)	SECURITISATION OR RE-SECURITISATION?	ROLE OF THE INSTITUTION: (ORIGINATOR / SPONSOR / ORIGINAL LENDER / INVESTOR)	SECURITISED EXPOSURES									TOTAL RISK-WEIGHTED EXPOSURE AMOUNT		SECURITISATION POSITIONS - TRADING BOOK			
			TOTAL AMOUNT	INSTITUTION'S SHARE (%)	TYPE	APPROACH APPLIED (SA/IRB/MIX)	NUMBER OF EXPOSURES	COUNTRY	ELGD (%)	(-) VALUE ADJUSTMENTS AND PROVISIONS	OWN FUNDS REQUIREMENTS BEFORE SECURITISATION (%)	BEFORE CAP	AFTER CAP	CTP OR NON-CTP?	NET POSITIONS		TOTAL OWN FUNDS REQUIREMENTS (SA)
															LONG	SHORT	SPECIFIC RISK

2. Template SEC4 from Pillar 3



		a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q
		Exposure values (by RW bands)					Exposure values (by regulatory approach)				RWA (by regulatory approach)				Capital charge after cap			
		≤20% RW	>20% to 50% RW	>50% to 100% RW	>100% to <1250% RW	1250% RW	IRB RBA (includin g IAA)	IRB SFA	SA/SSFA	1250%	IRB RBA (includin g IAA)	IRB SFA	SA/SSFA	1250%	IRB RBA (includin g IAA)	IRB SFA	SA/SSFA	1250%
1	**Total exposures**																	
2	Traditional securitisation																	
3	Of which securitisation																	
4	Of which retail underlying																	
5	Of which wholesale																	
6	Of which re-securitisation																	
7	Of which senior																	
8	Of which non-senior																	
9	Synthetic securitisation																	
10	Of which securitisation																	
11	Of which retail underlying																	
12	Of which wholesale																	
13	Of which re-securitisation																	
14	Of which senior																	
15	Of which non-senior																	

Tables

COREP return category	No of templates	Template no. based on COREP DPM	Reporting Frequency	Submission timing from reference dates
Capital Adequacy	6	C 01.00 to C 05.02	Quarterly	41 days
Group Solvency	2	C 06.01 and C 06.02	Quarterly	41 days
Credit, counterparty credit, settlement and securitization risk	20	C 07.00.a to C 15.00	Quarterly	41 days
Operational risk	5	C 16.00.a to C 17.02	Quarterly and Semi-annual	41 days
Market risk	7	C 18.00 to C 24.00	Quarterly	41 days
Credit value adjustment risk	1	C 25.00	Quarterly	41 days
Large exposures	6	C 26.00 to C 31.00	Quarterly	41 days
Sovereign exposures	2	C 33.00.a to C 33.00.b	Quarterly	41 days
Leverage	8	C 40.00 to C 47.00	Quarterly	41 days
Liquidity coverage ratio	30	C 51.00.a to C 54.00.w, C 72.00.a to C 76.00.w	Monthly	30 days
Net stable funding ratio	8	C 60.00.a to C61.00.x	Quarterly	41 days
Additional liquidity monitoring metrics	16	C 66.01.a to C 71.00.w	Monthly	15 working days

Table1 COREP-templates, timelines and frequencies (EBA, April 2017)[1]

[1] EBA reporting framework v2.7 https://www.eba.europa.eu/risk-analysis-and-data/reporting-frameworks/reporting-framework-2.7

Pillar 3 disclosure category	No of templates	Template no.	Reporting Frequency
Overview of risk management, key prudential metrics and RWA	4	KM1 to OV1	quarterly or annual
Linkages between financial statements and regulatory exposures	4	LI1 to PV1	annual
Composition of capitaland TLAC	6	CC1 to TLAC3	semiannual
Macro prudential supervisory measures	2	GSIB1 and CCyB1	semiannual or annual
Leverage ratio	2	LR1 and LR2	quarterly
Liquidity	3	LIQA to LIQ2	quarterly, semiannual or annual
Credit risk, counterparty credit risk and securitization	29	CRA to SEC4	quarterly, semiannual or annual
Market risk	7	MRA to MR4	quarterly, semiannual or annual
Interest risk	2	IRRBBA and IRRBB1	annual
Remuneration	4	REMA to REM3	annual

Table 2.Pillar 3 -templates and frequencies (BCBS, March 2017)[2]

[2] Standards Pillar 3 disclosure requirements – consolidated and enhanced framework. http://www.bis.org/bcbs/publ/d400.pdf

	C7	C8	C9	C10	C11	C12	C13	C14	OV	CR1	CR2	CCR	SEC	Sum
FI	28	40	120	0	12	48	134	40	22	5	11	0	72	532
FR	255	267	100	48	26	165	229	420	0	192	538	162	468	2870
R	199	696	203	194	0	585	1449	30	132	67	688	257	72	4572
Sum	482	1003	423	242	38	798	1812	490	154	264	1237	419	612	7974

Table 3. Contingency table with 3 categories/rows (FI, FR, R refer to three business functions. The 13 columns named C7, C8, etc., are templates or groups of templates from COREP and Pillar 3. The values are the number of the data points in each template sourced from FR, R and FI.)

Variables	Data Themes
CA1	Capital
VAP1	Value adjustment and provisions
EXP2	Risk exposures, including CR and CCR
SME2	Small and medium enterprise information
OB2	Obligor information, including number and grade
DET2	Business transaction details
CRM3	Credit risk mitigation
RWA3	Risk weighted assets or exposures
PD3	Probability of default
LGD3	Loss given default
EAD3	Exposure at default
EL3	Expected loss

Table 4. List of category variables and corresponding data themes

	C7	C8	C9	C10	C11	C12	C13	C14	OV	CR1	CR2	CCR	SEC	Sum

FI	CA1	0	0	1	0	12	0	52	30	22	0	0	0	72	189
	VAP1	28	40	119	0	0	48	82	10	0	5	11	0	0	343
FR	EXP2	211	173	55	27	2	165	229	220	0	192	326	106	468	2174
	SME2	44	13	45	0	0	0	0	0	0	0	0	0	0	102
	OB2	0	51	0	21	0	0	0	0	0	0	121	28	0	221
	DET2	0	30	0	0	24	0	0	200	0	0	91	28	0	373
R	CRM3	81	395	0	69	0	147	247	0	0	50	17	54	0	1060
	RWA3	85	91	71	29	0	96	190	20	132	16	231	93	72	1126
	PD3	0	30	40	22	0	0	0	0	0	1	151	28	0	272
	LGD3	0	52	22	22	0	0	0	10	0	0	91	28	0	225
	EAD3	33	90	59	26	0	342	1012	0	0	0	91	54	0	1707
	EL3	0	38	11	26	0	0	0	0	0	0	107	0	0	182
Sum		482	1003	423	242	38	798	1812	490	154	264	1237	419	612	7974

Table 5.Contingency table with 12 categories/rows

(The 12 row variables refer to data themes, and the 13 column variables refer to templates or group of templates. The values are the number of the data points in each template belonging to corresponding data themes).

Variables	Corresponding Templates	Source
C7	C 07.00	COREP
C8	C 08.01, C08.02	COREP
C9	C 09.01,C 09.02, C 09.03	COREP
C10	C 10.01, C 10.02	COREP
C11	C 11.00	COREP
C12	C 12.00	COREP
C13	C 13.00	COREP
C14	C 14.00	COREP
OV	OV1, HYP1, HYP2	Pillar 3
CR1	CR1,CR2,CR3,CR4,CR5	Pillar 3
CR2	CR6,CR7,CR8,CR9,CR10	Pillar 3
CCR	CCR1,CCR2,CCR3,CCR4,CCR5,CCR6,CCR7,CCR8	Pillar 3
SEC	SEC1,SEC2,SEC3,SEC4	Pillar 3

Table 6. List of column variables and corresponding templates and source

Eigenvalues

	Dim.1	Dim.2
Variance	0.25	0.07
% of var.	78.97	21.03

Column variables (the 10 first)

	Inertia*1000	Coordinates on Dim.1	Contribution	cos2	Coordinate on Dim.2	Contribution	cos2
C7	7.59	0.35	2.96	0.98	-0.05	0.24	0.02
C8	7.62	-0.23	2.56	0.84	-0.1	1.77	0.16
C9	40.49	-0.04	0.04	0	0.87	60.35	1
C10	6.98	-0.41	2.03	0.73	-0.25	2.8	0.27
C11	8.56	0.94	1.67	0.49	0.96	6.5	0.51
C12	11.04	-0.33	4.39	1	-0.01	0.02	0
C13	54.91	-0.49	21.6	0.99	0.05	0.92	0.01
C14	70.53	1.07	28.06	1	0.01	0.01	0
OV	11.34	-0.69	3.64	0.81	0.34	3.29	0.19
CR1	19.44	0.73	7.06	0.91	-0.23	2.53	0.09

Row variables

	Inertia*1000	Coordinates on Dim.1	Contribution	cos2	Coordinate on Dim.2	Contribution	cos2
FI	62.84	0.09	0.21	0.01	0.97	93.12	0.99
FR	153.72	0.65	60.13	0.98	-0.08	3.88	0.02
R	101.7	-0.42	39.67	0.88	-0.06	3	0.12

Table 7. Summary of CA result to 3-dimensional data. The squared cosine (cos2) indicates the contribution of a component to the squared distance of the observation to the origin.

Rule length distribution (lhs + rhs)

Size	2	3	4
Number	53	108	40

Min.	1st Qu.	Median	Men	3rd Qu.	Max.
2.000	2.000	3.000	2.935	3.000	4.000

Summary of quality measures:

	support	confidence	lhs.support	lift
Min.	0.05000	0.5000	0.05000	2.000
1st Qu.	0.05000	0.7000	0.05000	3.500
Median	0.05000	0.8889	0.64290	4.000
Mean	0.06354	0.8377	0.08060	5.277
3rd Qu.	0.64290	1.0000	0.07857	5.833
Max.	0.24286	1.0000	0.28571	20.000

Table 8.Summary of rules mining results.

rules	LHS	RHS	support	confidence	lift	order
1	{SEC1}	{SEC2}	0.05	1	20	2
2	{SEC2}	{SEC1}	0.05	1	20	2
5	{C 14-00,SEC1}	{SEC2}	0.05	1	20	3
6	{C 14-00,SEC2}	{SEC1}	0.05	1	20	3
15	{C 09-02,HYP2,CR6}	{CR7}	0.05	1	14	4
3	{SEC4}	{SEC3}	0.08	1	12.73	2
4	{SEC3}	{SEC4}	0.08	1	12.73	2
7	{C 14-00,SEC4}	{SEC3}	0.05	1	12.73	3
8	{C 14-00,SEC3}	{SEC4}	0.05	1	12.73	3
9	{C 13-00,SEC4}	{SEC3}	0.06	1	12.73	3
10	{C 13-00,SEC3}	{SEC4}	0.06	1	12.73	3
11	{C 12-00,SEC4}	{SEC3}	0.06	1	12.73	3
12	{C 12-00,SEC3}	{SEC4}	0.06	1	12.73	3
13	{C 12-00,C 13-00,SEC4}	{SEC3}	0.06	1	12.73	4
14	{C 12-00,C 13-00,SEC3}	{SEC4}	0.06	1	12.73	4
16	{C 08-02,C 10-01}	{C 10-02}	0.06	0.89	11.31	3
17	{C 08-01,C 10-01}	{C 10-02}	0.06	0.89	11.31	3
18	{C 08-01,C 08-02,C 10-01}	{C 10-02}	0.06	0.89	11.31	4
19	{HYP2,CR6}	{CR7}	0.06	0.8	11.2	3

Table 9.Rules with the higher lift scores.

rules	LHS	RHS	support	confidence	lift	order
1	{C 13-00}	{C 12-00}	0.24	0.97	3.89	2
2	{C 12-00}	{C 13-00}	0.24	0.97	3.89	2
3	{C 08-02}	{C 08-01}	0.19	0.96	3.37	2
4	{C 08-01}	{C 08-02}	0.19	0.65	3.37	2
5	{C 07-00}	{C 08-01}	0.16	0.65	2.26	2
6	{C 08-01}	{C 07-00}	0.16	0.55	2.26	2
7	{C 09-01}	{C 09-02}	0.15	0.7	3.5	2
8	{C 09-02}	{C 09-01}	0.15	0.75	3.5	2
9	{C 13-00,C 14-00}	{C 12-00}	0.12	1	4	3
10	{C 12-00,C 14-00}	{C 13-00}	0.12	1	4	3
11	{C 12-00,C 13-00}	{C 14-00}	0.12	0.5	2.92	3
12	{C 14-00}	{C 13-00}	0.12	0.71	2.83	2
13	{C 14-00}	{C 12-00}	0.12	0.71	2.83	2

Table 10.Rules with the higher support scores.