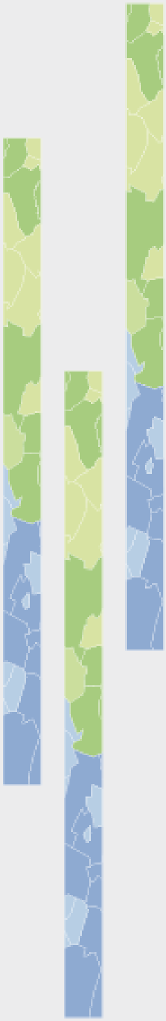
TECHNICAL REPORT

ESPON



A two-step approach to structure the ESPON 2013 DB by themes and sub-themes

MAIN RESULTS

- Database structures are used as source of information to determine, by means of visual grouping techniques, homogenous clusters of words
- We apply text mining methods to derive sub-themes. Here, ESPON evidence is used as textual data to extract potential keywords
- Two distinct parameters of data co-occurrence are employed to understand relational similarities
- We evaluate the explanatory power of words with an algorithm that weights the importance of each word in large corpora of textual data
- Mapping techniques of multidimensional scaling are used to depict similarities and ease interpretation
- Our results suggest that the ESPON 2013 DB should be structured in 7+1 themes and 27 sub-themes

ESPON 2013 DATABASE

MARCH 2011

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

53 pages

LIST OF AUTHORS

Nuno Madeira, University of Luxembourg
Geoffrey Caruso, University of Luxembourg

Contact

E-mail: nuno.madeira@uni.lu; Tel. +352 46 66 44 9691
E-mail: geoffrey.caruso@uni.lu; Tel. +352 46 66 44 6625

Table of contents

1. Introduction

The ESPON 2013 DB aims to improve the access to regional and spatial information. This process has been initiated by the previous ESPON Programme in order to increase the number of variables that may positively support the analysis of spatial structures and trends across European cities and regions (see, for instance, ESPON project 4.1.3).

The goal of this technical report is to structure the ESPON 2013 DB by themes. Besides, it complements the technical report entitled *"Towards an ESPON thesaurus? Some preliminary considerations for the thematic structuring of the ESPON database"* that seeks to derive themes and sub-themes from a corpus of textual data. Here, we argue that database structures, nomenclatures and taxonomies developed by other organisations should be taken into account when structuring the ESPON 2013 DB. The reason for this lies of the fact that many of these structures have established common themes that often aggregate similar data.

First, we focus themes and use that information to analyse similarities between the different database classifications. In addition, we employ matrix visualisation techniques to make the description more comprehensive.

The results should be used to further progress on the user interface prototype and hopefully constitute a robust basis to improve the performance of text mining methods (see previous technical report). Arguably, it is worth mentioning that methods employed in this report will only take into account statistical and geographical sources that have been used to develop indicators by ongoing ESPON projects. In other words, only indicators delivered up to mid-September 2010 will be considered in this analysis.

As a second step, we propose to link each indicator to a theme and sub-theme. Eventually, this process will facilitate harmonisation of codes – variable names – defined in an uncoordinated fashion by TPGs involved in ESPON projects. This is significant because it would offer some consistency to the entire database and assist other research projects when naming variables developed by ESPON to evaluate territorial trends, structures and policy impacts in Europe.

2. Research background and methodology

In this section we provide some research background and describe the methodology applied to determine themes that will eventually constitute the backbone structure of the ESPON 2013 DB.

2.1. Research background

As a first approach we assembled a list of first-level themes established by organisations from which ESPON projects normally obtain raw data, namely UNEP, EEA, EUROSTAT, OECD, UNESCO, WDI, and ILO[1]. This is meaningful because most of these database structures have provided and will continue to provide raw data both in terms of environmental and socio-economic topics to develop ESPON indicators and indices. With this regard, each word or expression used as a theme has been listed, evaluated in terms of similarity, and ultimately aggregated into similar themes. However, we must point out that the aggregation of words into thematic clusters has been purely inductive and based

[1] For more detailed information on each database classification, please visit the following Internet sites: UNESCO (http://stats.uis.unesco.org); ILO (http://laborsta.ilo.org); EUROSTAT (http://epp.eurostat.ec.europa.eu), OECD (http://www.oecd.org/statsportal), EEA (http://themes.eea.europa.eu), UNEP (http://geodata.grid.unep.ch), WDI (http://ddp-ext.worldbank.org).

on the semantic value of each theme. For detailed information, please see Appendix 4 and 5 to this report.

The dataset consists of 85 words or expressions taken from the 7 database structures. Each database has in average 18 first-level themes. Both UNEP and World Bank share the largest structure with 26 themes whereas UNESCO has structured its database with only 6 themes.

A prior step in this analysis is data preparation. The input data matrix is described by a binary (presence/absence) relationship model. That is, all values range between convergent (1) and divergent (0). Table 2.1 lists some of the words (rows) and database classifications (columns) employed in this analysis. If we take the first example, we would be able to understand that '*Tourism*' is considered as a first-level theme by UNEP and EEA while other databases do not devote the same attention to such topic. On the other hand, '*Unemployment*' has only been labelled as a first-level theme by ILO. This is reasonable due to the purposes of each database.

	UNEP	EEA	EUROSTAT	OECD	UNESCO	ILO	WPI
(…)							
Tourism	1	1	0	0	0	0	0
Trade	1	0	1	1	0	0	1
Transport	1	1	1	0	0	0	0
Unemployment	0	0	0	0	0	1	0
(…)							

Table 2.1 Short example of data input for analysis

In order to understand the structure generated by this binary matrix some graphical techniques have been applied to determine clusters, identify blocks within the matrix and increase visual perception of commonly used themes. Following the well-known methods developed by Bertin (1967), we explore the concept of matrix visualization and cluster analysis offered by generalized association plots, or GAP (Chen, 2002; Wu et al., 2008). This open source tool can be understood as recordable matrix to communicate data structures and patterns. Basically it offers the possibility to visualise raw data and display tabular quantities and relationships by means of colour-based representation. The output of such experiment is displayed in a rather natural, inductive perspective but sufficiently helpful to identify proximities between subjects and variables.

The proximity measure one can employ to relate objects in such an experiment depends on the data type (i.e. binary, nominal, ordinal, etc). Within this context, the choice of proximity measure has an effect on the association patterns which directly influences the visual representation of the interaction structure (Wu et al., 2008). GAP offers some specific measurements for asymmetric information. As our matrix corresponds to a binary data type (presence or absence of a theme in a given classification) we have applied Jaccard's coefficient.

2.2. Methodology

The choice of each database derives from the fact that ESPON evidence is strongly based on raw data provided by those above mentioned institutions. As a consequence, it seems appropriate to consider each database classification and validate by means of generalized association plots the degree of similarity and dissimilarity. The usefulness of such approach is to harmonise words and/or expressions used by some of the most prominent statistical databases and therefore provide end-users a common language of understanding.

The matrix visualization is illustrated by a series of images that explore correlation between themes (subjects) and databases (variables). In order to capture potential differences among those databases we decided to include the ESPON 2006 DB structure of first-level themes and identify specific features that could validate or refute our cluster analysis (see Appendix 1, 2, and 3). To this end, each exercise is illustrated by two matrices as an attempt to reveal potential difference. Clearly, some patterns can be discerned from those matrices. Next, we will explore and understand the structure embedded in each data matrix and determine a hierarchy of themes that could support the thematic structuring of the ESPON 2013 DB.

As a first step, we added to our correlation matrix the structure used by the previous ESPON database (ESPON, 2005). Surprisingly, some of the results indicate a weak correlation between ESPON 2006 DB and other structures. Even though, EUROSTAT has the strongest similarity whereas UNEP and UNESCO reveal less significant correlation coefficients. Somehow this reflects how crucial it would be for ESPON to be in accordance with main data providers.

In order to demonstrate the existing similarities between different classifications we employ a simple correlation analysis. The different goals defined for each database led to low correlation values. However, some interesting results emerged from this exercise. For instance, it is clear from the correlation matrix that EUROSTAT and OECD share the strongest correlation value (0.50). One reason that could be claimed to justify the degree of resemblance between these two structures is the nature of its content. Indeed, the fact that EUROSTAT and OECD collect and disseminate similar data for similar audiences has produced an impact on the classification of both databases.

The opposite scenario, i.e. weak correlation values, is rather frequent and little interpretation can be discerned. Still some assumptions must be done due to the degree of clearness, particularly among environmental databases. More precisely, the fact that those sources are committed to cover specific issues such as *environmental hazards*, *marine and coastal areas*, or *air pollution* (see Appendix 2 and 3) that often require detailed data also intensifies the number of discrepancies in most of the themes or categories adopted by each organisation. Preliminary results of this analysis are illustrated by Table 2.2.

	UNEP	EEA	EUROSTAT	OECD	UNESCO	ILO	WPI	ESPON 2006
UNEP	1.00							
EEA	0.22	1.00						
EUROSTAT	0.12	0.27	1.00					
OECD	-0.09	-0.12	0.50	1.00				
UNESCO	-0.10	-0.09	0.14	0.22	1.00			
ILO	-0.20	-0.11	-0.07	-0.10	-0.11	1.00		
WPI	-0.12	-0.15	0.22	0.24	-0.02	-0.07	1.00	
ESPON 2006	0.10	0.18	0.32	0.20	0.07	0.13	0.18	1.00

Table 2.2 Correlation matrix of the database classifications employed in this experiment.

Interesting enough in this analysis is the fact that environmental databases tend to be more detailed when compared with socio-economic databases. To a certain extent, this ensures a high level of accuracy and promotes its utility for large audiences. However, there is an enormous discrepancy on the content provided by each database on environmental topics. On the contrary, both EUROSTAT and OECD have defined a broad list of categories to search and retrieve socio-economic data. As a consequence, semantic similarities are higher and the degree of resemblance between these two entities is much stronger.

Despite the purpose and content of each database it is obvious that organisations do not give much importance to labelling harmonisation. Given the role of the ESPON 2013 Programme for policy advice and development, the ESPON 2013 DB project constitutes a major opportunity to demonstrate the advantages in establishing a harmonised thematic structuring that could rely on classifications defined by the main data providers but also taking into consideration the INSPIRE initiative for the creation of an European Spatial Data Infrastructure.

3. Matrix visualisation techniques for cluster analysis

The methodology applied in this section is partly based on visual grouping techniques to extract words from prominent database structures. We assume that such methods disclose an important capacity to define standards that eventually can lead to improved harmonisation and coherence of spatial concepts and therefore offer the possibility to organise knowledge for information retrieval by end-users. Next we discuss the results obtained with GAP to determine clusters and identify blocks.

3.1. Discussion of preliminary results

The figures presented in Table 3.1 demonstrate a clustering of words. The ranking of words is defined by GAP and includes the structure used in the previous ESPON Programme. The ranking has no absolute meaning but the relative position of words is useful to interpret. GAP ranking is actually the result of a permutation of words so that words that share a similar pattern of presence/absence within the different classifications are positioned in neighbouring rows. Here, we used the single linkage algorithm to obtain the block structure of rows from the permutation. Appendices 1 and 2 to this report display some of the techniques to help identify blocks. Despite its value in terms of matrix visualisation we will give a primary focus to Appendix 1.

Our initial assumption is that GAP offers very helpful features to interpret data matrix association, patterns and ultimately behaviours. This helped to identify some of the key ideas underlying matrix visualisation needs, namely in terms of adopting a practical solution to display matrices. In fact, the main advantage of such tool corresponds to what Wilkinson & Friendly (2009) designated by *cluster heat maps*. The expression itself is very fortune because it gives the idea of clusters by shading association. That is, data matrices structured by similarity and/or dissimilarity to facilitate analysis and interpretation.

In this section we report our results with GAP (Wu et al., 2008). The goal is to illustrate by means of correlation matrices relevant patterns that could easily be interpreted and communicated. More precisely, we propose to find relatively homogeneous clusters of themes. In order to enrich our analysis the number of citations by theme will also be taken into account. Then, we discuss the results from this experiment to propose a first set of themes. Ultimately, these results are compared and clusters interpreted with respect to the indicators delivered up to date (i.e. mid-September 2010).

The preliminary results have provided substantial information to comprehend our data collection. According to Appendix 1 it became clear that certain themes are very representative to the different databases while others are less visible. For instance, if we consider the bottom right hand corner of the first figure on Appendix 1 we observe that the correlation of certain themes (subjects) is very strong among the different database structures used in this experiment. Themes such as *Agriculture*, *Population, Transport* or *Energy* are exceptionally transversal and consequently among the most-cited categories established by certain database classifications. This is significant and somehow justifies the need for adopting such themes within the ESPON 2013 DB.

On the contrary, the second figure on Appendix 1 does not include any reference to the ESPON 2006 DB structure. This was intentional as explained above. Indeed, after computing data the association matrix has slightly changed its appearance. Some of the themes have gained more visibility while others expressed a reverse tendency. However, it should be stressed that the primary group of four themes identified in the previous matrix has been kept very alike (i.e. *Energy*, *Transport*, *Population*, and *Agricultur*e). Similarly, we have identified a less prominent group of themes, mostly clustered on environmental topics, but totally disconnected from the above mentioned cluster. Themes such as *Tourism*, *Land Use*, *Climate*, *Resources* or *Health* loose their importance if not included in the same matrix as ESPON 2006 DB.

Surprisingly enough in this experiment is the fact in both matrices the number of citations is fairly similar, respectively 25% and 28.6%. Two themes, however, react in a different way and demonstrate common behaviours. Both *Tourism* and *Land Use* assume different raking positions when GAP is employed and somehow the percentage of citations reflects that situation. This is extremely relevant because it justifies the ranking of each theme. Ultimately, it confirms that *Tourism* and *Land Use*, two themes credited to the previous ESPON database, are not so important when considering the entire group of words or expressions used in this experiment. An opposite dynamic is observed with *Trade* and *Environment*. Both themes are cited as much as those observed in the first cluster but apparently emerge too disconnected from the structure if the ESPON 2006 DB is considered. Despite this situation, it is clear that such themes should be aggregated to the first set of themes for the ESPON 2013 DB. Besides, it would compensate some of the environmental-oriented themes identified previously (i.e. *Water, Climate, Consumption*, *Resources*).

The results summarized by Table 3.1 reveal as well other groups of themes that may require further attention. The main feature of the fourth group is related with the predominant focus on socio-economic issues. Themes such as *Finance*, *Development*, *Science*, *Infrastructure* or *Education* assume greater importance within this cluster. On one hand, this is essentially due to the ranking defined by GAP when grouping themes that intersect both OECD and EUROSTAT database structures. On the other, it justifies the fact that most of these themes are linked to economic, social and development-oriented data. Nevertheless, it is also clear from Table 3.1 that an independent subgroup emerges within this primary group of themes. Indeed, it seems that the choice of computing a correlation matrix without including the ESPON 2006 DB structure has produced some significant impacts on the permutation result, especially on the position of *Technology*, *Fisheries* and *Industry*. Our interpretation is that those themes are strongly linked with the classification adopted by EEA and the motivation for this behaviour seems to be related to the fact that ESPON has not been considered in one of those occasions.

From this point onwards the structure is much more balanced both in terms of ranking and number of citations. This means that little interpretation can be discerned if the ESPON 2006 DB structure is employed by one of the correlation matrices. Next, we argue that those less prominent themes should be included or grouped within bigger groups since most of them are often related to a specific theme. This process has been developed in a rather inductive way and merely based on the semantic value or weight attributed to each theme. That is, the meaning of a given word (or expression) will define its value or weight when compared with themes and therefore determine the level of closeness.

As stated initially, this section justifies the choice of aggregating some themes that otherwise would be completely disconnected from our analysis. Consequently, we should stress that this experiment has to a considerable extent been influenced by the level of semantic closeness to other major themes previously identified. Against this background, it seems obvious that an important set of less prominent terms (or expressions) should be treated as environmental-oriented issues. A strong argument to support this view is related to the fact that most of those themes derive from environmental database structures such as EEA or UNEP. Thus, it is not surprising that our aggregation method considered domains on *Biodiversity*, *Waste*, *Elevation*, or *Slopes* as traditional environmental issues. The same applies to socio-economic issues largely labelled as integrative components. For instance, we noticed that *Taxation*, *Market Regulation*, *Employment*, *Labour*, or *Wages* can be understood as basic socio-economic themes that characterize the diversity of data published and disseminated by OECD or ILO on their own websites.

Table 3.1 GAP ranking of words or expressions used as a theme

Themes	GAP Ranking (including ESPON 2006)	GAP Ranking (excluding ESPON 2006)	Number of citations, including ESPON 2006 (%)	Number of citations, excluding ESPON 2006 (%)	Groups
Agriculture	1	1	62.5	57.1	
Population	2	2	75.0	71.4	
Transport	3	5	50.0	42.9	
Energy	4	6	50.0	42.9	(1)
Tourism	5	17	37.5	28.6	
Land use	6	19	37.5	28.6	
Climate	7	14	25.0	28.6	
Water	8	13	25.0	28.6	
Urban	9	15	25.0	28.6	
Consumption	10	16	25.0	28.6	
Resources	11	18	25.0	28.6	
Health	12	20	25.0	28.6	(2)
Trade	13	4	50.0	57.1	
Environment	14	3	62.5	71.4	(3)
Finance	15	11	37.5	42.9	
Development	16	22	37.5	28.6	
Social	17	10	50.0	42.9	
Regional	18	26	25.0	28.6	
Science	19	12	37.5	42.9	
Technology	20	9	62.5	57.1	
Fisheries	21	8	37.5	42.9	
Industry	22	7	37.5	42.9	
Communication	23	21	37.5	28.6	
Infrastructure	24	25	37.5	28.6	
Economy	25	24	25.0	28.6	
Education	26	23	37.5	42.9	(4)
Air	27	27	12.5	14.3	
Biodiversity	28	28	12.5	14.3	
Chemicals	29	29	12.5	14.3	
Coastals	30	31	12.5	14.3	
Waste	31	30	12.5	14.3	
Soil	32	32	12.5	14.3	
Seas	33	33	12.5	14.3	
Scenarios	34	34	12.5	14.3	
Pollution	35	35	12.5	14.3	
Noise	36	36	12.5	14.3	
Welfare	37	60	12.5	14.3	
Demography	38	61	12.5	14.3	
Taxation	39	62	12.5	14.3	
Services	40	63	12.5	14.3	
Productivity	41	64	12.5	14.3	
Patents	42	65	12.5	14.3	
Market regulation	43	66	12.5	14.3	
Globalisation	44	68	12.5	14.3	
Information	45	67	12.5	14.3	
Boundaries	46	49	12.5	14.3	
Vegetation	47	50	12.5	14.3	
Elevation	48	52	12.5	14.3	
Threatened (species)	49	51	12.5	14.3	
Slopes	50	53	12.5	14.3	
Fertilizer	51	57	12.5	14.3	
Food (supply)	52	59	12.5	14.3	
Pesticides	53	54	12.5	14.3	
Marine	54	55	12.5	14.3	
Land cover	55	56	12.5	14.3	
Hazards	56	58	12.5	14.3	
Employment	57	44	25.0	14.3	
Labour	58	48	37.5	28.6	
Household	59	39	37.5	28.6	
Wages	60	40	12.5	14.3	
Consumer price (indices)	61	42	12.5	14.3	
Unemployment	62	41	12.5	14.3	
Strikes & lockouts	63	43	12.5	14.3	
Occupational (injuries)	64	45	12.5	14.3	
International labour migration	65	46	12.5	14.3	
Hours of work	66	47	12.5	14.3	
Wealth	67	-	12.5	-	
Spatial typologies	68	-	12.5	-	
Research	69	-	12.5	-	
Public sector	70	-	12.5	-	
Culture	71	37	12.5	14.3	
Literacy	72	38	12.5	14.3	
Balance of payments	73	69	12.5	14.3	
Exchange rates & prices	74	70	12.5	14.3	
External debt	75	72	12.5	14.3	
Governance	76	73	12.5	14.3	(5)

For those terms (or expressions) where uncertainties arise we adopted a more pragmatic solution. Themes like *Globalisation*, *Governance*, or *Welfare* which may be interpreted as very general concepts with meanings that often gravitate between different subjects, we decided to analyse what type of data was being labelled as such. Here, we noticed that such themes have not been equally considered by the database structures employed in this experiment. Somehow, this explains the singularity and different purposes attached to each database.

3.2. Towards a first set of themes

The thematic structure of the ESPON 2013 DB should not be seen as a normative approach, but rather as a descriptive one. However, the choice of themes itself is very crucial for the success of the ESPON 2013 Programme because it offers the possibility to support policy developments which, in turn, can and will be used by different target groups who wish to promote policy documents, technical reports, or academic studies. Moreover, data publically available for retrival can be used as a source for developing trends and scenarios.

This represents a significant gaining for policy development on European spatial planning but most likely is subject of criticism. Indeed, one could ask if this theme or that were emphasized more, or if an attempt was made to add one theme or another. We believe that our preliminary results should be seen as images of the future or as elements that correspond to the needs of a particular moment. We listed below a first set of themes to help end-users to understand the structure we propose for the ESPON 2013 DB. Taking into consideration the methodology applied in this experiment, we labelled themes as follows:

01. Agriculture and Fisheries

02. Demography

03. Transport

04. Energy and Environment

05. Land Use

06. Social Affairs

07. Economy

99. Cross-Thematic and Non-Thematic Data

9901 Integrative indices, typologies and scenarios

9999 Geographical objects

Table 3.2 Themes proposed to structure the ESPON 2013 DB

This list assembles themes used by some of the main data providers. Occasionally, the meaning of the word derives from similar terms or expressions. This has been the case for *Social Affairs* that often recalls societal-related issues that have great effects on many members of those societies and, for that reason, considered to be problems (e.g. poverty, unemployment) or subjects that need further improvement (e.g. healthcare, education). We also add a group to cover cross-thematic and non-thematic data. A first sub-theme includes variables that mix themes on purpose (e.g. integrative indicators, complex typologies, scenarios). The second sub-theme refers to base maps (administrative units) and other geographical objects (e.g. grids, cities, networks) or spatial delineations (e.g. morphological zones, functional areas).

Those themes that have not been mentioned in this list should be considered as less interesting for the moment, although this assumption should not be taken as granted.

Besides, it is not feasible to address all the relevant political, environmental or social aspects. Nevertheless, we can still consider different approaches to conjecture about the degree to which different topics will develop and gain more or less visibility. Here, we argue that text mining methods and tools have the capacity to identify key words on documents that both employ and communicate ESPON evidence and results. We assume that such approach would contribute to a comprehensive thematic structuring of the ESPON 2013 DB. For the moment, it is not obvious that this analysis will introduce new themes or sub-themes within the predefined structure. The emphasis on a particular theme also depends on other variables such as data deliveries (i.e. indicators, indices, typologies), but also the demand from users and potential users, or even EU policy agenda. Whether this occurs or not, many other themes and sub-themes are likely to be added to the ESPON 2013 DB.

4. Text mining methods and visualisation tools

In the previous section of this report we assume that database structures adopted by international organisations represent an important source of information to extract themes. For this purpose, we apply a visual grouping technique to illustrate, by means of correlation matrices, homogenous clusters of words. The rationale defined for sub-themes is slightly different. Here, we use a large collection of textual data to extract potential keywords to label sub-themes. More concretely, for each of the themes that emerged from our experiment with GAP we employ text mining methods and tools on documents related to each of those themes.

The goal of text mining is to find patterns across textual data and, therefore, derive new information. Such methods enable users to identify keywords that, inductively, create thematic overviews of large text collections. Against this background, we argue that text mining methods may positively support the thematic structuring of the ESPON 2013 DB. It is accepted that ESPON introduced new vocabulary of spatial concepts which, in turn, have definitely influenced the terminology adopted by EU institutions. We make use of this evidence to extract keywords from qualitative and unstructured data, in particular ESPON scientific reports but also texts delivered by EU institutions that use or make reference to ESPON evidence. In order to further enrich this approach we depict the results with mapping techniques of multidimensional scaling.

4.1. Methods

The goal of our investigation is to identify keywords on textual data according to their co-occurrence and use that information to conveniently structure the ESPON 2013 DB. As explained in the introductory part of this report, our contribution will only focus on keywords that can be used as sub-themes. For this purpose, we employ the findings of the previous section where we propose a list of 7+1 first-level themes. Primarily, it is important to identify documents that potentially address each of the themes proposed in the last section of this report and, secondly, ensure that we integrate ESPON reports with evidence-based knowledge on European territorial potentials and dynamics.

The most challenging task before applying text mining tools is data preparation (Berry, 2004; Weiss et al. 2005). Due to the fact that textual data is unstructured and often arranged inconveniently it is necessary to follow certain procedures to ensure some consistency to the overall process. The first step is obviously to collect data. In our case this represents any document, study or policy note addressing ESPON evidence and results.

For this purpose, we have initially identified 27 final project reports delivered by the ESPON 2006 Programme. Our desk research expanded then to reports published in the current ESPON Programme and documents released by other sources that offer a wide range of perspectives to ESPON knowledge (e.g. European Parliament, European Commission). In total, we have collected 53 documents (see Annexes 6-12). Altogether, these documents constitute a large corpus of textual data that needs to be structured as efficiently as possible before applying any methodological approach.

Similarly, we have to bear in mind that textual data is a complex conjunction of words and phrases that frequently need to be considered as a whole. There is a quite huge amount of dependency that should not be ignored. Moreover, it is also important to overcome word and semantic ambiguities that may adversely influence our analysis. To this end, the usability offered by the software WordStat is quite straightforward and no additional expertise is needed (Lewis, 1999; Davi et al., 2005). The pre-processing of our text collection took into account some of the features offered by this text mining module, particularly with regard to stop-word lists and lower case conversion.

One of the most interesting features provided by this software is the compilation of non information-bearing words that basically exclude terms without any predictive capability, such as articles, pronouns or prepositions. These words are often characterised as noise data and hardly add new information. Besides, it is also possible to add more words to this dictionary of stop-words and improve the accuracy of the corpus for analysis. It should also be mentioned that WordStat merely records the number of times a word appears within a text regardless the content of a sentence or paragraph. After computing data there is a wide variety of ways in which the result can be displayed. The most basic output offered by this application is the word frequency distribution. This knowledge will constitute the basis to explain our results.

As a first step, we apply a pre-defined list on non-information bearing words with no semantic value. Next, we make use of word lemmatization to reduce inflectional form of words to a common root word and ultimately exclude words based upon a frequency criterion. With this regard, we suggest that words below 100 occurrences should be ignored from our analysis. This option exposed a significant number of words that allowed us to further analyse the knowledge structure in text collections.

However, several authors state that words with high frequency distribution do not offer a solid basis for analysis. With this regard, Luhn (cited by Blanchard, 2007: 309) says that ‘mid-frequency words in the distribution have both high significance and high discriminating power’. This means that words above an upper cut-off and below a lower cut-off should be removed and, as a consequence, more effort should be added to those that have a mid-frequency (see Figure 1). We took into consideration these aspects and defined a threshold from the estimated densities in order to make emerge other words with explanatory power. The threshold for words with high explanatory power is based on the difference between term frequencies. In other words, the highest gap determines the cut-off.

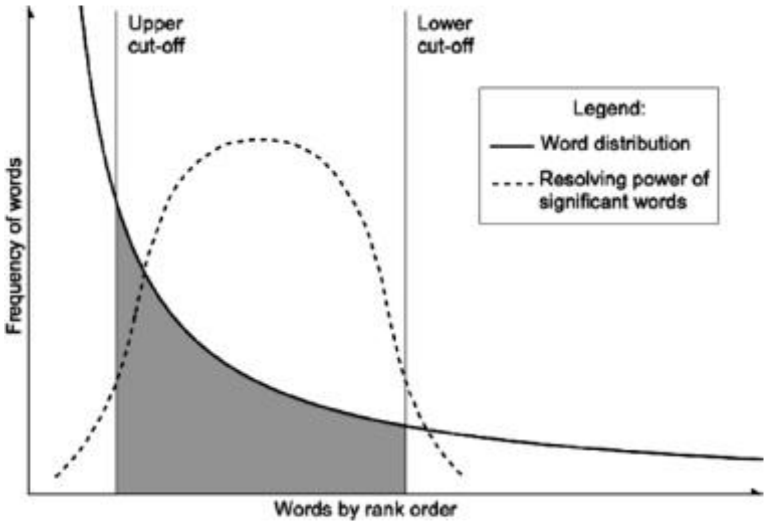


Figure 1: Illustrative plot of word distribution. The power of significant words in higher between the two cut-offs (in grey). Source: Luhn (cited in Blanchard, 2007: 310).

In order to further evaluate the explanatory power of terms in large corpora of textual data, we decided to use TF*IDF (i.e. Term Frequency * Inverse Document Frequency). Basically, this scoring algorithm is a weight that measures the importance of each term in a collection of documents. As a result, the importance increases proportionally to the number of times a term appears in a specific document but is counterbalanced by the frequency of the term in the same corpus or text collection. To a certain extent, TF*IDF algorithm computes the relevance of a document with respect to a particular term. Here we limited our analysis to the 100 most significant terms.

In addition, other features have been applied. Our initial assumption also questioned the ability to use such tools to determine co-occurrence analysis on data derived from a large textual corpus. This function is available on WordStat and it can be used in a variety of ways. For this experiment we have considered two distinct parameters of data co-occurrence – paragraph and group of 5 words. The output file is square matrix where the rows and columns refer to the words and the value of each cell corresponds to the number of co-occurrences. However, it is important to point out that the methods described in this part of the report are solely applied for each of the seven themes that came out from our analysis with database structures. That is, we selected a set of documents for each theme and, based on these documents, prepared data for empirical analysis.

Ultimately, we employ mapping techniques to depict similarities. According to Yang et al (2008), one of the crucial needs in text mining is the ability to visualize relationships between words extracted from large text collections. In order to simply the analysis of data co-occurrence for words with both explanatory power and non-explanatory power we apply a mapping technique of multidimensional scaling to display information, i.e. VOSviewer (Eck & Waltman, 2007; Eck & Waltman, 2008; Eck et al, 2010). The comparison of these two approaches is also subject to discussion.

4.2. Results

In this section we describe the results obtained for the seven themes that came out from our analysis with database structures. In addition, we review the rationale defined for the theme on both cross-thematic and non-thematic data. For each theme and sub-theme we provide an ordinal code of, respectively, two and four digits to simplify membership. In addition, we coded the last theme with the number 99 so that more themes can be added to the ESPON 2013 DB. Apart from the theme on cross-thematic and non-thematic data, the interpretation of the other themes solely refers to the annexes present in this report.

4.2.1. Agriculture and Fisheries

The maps illustrated in Figure 1a and 1b correspond to the density view of data co-occurrence in a multidimensional scaling. For data on 'Agriculture and Fisheries', we selected 7 texts that combine ESPON evidence and results on issues related to the above mentioned topic (see Appendix 6).

The comparison of these maps suggests that words in Figure 1a are less concentrated that Figure 1b. Somehow this can be explained by the dimension of each parameter for data co-occurrence. Despite the relevance of this aspect what is interesting here is that, in both cases, several densities can be depicted. The main densities aggregate more or less the same number of words and therefore similarities are very alike. This is quite evident for inflectional variants of words, such as 'territorial', 'develop', and 'impact', or 'structur', 'farm', and 'agricultur'.

As stated in the methodology part of this report, the threshold for words with high explanatory power is based on the difference between word frequencies. As it can be seen in Figure 1b, such approach increases the importance of other words, such as 'environment', 'farm' and 'product'. Despite the significance of this adjustment to identify an explanatory power on words with mid-frequency distribution, the exclusion of anchor words produces and effect that suggests the concentration of many words in just one individual density. This is particularly clear on both parameters.

The ambiguity of the results observed in the previous maps is now reduced to clear and visible densities of words. In addition, it also indicates that terms with high frequency distribution may also discloser similar TF*IDF importance. That seems to be the case for 'territori' and 'impact'[2]. If we look to the results in more detail, it also becomes evident that rural development and the Common Agricultural Policy receive similar attention.

The number of sub-themes as well as the labelling to describe the content of each sub-theme is therefore based on textual data co-occurrence by means of visualisation techniques in VOSviewer. The sub-themes related to 'Agricultural and Fisheries' looks as follows:

01 AGRICULTURE AND FISHERIES

0101 Farm structure (e.g. farm type, income from farming, organic farming)

0102 Livestock (e.g. livestock output)

0103 Aquaculture and sea fisheries (e.g. aquaculture resources in coastal areas)

0104 Forestry (e.g. production, consumption, and import/export products)

0105 Rural characteristics (e.g. rural employment, rural access to services)

The first sub-theme is designed to incorporate data on farm type, size of farms, income from farming and other related components, such as organic farming. The decision to set out this sub-theme derives from the relatedness between 'farm', 'agricultur', and 'structur' (see Figure 1a and 1b). The third sub-theme is meant to integrate data on 'aquaculture and sea fisheries' and its presence is justified by the high TF*IDF distribution observed for the word 'aquacultur'.

The same applies for 'Forestry' and 'Rural characteristics'. Here, the sub-theme on 'Forestry' is meant to cover data on production, consumption, and trade, while the other reflects the increasing diversity of rural areas in Europe that demand a better understanding of the development opportunities and challenges for the countryside. Therefore, the goal is to address the differences and similarities in regional economic structures through indicators and indices on diversity patterns, including access to services, rural employment, and core-periphery structures.

4.2.2. Demography

The capacity of visualizing and analysing data co-occurrence matrices with VOSviewer reduces the complexity inherent to the content of texts. The collection of documents in the field of 'Demography' (see Appendix 7) indicates a strong emphasis on subjects associated with the following root words: 'popul', 'migrat', and 'labour'. The co-occurrence of such words, regardless the parameter used in this analysis, is quite evident on Figure 2a. Hence, one could assume that the presence of such words is highly important for analysis and eventually contribute to the definition of sub-themes. This assumption is not completely accurate and, as explained previously, some adjustments should be added to our research by withdrawing those words from our corpus and therefore identify other potential words with explanatory power. If we look to both maps in Figure 2b, we recognize inflectional

[2] Of course, documents chosen to integrate the corpus for analysis might have influenced the results. But, in any case, this can be explained by the strong emphasis given to ESPON projects on tools and methodologies to evaluate the effects of agricultural policies on the EU territory.

forms of specific words strongly related to the theme on ‘Demography’. This is the case for ‘fertile’, ‘develop’, ‘scenario’, ‘trend’, and ‘rural’.

But what would happen to our dataset if we just consider words that emerge frequently in a document but do not have the same frequency in the remainder of the corpus? In other words, what would be the result of our analysis if we just keep a maximum of 100 words based on TF*IDF distribution? With this regard, some of the major words identified previously can still be seen in Figures 2c and 2d. Despite the usefulness of this approach, Figure 2c does not improve the distribution observed in Figure 2d. By excluding words that demonstrate high TF*IDF distribution, the density view in both maps is rather similar. Somehow this feature suggests that our corpus is not sensitive to such boundaries and words with high frequency distribution tend to be dispersed more homogenously across documents (e.g. ‘labour’, immigrate’, ‘depopula’).

Having said that, it is important to interpret the results and derive sub-themes related to this specific theme. Some of the terms are implicitly associated with migration (e.g. immigration, migration replacement) or natural change in population (i.e. fertility, mortality, life expectancy) while others demonstrate close links to demographic ratios, trends, and scenarios in rural, urban and peripheral areas. Surprisingly enough we observe a constant correlation between ‘labour’ and ‘forc’. The fact that these two inflectional forms of words disclose a strong correlation is normal, what is not normal is to see these words isolated from other words equally relevant, such as ‘ageing’ and ‘decline’. The topic itself has been on the EU agenda for some years now and many policy measures have been implemented to address the issue of ageing labour force and ageing in general.

Despite the feasibility of text mining techniques in analysing documents related to specific themes or subjects, we would expect to find other words with high explanatory power. This seems to be the case for ‘household’, ‘health’, or ‘dependency’. Nevertheless, we believe the information collected for analysis is solid and robust to extract sub-themes. The structure of second-level themes related to ‘Demography’ looks as follows:

02 DEMOGRAPHY

0201 Population structure (e.g. age distribution by group and gender)

0202 Natural changes (e.g. fertility, mortality, life expectancy)

0203 Households (e.g. number of households)

0204 Migrations (e.g. migration replacement, high-skilled labour migration)

The sub-theme related to ‘population structure’ is meant to understand how EU cities, regions and other territories are made up of people of different ages, and of males and females. Another sub-theme that emerged from our analysis concerns ‘natural changes’. This topic is widely covered by statistical databases and is intended to describe the difference between the number of births and deaths. On the contrary, ‘households’ refers to an individual or a group of individuals who occupy the same dwelling. Here, it is suggested that TPGs should integrate data on persons per households, expenditures per household, but also median income per household member. Finally, the sub-theme ‘migrations’ is expected to cover data related to international migrations, including migration replacement and high-skilled labour migration. Alongside natural changes, the balance of international migrations is an important component of population growth. In some cases, negative natural population growth can be fully compensated by migration surplus.

The analysis performed on documents related to ‘Demography’ sets out four distinct sub-themes. These sub-themes try to cover the most significant topics present in the corpus for our initial experiments. It is worth mentioning, however, that this proposal only takes into consideration textual data delivered up to now on ESPON evidence and results, including inception, interim and final reports of both ESPON 2006 and ESPON 2013 Programmes. In the future, more documents should enrich the corpus of this theme and eventually extract new terms and label other potential sub-themes.

4.2.3. Transport

Transport-related issues have always represented a major issue for analysis and discussion within the ESPON Programme. Our corpus features 6 documents that explore subjects related to the transport sector in Europe (see Appendix 8). The major topics under discussion stem from the need to monitor the EU transport policies, such as mobility, accessibility, sustainable transport policies, impacts on the environment, competitiveness of the economy, or leverage effects on EU territorial development. With this regard, a series of applied research projects commissioned by the ESPON 2006 Programme dedicated a large attention to the Trans-European Transport Network (TEN-T) Initiative. Its importance for competitiveness and growth has been recognized as one of the most fundamental initiatives to set out a proper EU transport policy. The findings of these studies enriched many policy discussions on the developments of the TEN-T with new data on transport networks and traffic flows, including performance indicators, typologies, and scenario-based projections.

The mapping perspective of data co-occurrence in Figure 3a varies according to the parameter. To a large extent, it is correct to say that we observe the same densities. For instance, 'transport', 'access', 'impact', or 'network' confirm their importance. But when we exclude these terms from our dataset the result conveys new densities or clusters of words. This is the case for 'infrastructur', 'develop', 'model', 'scenario', 'road', and 'rail'. If we now focus our analysis on terms with high TF*IDF distribution, we see that some of the terms identified previously become less visible or, inclusively, unobserved. Due to the indexing rationale behind TF*IDF measure, words like 'flow', 'model', 'ten', or 'gdp' gain more visibility. To some extent, this situation increases the robustness of the dataset by focusing on terms with high levels of explanatory power. As it can be seen in Figure 3c and 3d, the best terms to label sub-themes correspond to 'model', 'scenario', and 'ten'.

Clearly, one of the best densities that can be identified in both Figure 3c and 3d correspond to a group of words that include 'air', 'rail', 'maritim', 'traffic', and 'flow'. Somehow this shows a strong focus of our corpus on transportation systems. Similarly, it seems that 'scenario', 'model', and 'forecast' suggest a distinct density. In this case, one could say that the impact of EU transport policy compromises medium and long-term scenarios. Also important, but less merged with other words, is 'ten'. Here the word corresponds to Trans-European Transport Network (TEN-T) and most of the maps depict data without making any sort of similarity linkage to other potential words. The only exception are the maps of data co-occurrence based on TF*IDF distribution.

In sum, we believe our analysis discloses some hidden knowledge on a dataset of text information related to 'transport' and, by doing so, facilitates the decision-making process by reducing uncertainly and doubt on words to label sub-themes. Against this background, we suggest that the above-mentioned theme of the ESPON 2013 DB should integrate the following sub-themes:

03	TRANSPORT
0301	Accessibility (e.g. performance indicators, multimodal accessibility)
0302	Flows (e.g. vehicles, passengers, goods, freight)
0303	Infrastructure (e.g. transportation systems, railways, airports, harbours)

To a certain extent, the proposal of sub-themes to allocate data in the field of 'Transport' is similar to the one suggested in the ESPON 2006 Database (ESPON, 2005). Nevertheless, we believe that such proposal combines a comprehensive structure of transport information, ranging from data on flows of vehicles, passengers and goods to infrastructures, safety, and investments in the transport sector. It considers as well data that could be delivered on sustainable development, modal split, and environmental impact indicators, including the contribution of each mode of transport, used alone and in combination with others.

4.2.4. Energy and Environment

We assume that energy and environment are complementary and, in many ways, essential for sustainable development. Several policy documents delivered by EU institutions state that sustainable development corresponds to the improvement of citizen's quality of life while reducing the use of natural resources and pressures on the environment (CEC, 2001; EEA, 2002). However, the quality of life is enhanced by costly energy services. The main question, according to these institutions, is how to make use of available energy resources without preventing the needs of future generations. In order to meet the right balance it is necessary to consider other aspects, such as climate change, loss of biodiversity, or ozone layer depletion.

The empirical comparison of data co-occurrence for the two parameters (i.e. paragraph and group of 5 words) is far too similar. The densities observed in both maps only give prominence to 'energy' (Figure 4a). This result was expected, in part, due to the high presence of inflectional forms of words related to this specific word. However, when we add this word to an exclusion list the picture obtained is rather different as other densities emerge. This seems to be the case for 'climat' and 'chang', 'environment' and 'sustain', or 'urban', 'transport' and 'demand'. It is also interesting to note the presence of a density related to industry, biomass and fossil fuels. Here we noticed that replacing fossil fuels by sustainable-produced biomass is seen has a safe method to reduce CO_2 emissions to the atmosphere and therefore the negative impact on the environment (Gustavsson, 1995; Forsberg, 2002).

Let us now consider the results of data co-occurrence based on TF*IDF distribution in somewhat more detail. The density view of both maps is meaningful because it shows the presence of some terms identified previously and related to nuclear, fossil, and renewable energy sources (e.g. 'oil', 'coal', 'wind', 'solar', 'nuclear', 'thermal'). The difference is that now we just consider terms with high explanatory power and try to depict the information using distinct parameters of linguistic discourse. Here, it is visible the presence of terms that express concern about the subject in analysis, such as 'sensit' or 'vulnerab', or even terms that call for adaption, such as 'adapt'. This is even more evident for densities that combine 'household', 'gdp', and 'employ' to illustrate some of the possible effects emanating from climate change. In a way, the combination of these terms is understandable, especially if we consider that low income households tend to live in areas with low GDP growth, high unemployment rates and therefore more likely to be affected by climate change, and have far less ability to move or make the necessary adjustments to their living conditions.

After conducting the co-occurrence analysis in our corpus, we propose two sub-themes to structure data delivered by TPGs in the field of 'Energy & Environment'. The sub-theme structure and inheritance is the following:

04 ENERGY AND ENVIRONMENT

0401 Energy and resources (e.g. renewable, nuclear, and fossil energies)

0402 Environmental facets of climate change (e.g. GHG emissions, air pollution)

The first sub-theme is intended to include data on energy and resources, including renewable, nuclear, and fossil energies. The second sub-theme should incorporate distinct features or elements that actively contribute to climate change (e.g. air and soil pollution, biodiversity loss, water management, greenhouse gas emissions).

4.2.5. Land Use

Land use refers how the earth's surface is used, including the location, type and design of human development. As a result, land use patterns have diverse economic, social and environmental impacts. In the previous ESPON Database, 'Land Use' is defined as a first-level theme but its inheritance is somehow vague in the description[3].

The collection of documents used for analysis is not very substantial. Nevertheless, we have managed to gather some documents based on research activities conducted by EU institutions or commissioned to universities and research institutions on behalf of the funding entity (see Appendix 10). The visualization of similarities of terms extracted from such context provided a better understanding of data co-occurrence. In Figure 5a we can see the pattern of similarities between terms with high frequency distribution for documents related to 'Land Use'. Clearly, 'urban', 'model', 'land', and 'area' are among those. However, the knowledge obtained from the exclusion of these terms conveys other relationships. This seems to be the case for 'chang', 'impact', 'environment' and 'develop'. The same applies for 'agricultur', 'produc', and 'cropland'. Less visible, but still important, is relationship between 'transport', 'sprawl', and 'energi'. The term 'scenario', itself very important in land use discussions, appears completely isolated from the other main densities. Overall, the maps presented in Figure 5b disclose relevant information for analysis.

Among the visible interactions established by our corpus in the field of 'Land Use', it is possible to identify a sequence of terms that correspond to changes in land-use for both rural and urban settings. This facet seems to be evident for terms like 'chang', 'impact', 'rural' and 'urban'. In fact, most of the impacts related to land-use have an effect in rural and urban contexts. This also holds true for socio and economic factors. In addition to these, the term 'scenario' also discloses some significance. The ability to forecast land-use scenarios is essential to better understand dynamic processes which are determined by a range of driving forces, including demographic, socio-economic, and environmental change.

If we consider the strong emphasis given to these terms and its similarity, one could assume that the focus of our corpus is oriented to land-use changes, impacts and scenarios. As a consequence, we propose the following structure of sub-themes on 'Land Use' data:

05	LAND USE
0501	Land use and land cover types (e.g. CORINE Land Cover, GMES)
0502	Urban land use attributes and changes (e.g. LUZ, Urban Atlas)
0503	Rural land use attributes and changes (e.g. Natura 2000)

The first sub-theme has not been defined with text mining tools. However, its purpose is to integrate data related to CORINE Land Cover (CLC), Natura 2000, and the Urban Atlas Initiative. The other two sub-themes derive from a qualitative description of term similarity maps within a corpus of documents in the field of 'Land Use' and are meant to integrate data on changes, including indicators and indices on land use changes and impacts.

[3] Two sub-themes have been defined to structure data in the field of 'Land Use'. These are: '111 Natural resources' and 'Land use' (ESPON, 2005).

4.2.6. Social Affairs

The theme related to 'Social Affairs' is meant to cover data on social, economic and cultural issues with an emphasis on employment, labour market, income, living conditions, and poverty. Our collection of documents is based on ESPON evidence and results and, alternately, findings from other sources used by TPGs while undertaking research in this sort of topics. In total, we collected 12 documents from different sources, ranging from studies, policy notes and technical reports (see Appendix 11).

Unfortunately, in this case, the density maps of data co-occurrence do not add any relevant information. This means that the maps on both parameters are dominated by terms with little explanatory power and therefore terms with the highest frequency distribution should be added to an exclusion list. This seems to be the example for 'polici' and 'social' (see Figure 6a). However, if we remove these terms from our analysis the density map will depict other similarities offering a better understanding of how inflectional forms of words are associated with each other. Here, a special mention should be made for 'labour' and 'employ', econom' and 'develop', and 'indic' and 'cultur' (see Figure 6b).

The capacity of visualising similarities in a multidimensional scaling dramatically increases with the TF*IDF scoring algorithm. As it can be seen in Figure 6c and 6d, the knowledge structure is relatively easy to comprehend. Somehow, these maps suggest that TF*IDF measure reinforces the importance of terms less visible within the corpus. Besides, it clearly differentiates the major similarities. For instance, 'cultur' and 'heritag' reveal a distinct similarity. The same applies for 'labour' and 'job', 'household' and 'health', or 'incom' and 'famili'. Equally relevant is the presence of 'poverti'. Most of these similarities underline the rationale behind the theme in the field of 'Social Affairs' and, as a consequence, facilitate the definition of sub-themes.

The thematic structure designed for the ESPON 2006 DB suggests that data on employment and labour market should be disconnected from social exclusion (e.g. poverty) (ESPON, 2007). However, the results of this experiment indicate the opposite meaning that both should integrated and, if possible, include data on similar issues, such as living conditions and health systems.

06	SOCIAL AFFAIRS
0601	Education (e.g. training, lifelong learning)
0602	Labour market (e.g. labour force, labour costs, economic inactivity, earnings)
0603	Living conditions (e.g. poverty, social exclusion, health systems)
0604	Culture (e.g. socio-cultural activities, cultural consumption)

The sub-theme on 'education is designed to integrate data on training and lifelong learning while 'labour market' is more focused on economic inactivity, average earnings, and productivity. Within the same structure we suggest a second-level theme related to 'living conditions'. Ideally, this sub-theme will serve the purpose of integrating data on poverty, social exclusion as well as other types of living conditions, including data on health systems. Finally, our experiment with text mining also suggests a strong emphasis on issues related with culture and heritage. With this respect, we propose a sub-theme to allocate data on socio-cultural activities, cultural consumption and participation.

4.2.7. Economy

The economic analysis of EU strengths and weaknesses is of great importance to understand the policy designed by the Lisbon Strategy. According to the European Council its goal was to make the EU 'the most competitive and dynamic knowledge-based economy in the world capable of sustainable economy growth with more and better jobs and greater social cohesion' (CEC, 2000).

Despite joint efforts to achieve these goals only a small number of actions could have been fully implemented. One of the recent drawbacks to justify the moderate success of this initiative is the serious economic crisis that hit Europe and its citizens. As a response to such event, the European Commission has launched the 'Europe 2020 Strategy' in order to re-adapt the Lisbon Strategy to new challenges (CEC, 2010).

Access to accurate data is therefore of crucial importance to comprehend, for instance, the role of R&D, innovation, and patents to boost competitiveness in Europe. With this regard, the ESPON 2013 Database has the opportunity to structure both statistical and geographical data related to these topics.

The analysis undertaken to extract terms and label sub-themes in the field of 'Economy' is based on a dataset of 10 documents that combine ESPON evidence and results (see Appendix 12). The visualisation of knowledge structure created by these documents in a VOSviewer environment is meant to capture the similarity degree of terms with both explanatory power and non-explanatory power. As is can be seen in Figure 7a and 7b, the keyword co-occurrence of inflectional words is quite obvious for 'develop' and 'econom'. The same applies for 'innov' and 'research', or 'fund' and 'structur'.

Surprisingly enough, the exclusion of terms with high frequency distribution (i.e. 'develop' and 'econom') does not demonstrate the expected impact on the density view of data co-occurrence and, therefore, its structure neither changes nor generate unseen similarities. In this sense, the major subgroups identified previously can still be seen in the VOSviewer maps of Figure 7b. It is particularly interesting to observe the strong correlation between 'servic', 'industri', and 'capit', the continuous emphasis on 'knowledge', 'innov', and research', or the association established between 'territorial' and 'structur'.

The application of the TD*IDF measure also plays a decisive role in this analysis. The VOS maps presented in Figure 7c depict the knowledge structure of data co-occurrence based on TF*IDF distribution and, Figure 7d, illustrates what would happen if we added the most significant TF*IDF terms to an exclusion list. In general, our results do not add more information than what we have so far exposed. Therefore, the relationship established by terms with high TF*IDF distribution did not extract unknown and potential information. The terms identified by the preceding experiment confirm this evidence. Hence, we propose the following sub-themes in the field of 'Economy':

07 ECONOMY

0701 Aggregated accounts (e.g. GDP, purchasing power parities, balance of payments)

0702 Employment (e.g. employment, unemployment, long-term unemployment)

0703 Production and costs per sector (e.g. production of manufactured goods)

0704 Research and innovation (e.g. R&D expenditure, ICT research, patents)

The knowledge structure of our corpus suggests the presence of four sub-themes. The first sub-theme is the less visible in our maps. Still, we decided to introduce it so that TPGs involved in ESPON projects may allocate data on GDP and its main components. The other three sub-themes emerged more implicitly due to the high similarity or relatedness of terms and its clear organisation in clusters. Therefore, the sub-theme on 'employment' should allocate data on employment and unemployment. The sub-theme related to 'production and costs per sector' is meant to incorporate data on production of manufactured goods. Finally, for 'research and innovation' TPGs are encouraged to integrate date on R&D expenditure, ICT research, patents, and public investments.

4.2.8. Cross-Thematic and Non-Thematic Data

This theme is meant to cover both cross-thematic and non-thematic data. A first sub-theme should include variables that mix themes (e.g. integrative indicators; complex typologies; trends and impacts on both CAP and TEN-T; scenario-based projections on urban development; environmental, social, and economic concerns associated with lad-use changes). In addition, it should integrate indicators and indices aimed at evaluating the sensitivity and vulnerability impacts. The second subset refers to base maps (i.e. administrative units) and other geographical objects (e.g. grids, cities, networks) or spatial delineations (e.g. morphological zones, functional areas). Its current structure looks as follows:

99	CROSS-THEMATIC AND NON-THEMATIC DATA
9901	Integrative indices, indicators and scenarios (e.g. typologies, scenarios)
9999	Geographical objects (e.g. administrative units, grids, networks)

We should stress that this proposal is open to more sub-themes. This assumption is also valid for the themes described previously. However, its expansion should depend on datasets delivered by future ESPON projects.

5. Conclusions and future work

In this technical report, we have proposed a two-step approach to structure the ESPON 2013 DB by themes and sub-themes. First, we assume that database structures constitute an important resource of information. To a certain extent, this knowledge reflects the structure of the previous database and will certainly influence the current developments. For this purpose, we apply a visual grouping technique to illustrate, by means of correlation matrices, homogeneous clusters of words. The findings of our experiment constitute the basis to derive a first set of themes and eventually facilitate data allocation.

In the second part of the report we define sub-themes. Here, we believe the demand from the ESPON 2013 DB end-users will correspond to immediate, easy and practical access to datasets. In order to meet this request, we explore the potentialities offered by text mining methods and tools. The rationale of this approach is to find patterns across textual data and generate simple overviews of large text collections.

Our collection of data is based on ESPON evidence and for each theme we demonstrate that is possible to shed some light on ways to further progress in this field. The definition of sub-themes has been data-driven. This assumption has greatly beneficiated from mapping techniques of multidimensional scaling to ease the interpretation of relational similarities that came out from data co-occurrence with both explanatory power and non-explanatory power. The identification of these patterns suggests that the ESPON 2013 DB should be structured in 27 sub-themes unveiling the inheritance of 7+1 themes meaning that we would add a last theme to cover cross-thematic and non-thematic data (see Appendix 13).

Ultimately, it is necessary to allocate data into themes and sub-themes. For this purpose, we have considered data from of the ESPON 2006 Programme and data delivered up to date, i.e. mid-September 2010. During the course of this analysis we also suggest a potential second theme and sub-theme that somehow could improve classification and data retrieval. If some doubts subsist in our evaluation we propose other words to describe data. Hopefully this rather inductive process will rationalize the ability to restrict a search query when looking for specific datasets and consequently allow end-users to achieve greater level of precision and recall.

Our findings provide a new approach to data structure but also point out to considerable future work, of both empirical and conceptual nature. At the empirical level, it is clear that we need to refine our understanding of what is being measured to better allocate each variable to a specific theme and sub-theme. The quality of metadata is of course crucial in that regard. But, perhaps more fundamentally, there are some open questions at the conceptual level. Here, future work should validate the usability of this method. Taking into consideration that ESPON will launch new calls for proposals it is important to understand what kind of data future projects will deliver. If there's no appropriate theme and sub-theme, new analysis should be performed on text collections to improve the quality of the ESPON 2013 DB thematic structure. Hence, some of the difficulties that might emerge in the future should be further investigated by means of text mining methods.

References

Berry, M. [ed.] (2004). Survey of text mining. Clustering, classification, and retrieval. Springer: New York.

Blanchard, A. (2007) Understanding and customizing stopword lists for enhanced patent mapping. World Patent Information, 29 (1), 308-316.

CEC (2001) Environment 2010: Our future, our choice. Communication from the Commission to the Council. Commission of the European Communities: Brussels.

CEC (2010) Europe 2020. A European Strategy for Smart, Sustainable and Inclusive Growth. Commission of the European Communities.

Davi, A.; Haughton, D.; Nasr, N. et al (2005). A review of two text mining packages: SAS TextMining and WordStat. *American Statistician*, 59(1), 89-103.

Eck, N. & L. Waltman (2007). VOS: a new method for visualizing similarities between objects. In H.-J. Lenz and R. Decker [ed.]. Advances in Data analysis: Proceedings of the 30th Annual conference of the German Classification Society. Studies in Classification, Data Analysis, and Knowledge Organization. Springer: Heidelberg, 299-306.

Eck, N. & L. Waltman (2009). A computer program for bibliometric mapping. In B. Larsen and J. Leta [eds.]. Proceedings of the XII International Conference on Scientometrics and Informetrics, 886-897.

Eck, N. et al (2010) A comparison of two techniques for bibliographic mapping: Multidimensional scaling and VOS. *Journal of the American Society for Information Science and Technology*, 61(12), 2405-2416.

EEA (2002) Energy and environment in the European Union. Environmental Issue Report No 31. European Environment Agency: Copenhagen.

ESPON (2005). Integrated Tools for European Spatial Development. ESPON 3.1 Project. Luxembourg: ESPON Coordination Unit, pp. 141-174.

ESPON (2006) Feasibility study on monitoring territorial monitoring territorial development based on ESPON key indicators. ESPON Project 4.1.3 Final Report. BBR: Bonn.

European Council (2000) Presidency conclusions. Lisbon European Council, 23-24 March 2000. European Council: Brussels (available for download on www.consilium.europa.eu).

Forsberg, C. (2008) Sustainability by combining nuclear, fossil, and renewable energy sources. *Progress in Nuclear Energy*, 1-9.

Gustavsson, L. et al (1995) Reducing CO_s emissions by substituting biomass for fossil fuels. Energy 20(11), 1097-1113.

Lewis, B. (1999). Simstat with Wordstat: A Comprehensive Statistical Package with a Content Analysis Module. *Field Methods*, 11(2), 166-179.

Weiss, S.; Indurkhya, N.; Zhang, T. and F. Damerau (2005). Text Mining. Predictive Methods for Analysing Unstructured Information. Springer: New York.

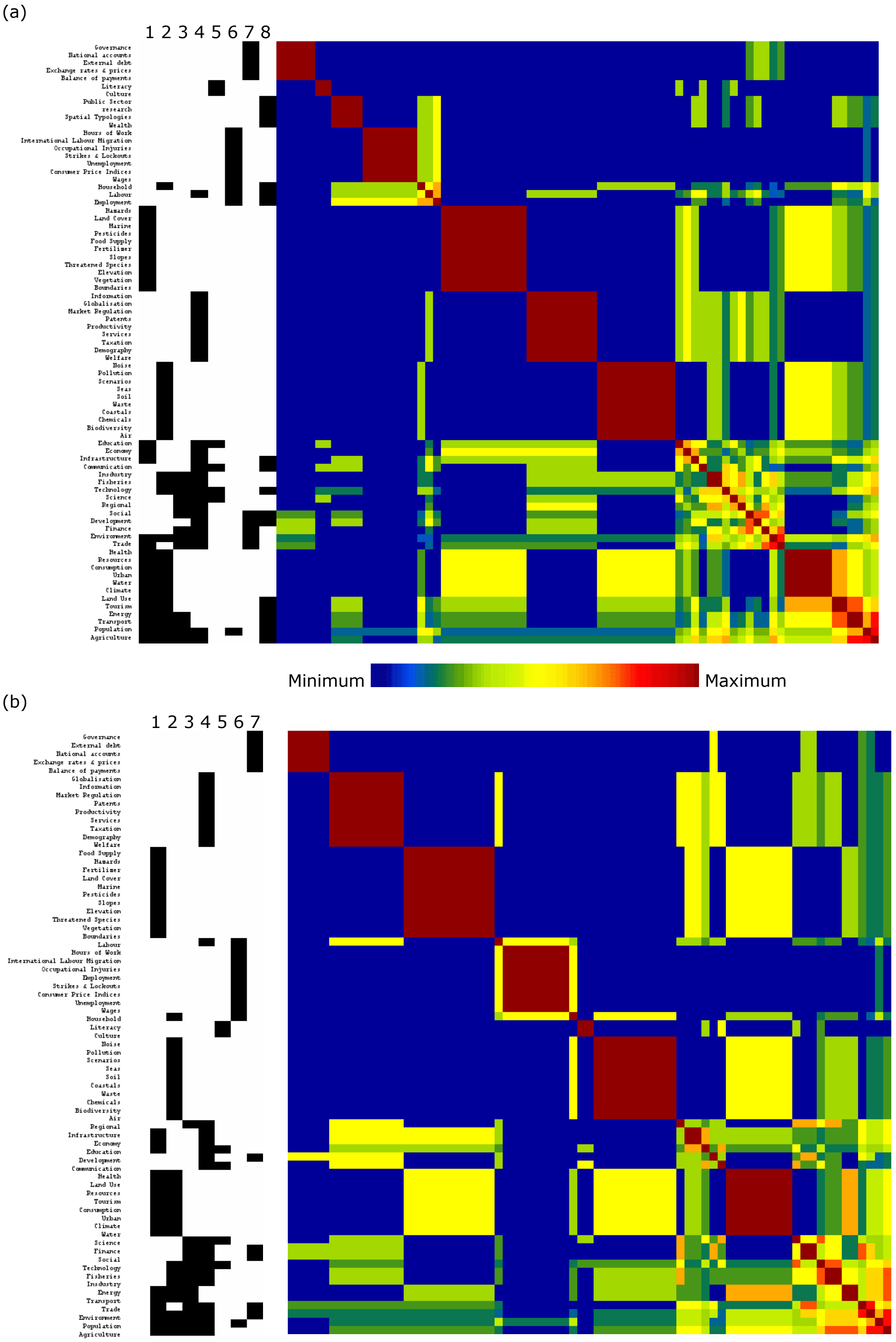
Chen, C.-H. (2002). Generalized Association Plots: Information Visualization via Iteratively Generated Correlation Matrices. *Statistica Sinica*, 12(1), 7-29.

Wu, H.-M., et al. (2008). GAP: A graphical environment for matrix visualization and cluster analysis. *Computational Statistics & Analysis*, in press.

Yang, Y.; Akers, L.; Klose, T. and Yang, C. (2008) Text mining and visualization tools. Impressions of emerging capabilities. *World Patent Information*, 30 (1), 280-293.

Appendix 1

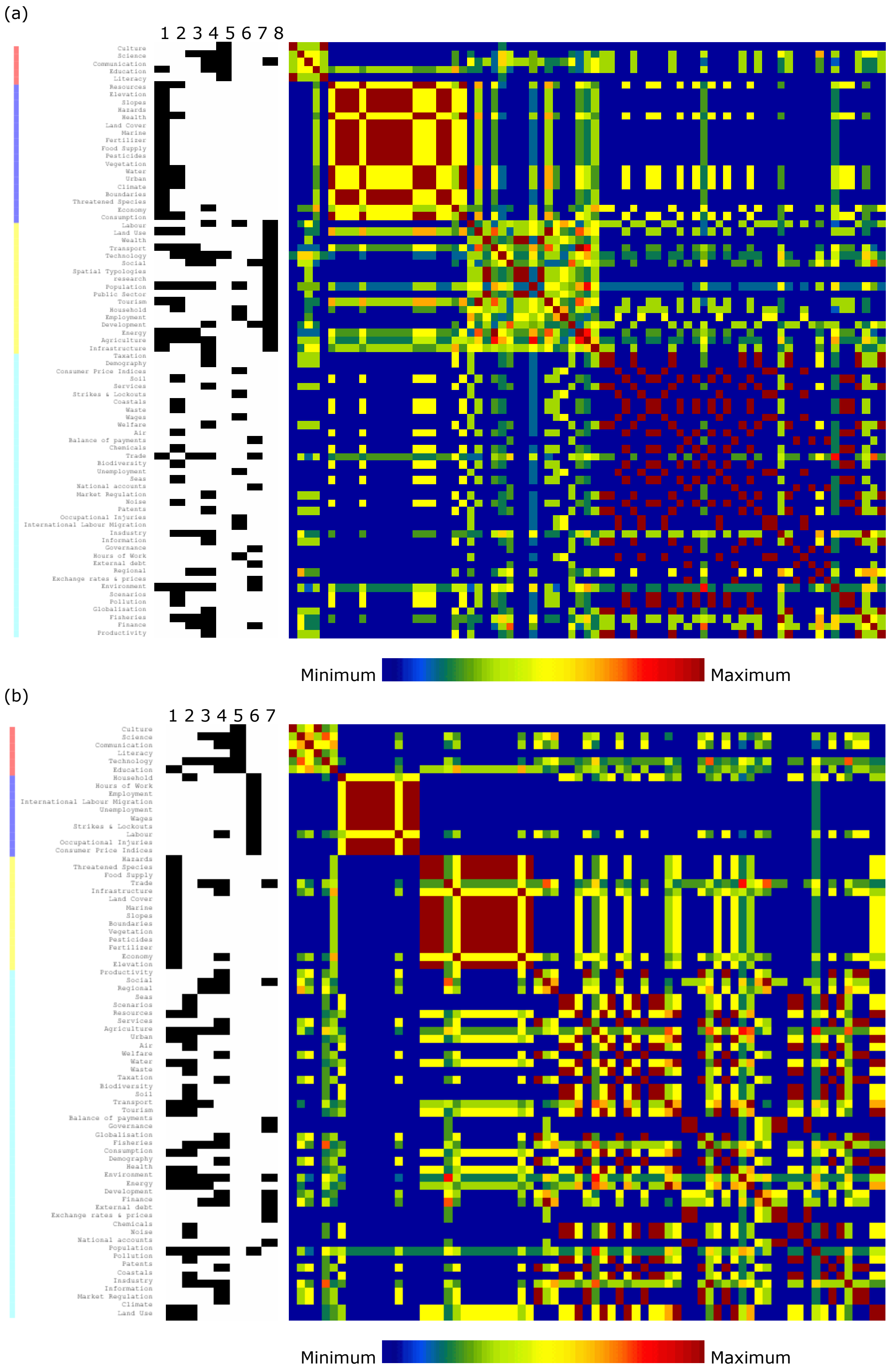
Matrix visualisation in GAP environment of words used by statistical databases to label themes. Matrix (a) includes the structure of the ESPON 2006 DB while (b) ignores it



Note: (1) UNEP, (2) EEA, (3) EUROSTAT, (4) OECD, (5), UNESCO, (6) ILO, (7) WPI, (8) ESPON 2006.

Appendix 2

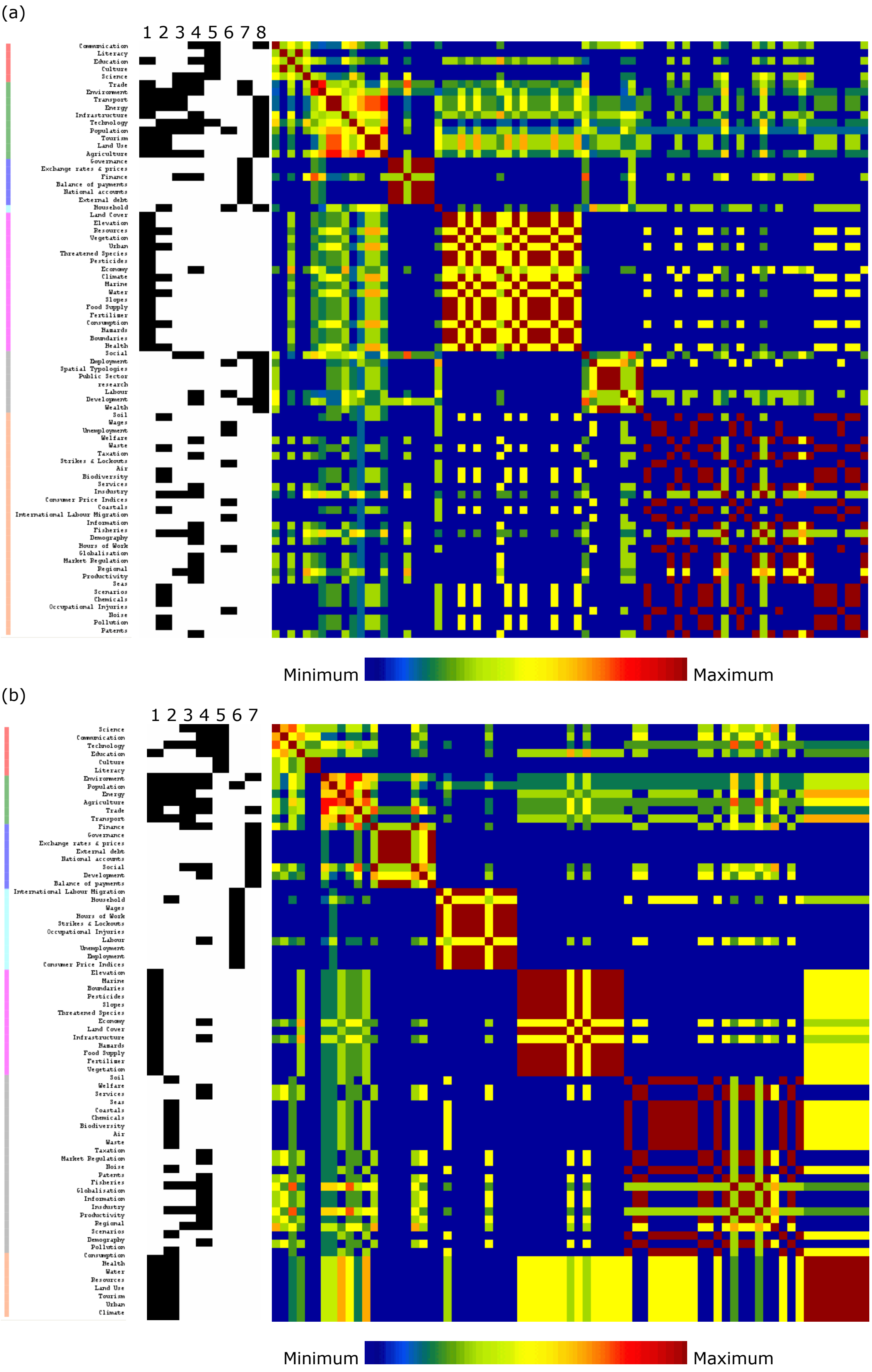
Cluster analysis based on Jaccard’s coefficient in GAP environment. Matrix (a) includes the structure of the ESPON 2006 DB while (b) ignores it



Note: (1) UNEP, (2) EEA, (3) EUROSTAT, (4) OECD, (5), UNESCO, (6) ILO, (7) WPI, (8) ESPON 2006.

Appendix 3

Cluster analysis based on Jaccard's coefficient in GAP environment. Matrix (a) includes the structure of the ESPON 2006 DB while (b) ignores it



Note: (1) UNEP, (2) EEA, (3) EUROSTAT, (4) OECD, (5), UNESCO, (6) ILO, (7) WPI, (8) ESPON 2006.

Appendix 4
Database structures ordered by themes

UNESCO

1. Education
2. Science & Technology
3. Culture & Communication
4. Literacy

ILO

1. Economically Active Population
2. Employment
3. Unemployment
4. Hours of Work
5. Wages
6. Labour Cost
7. Consumer Price Indices
8. Occupational Injuries
9. Strikes and Lockouts
10. Household Income and Expenditure
11. International Labour Migration

EUROSTAT

1. General and Regional Statistics
2. Economy and Finance
3. Population and Social Conditions
4. Industry, Trade and Fisheries
5. External Trade
6. Transport
7. Environment and Energy
8. Science and Technology

OECD

1. General Statistics
2. Agriculture and Fisheries
3. Demography and Population
4. Development
5. Economic Projections
6. Education and Training
7. Environment
8. Finance
9. Globalisation
10. Health
11. Industry and Services
12. Information and Communication Technology
13. International Trade and Balance of Payments
14. Labour
15. Monthly Economic Indicators
16. National Accounts
17. Prices and Purchasing Power Parities
18. Productivity
19. Public Sector, Taxation and Market Regulation
20. Regional Statistics
21. Science, Technology and Patents
22. Social and Welfare Statistics

EEA

1. Agriculture
2. Air
3. Biodiversity Change
4. Chemicals
5. Climate Change
6. Coastals and Seas
7. Energy
8. Environmental Scenarios
9. Fisheries
10. Households
11. Human Health
12. Industry
13. Natural Resources
14. Noise
15. Policy Analysis
16. Population and Economy
17. Regions
18. Soil
19. Tourism
20. Transport
21. Urban Environment
22. Waste
23. Water

UNEP

1. Agricultural Production
2. Boundaries
3. Climate
4. Economy
5. Education
6. Elevation and Slopes
7. Emissions of GHG and ODS
8. Energy Consumption and Production
9. Environmental Hazards
10. Fertilizer & Pesticides
11. Food Supply & Caloric Intake
12. Health
13. Infrastructure
14. Land Use
15. Marine and Coastal Areas
16. Population
17. Private Consumption
18. Protected Areas and Environmental Protection
19. Technological Hazards
20. Total and Threatened Species
21. Tourism
22. Trade Balances
23. Transport
24. Urbanisation
25. Vegetation and Land Cover
26. Water Consumption and resources

WORLD BANK

1. Agriculture
2. Aid
3. Childhood Development
4. Debt
5. Education
6. Environment
7. Finance
8. Gross Domestic Production
9. Gender
10. Globalisation
11. Governance
12. Health
13. Information Technology
14. Infrastructure
15. Industry
16. Labour & Employment
17. Macroeconomics & Growth
18. Population
19. Poverty
20. Purchasing Power Parity
21. Private Sector
22. Public Sector
23. Rural Development
24. Social Development
25. Trade
26. Urban Development

Appendix 5

Words employed to label themes in database structures

Words (or expressions)	UNEP	EEA	EUROSTAT	OECD	UNESCO	ILO	WPI	ESPON 2006
Agriculture	🟩	🟩	🟩	🟩			🟩	🟩
Aid							🟩	
Air		🟩						
Balance of payments							🟩	
Biodiversity		🟩						
Boundaries	🟩							
Chemicals		🟩						
Childhood							🟩	
Climate	🟩	🟩						
Coastals		🟩						
Communication				🟩	🟩			🟩
Consumer Price Indices						🟩		
Consumption	🟩	🟩						
Culture					🟩			
Demography				🟩				
Development				🟩			🟩	🟩
Economy	🟩			🟩				
Education	🟩			🟩	🟩		🟩	
Elevation	🟩							
Employment						🟩	🟩	🟩
Energy	🟩	🟩	🟩					🟩
Environment	🟩	🟩	🟩	🟩			🟩	🟩
Exchange rates & prices							🟩	
External debt							🟩	
Fertilizer	🟩							
Finance			🟩	🟩			🟩	
Fisheries		🟩	🟩	🟩				
Food Supply	🟩							
GDP							🟩	
Gender							🟩	
Globalisation				🟩			🟩	
Governance							🟩	
Hazards	🟩							
Health	🟩	🟩					🟩	
Hours of Work						🟩		
Household		🟩				🟩		🟩
Information				🟩			🟩	
Infrastructure	🟩			🟩			🟩	🟩
Insdustry		🟩	🟩	🟩			🟩	
International Labour Migration						🟩		
Labour				🟩		🟩	🟩	🟩
Land Cover	🟩							
Land Use	🟩	🟩						🟩
Literacy					🟩			
Macroeconomics							🟩	
Marine	🟩							
Market Regulation				🟩				
National accounts							🟩	
Noise		🟩						
Occupational Injuries						🟩		
Patents				🟩				
Pesticides	🟩							
Pollution		🟩						
Population	🟩	🟩	🟩	🟩		🟩	🟩	🟩
Poverty							🟩	
PPP							🟩	
Productivity				🟩				
Public Sector							🟩	🟩
Regional			🟩	🟩				
Research								🟩
Resources	🟩	🟩						
Rural							🟩	
Scenarios		🟩						
Science			🟩	🟩	🟩			
Seas		🟩						
Services				🟩				
Slopes	🟩							
Social			🟩	🟩			🟩	🟩
Soil		🟩						
Spatial Typologies								🟩
Strikes & Lockouts						🟩		
Taxation				🟩				
Technology		🟩	🟩	🟩	🟩		🟩	🟩
Threatened Species	🟩							
Tourism	🟩	🟩						🟩
Trade	🟩		🟩	🟩			🟩	
Transport	🟩	🟩	🟩					🟩
Unemployment						🟩		
Urban	🟩	🟩					🟩	
Vegetation	🟩							
Wages						🟩		
Waste		🟩						
Water	🟩	🟩						
Wealth								🟩
Welfare				🟩				

Appendix 6

Reports taken into account for text mining purposes on 'Agriculture and Fisheries'

Code	Report
report#01	European Parliament (2007) Regional Dependency on Fisheries. European Parliament: Brussels.
report#02	European Parliament (2007) Reflection on the possibilities for the future development of the CAP. European Parliament: Brussels.
report#03	ESPON (2005) The territorial impact of CAP and Rural Development Policy. ESPON Project 2.1.3 Final Report. Arkleton Institute: Aberdeen.
report#04	ESPON (2005) Territorial Impacts of European Fisheries Policy. ESPON Project 2.1.5 Final Report. NIBR: Oslo.
report#05	ESPON (2010) European Development Opportunities for Rural Areas. EDORA Draft Final Report. UHI Millenium Institute: Inverness.
report#06	ESPON (2010) Territorial Impact for Transport and Agricultural Policies. TIP TAP Final Report A/B. Politecnico di Milano: Milan.
report#07	ESPON (2010) Territorial Impact for Transport and Agricultural Policies. TIP TAP Final Report C. Politecnico di Milano: Milan.

Appendix 7

Reports taken into account for text mining purposes on ‘Demography’

Code	Report
report#08	European Parliament (1999) Regional development in less-densely populated regions in the EU. European Parliament: Brussels.
report#09	ESPON (2005) The spatial effects of demographic trends and migration. ESPON Project 1.1.4 Final Report. ITPS: Stockholm.
report#10	ESPON (2010) Demographic and migratory flows affecting European regions and cities. DEMIFER Draft Final Report. NIDI: The Hague.
report#11	CEC (2006) The Demographic Future of Europe: From Challenge to Opportunity. Commission of the European Communities: Brussels.
report#12	CCE (2007) Europe’s Demographic Future: Facts and Figures. Commission of the European Communities: Brussels.

Appendix 8

Reports taken into account for text mining purposes on 'Transport'

Code	Report
report#13	European Parliament (2006) The Impact of Trans-European Networks on Cohesion and Employment. European Parliament: Brussels.
report#14	ESPON (2004) Transport services and networks: Territorial trends and basic supply on Infrastructure for Territorial Cohesion. ESPON Project 1.2.1 Final Report. University of Tours: Tours.
report#15	ESPON (2005) Territorial Impact of EU Transport and TEN policies. ESPON Project 2.1.1 Final Report. Spiekermann & Wegener: Dortmund.
report#16	ESPON (2009) Territorial Impact package for Transport and Agricultural Policies. TIPTAP Draft Final Report. Politecnico de Milano: Milan.
report#17	ESPON (2007) Update of selected potential accessibility indicators. Final report. Spiekermann & Wegener: Dortmund.
report#18	CEC (2007) Trans-European Networks: Towards an integrated approach. Commission of the European Communities: Brussels.

Appendix 9

Reports taken into account for text mining purposes on 'Energy and Environment'

Code	Report
report#19	ESPON (2005) Territorial Trends and Policy Impacts in the field of EU Environmental Policy. ESPON 2.4.1 Final Report. Geological Survey of Finland: Helsinki.
report#20	ESPON (2005) Territorial Trends of Energy Services and Networks and Territorial Impact of EU Energy Policy. ESPON 2.1.4 Final Report. CEEETA: Lisbon.
report#21	ESPON (2010) Climate Change and Territorial Effects on Regions and Local Economies. ESPON CLIMATE Interim Report. TU Dortmund: Dortmund.
report#22	ESPON (2010) Regions at Risk of Energy Poverty. ReRisk Draft Final Report. Inasmet-Tecnalia: Donostia/San Sebastian.
report#23	European Parliament (2007) Using Sustainable and Renewable Energies in the context of Structural Policy 2007-2013. European Parliament: Brussels.
report#24	European Parliament (2006) Energy and Structural and Cohesion Policies. European Parliament: Brussels.
report#25	European Parliament (1998) Sustainable Development: A Key Principle for European Regional Development. European Parliament: Brussels.
report#26	European Parliament (2003) The Enlargement Process of the EU: Consequences in the Field of Environment. European Parliament: Brussels.
report#27	European Parliament (2008) The Challenge of Climate Change for Structural and Cohesion Policies. European Parliament: Brussels.

Appendix 10

Reports taken into account for text mining purposes on ‘Land Use’

Code	Report
report#28	EEA (2006) Urban sprawl in Europe. The ignored challenge. EEA Report No 10/2006. European Environment Agency: Copenhagen.
report#29	EEA (2007) Land-use scenarios for Europe: qualitative and quantitative analysis on a European scale. EEA Technical Report No 9/2007. European Environment Agency: Copenhagen.
report#30	EEA (2009) Ensuring quality of life in Europe’s cities and towns. Tackling the environmental challenges driven by European and global change. EEA Report No 5/2009. European Environmental Agency: Copenhagen.
report#31	DG Environment (2008) Modelling of EU Land-use choices and environmental impacts. Scoping study. BIO Intelligence Services: Ivry-sur-Seine.

Appendix 11

Reports taken into account for text mining purposes on 'Social Affairs'

Code	Report
report#32	ESPON (2006) The role and spatial effects of cultural heritage and identity. ESPON Project 1.3.3 Final Report. Universita' degli Studi Ca' Foscari: Venice.
report#33	ESPON (2006) Territorial dimension of the Lisbon-Gothenburg strategy. ESPON Project 3.3 Final Report. Universita' degli Studi 'Tor Vergata' di Roma. CEIS: Rome.
report#34	European Parliament (2005) Adaption of Cohesion Policy to the enlarged Europe and the Lisbon and Gothenburg objectives. European Parliament: Brussels.
report#35	ESPON (2006) Preparatory study on social aspects of EU territorial development. ESPON Project 1.4.2 Final report. OIR: Vienna.
report#36	European Parliament (2007) Impact of Accession on the Labour Markets of the New Member States. European Parliament: Brussels.
report#37	European Parliament (2007) The role of minimum income for social inclusion in the European Union. European Parliament: Brussels.
report#38	European Parliament (2009) Indicators of Job Quality in the European Union. European Parliament: Brussels.
report#39	European Parliament (2010a) The link between job creation, innovation, education and training: An assessment of policies pursued at EU level. European Parliament: Brussels.
report#40	European Parliament (2010b) Structural and Cohesion Policies following the Treaty of Lisbon. European Parliament: Brussels.
report#41	European Parliament (2010c) Social Policy Agenda. Directorate-General for Internal Policies, European Parliament: Brussels.
report#42	European Parliament (2010d) Mobility and Integration of People with Disabilities into the Labour Market. European Parliament: Brussels.
report#43	European Parliament (2010e) EU Cooperation in the field of Social Inclusion. European parliament: Brussels.

Appendix 12

Reports taken into account for text mining purposes on 'Economy'

Code	Report
report#44	ESPON (2005) The Territorial Impact of EU Research and Development Policies. ESPON Project 2.1.2 Final Report. ECOTEC: Birmingham.
report#45	ESPON (2006) Territorial dimension of the Lisbon-Gothenburg strategy. ESPON Project 3.3 Final Report. Universita' degli Studi 'Tor Vergata' di Roma. CEIS: Rome.
report#46	ESPON (2007) Identification of spatially relevant aspects of information society. ESPON Project 1.2.3 Final Report. EUROREG: Warsaw.
report#47	ESPON (2006) Territorial impacts of EU economic policies and location of economic activities. ESPON Project 3.4.2 Final Report. IGEAT: Brussels.
report#48	CEC (2002) First progress report on economic and social cohesion. Commission of the European Communities: Brussels.
report#49	CEC (2003) Second progress report on economic and social cohesion. Commission of the European Communities: Brussels.
report#50	CEC (2005) Third progress report on cohesion: towards a new partnership for growth, jobs and cohesion. Commission of the European Communities: Brussels.
report#51	CEC (2006) The growth and jobs strategy and the reform of European cohesion policy. Fourth progress report on cohesion. Commission of the European Communities: Brussels.
report#52	CEC (2008) Fifth progress report on economic and social cohesion. Growing regions, growing Europe. Commission of the European Communities: Brussels.
report#53	CEC (2009) Sixth progress report on economic and social cohesion. Commission of the European Communities: Brussels.

Appendix 13

Overview of the ESPON 2013 DB thematic structure by themes and sub-themes

01 AGRICULTURE AND FISHERIES

0101 Farm structure (e.g. farm type, size of farms, income from farming, organic farming)

0102 Livestock (e.g. livestock output)

0103 Aquaculture and sea fisheries (e.g. aquaculture resources in coastal and marine areas)

0104 Forestry (e.g. production, consumption, import/export products)

0105 Rural characteristics (e.g. rural employment, rural access to services)

02 DEMOGRAPHY

0201 Population structure (e.g. age distribution by group and gender)

0202 Natural changes (e.g. fertility, mortality, life expectancy)

0203 Households (e.g. number and sizes of households)

0204 Migrations (e.g. immigration, migration replacement, high-skilled labour migration)

03 TRANSPORT

0301 Accessibility (e.g. performance indicators, multimodal accessibility)

0302 Flows (e.g. vehicles, passengers, goods, freight)

0303 Infrastructures (e.g. transportation systems, railways, airports, harbours)

04 ENERGY AND ENVIRONMENT

0401 Energy and resources (e.g. renewable, nuclear, and fossil energies)

0402 Climate change (e.g. GHG emissions, air pollution)

05 LAND USE

0501 Land use and land cover types (e.g. CORINE Land Cover, GMES)

0502 Urban land use attributes and changes (e.g. LUZ, Urban Atlas)

0503 Rural land use attributes and changes (e.g. Natura 2000)

06 SOCIAL AFFAIRS

0601 Education (e.g. training, lifelong learning)

0602 Labour market (e.g. labour force, labour costs, economic inactivity, earnings)

0603 Living conditions (e.g. poverty, social exclusion, health systems)

0604 Culture (e.g. socio-cultural activities, cultural consumption)

07 ECONOMY

0701 Aggregated accounts (e.g. GDP, balance of payments)

0702 Employment (e.g. employment, unemployment, long-term unemployment)

0703 Production and costs per sector (e.g. production of manufactured goods)

0704 Research and innovation (e.g. R&D expenditure, ICT research, patents, investments)

99 CROSS-THEMATIC AND NON-THEMATIC DATA

9901 Integrative indices, indicators and scenarios (e.g. typologies, scenarios)

9999 Geographical objects (e.g. administrative units, grids, networks)

Figure 1a: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a window of 5 words (right)

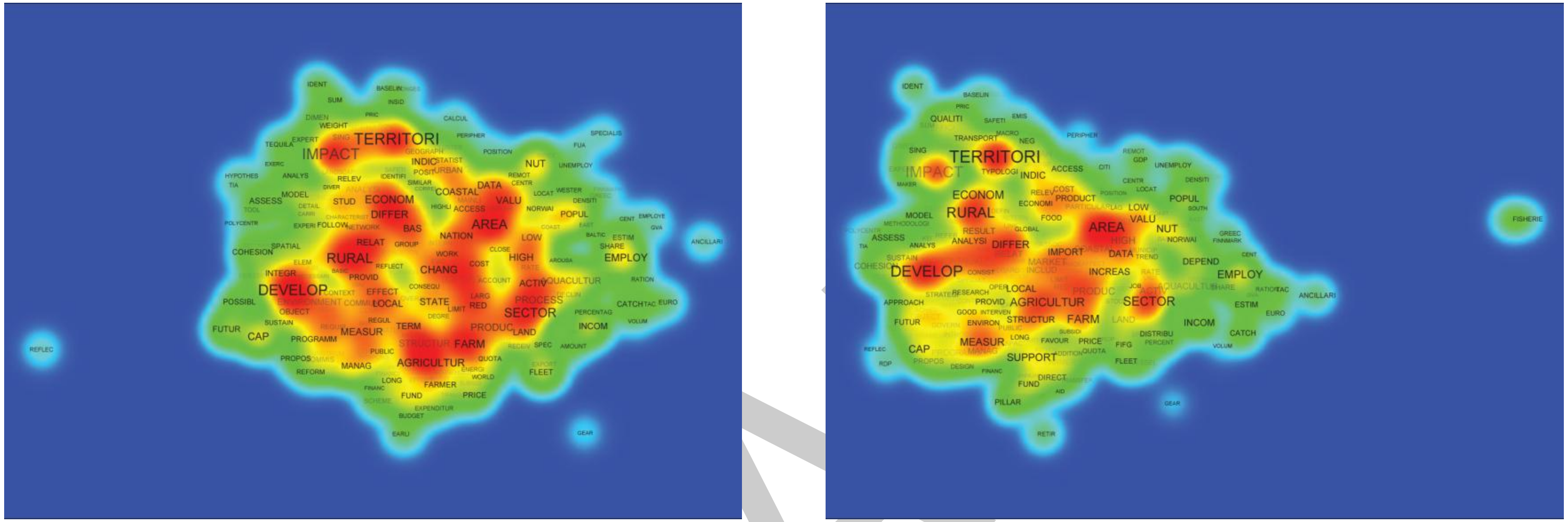
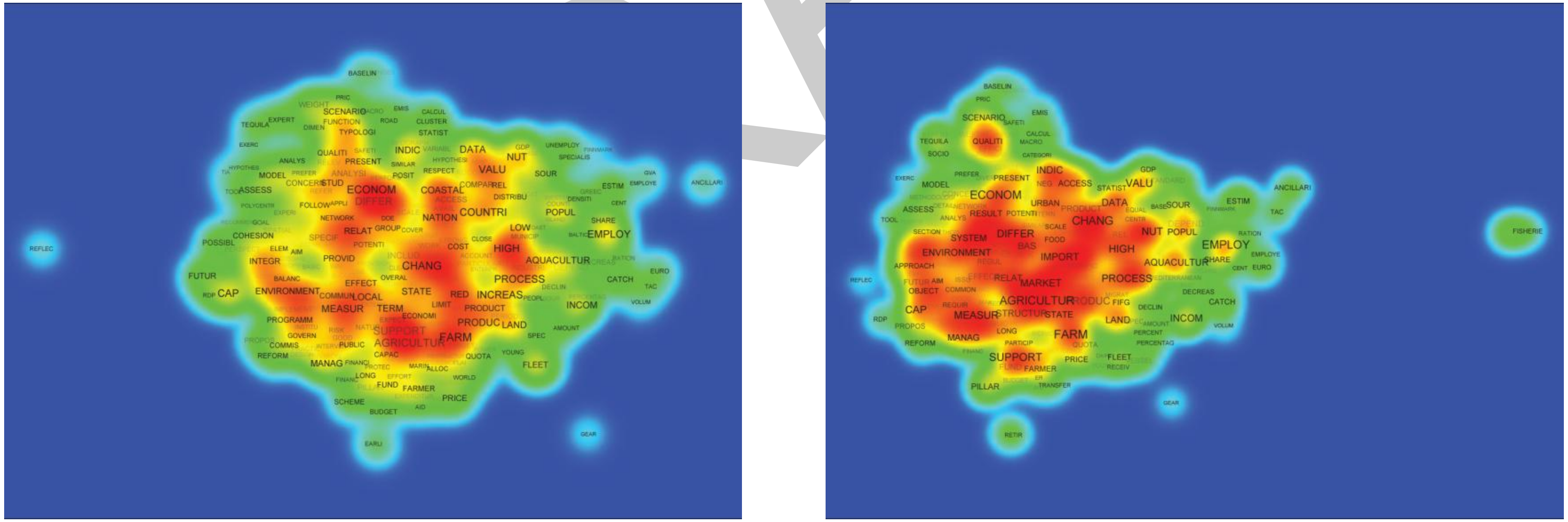


Figure 1b: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a window of 5 words (right)



Note: The words ‘territori’, ‘develop’, ‘impact’, ‘area’, ‘rural’, and sector’, present in (1a), were added to a exclusion list in (1b). TF: Term Frequency.

Figure 1c: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a window of 5 words (right)

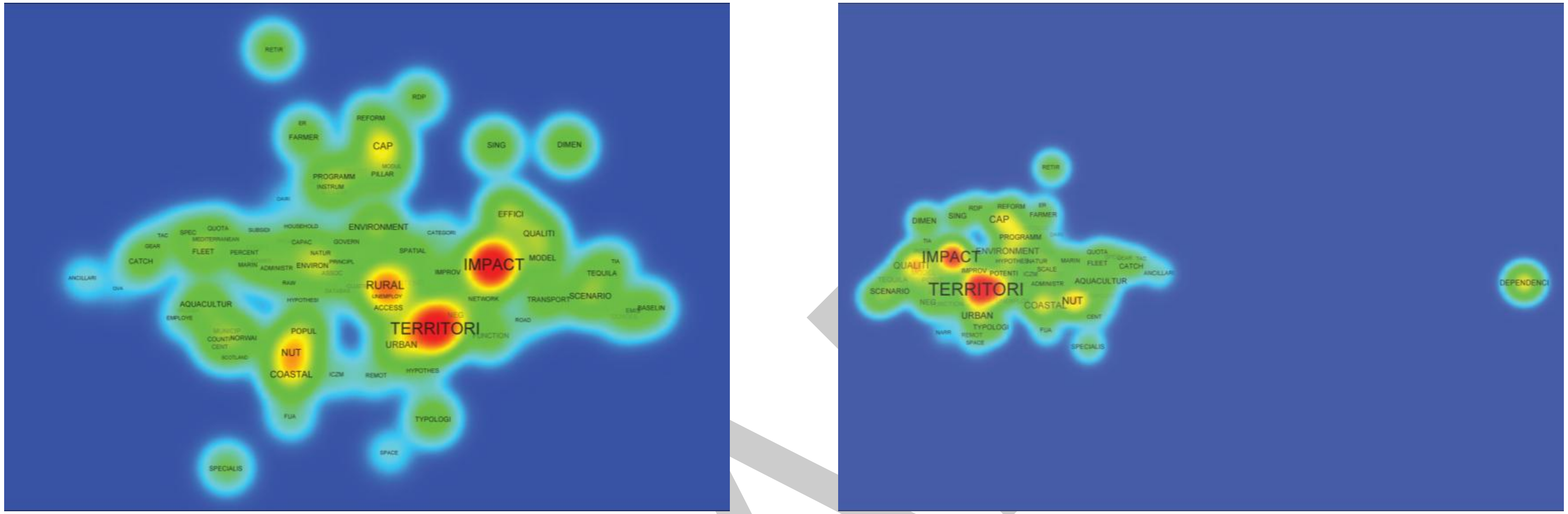
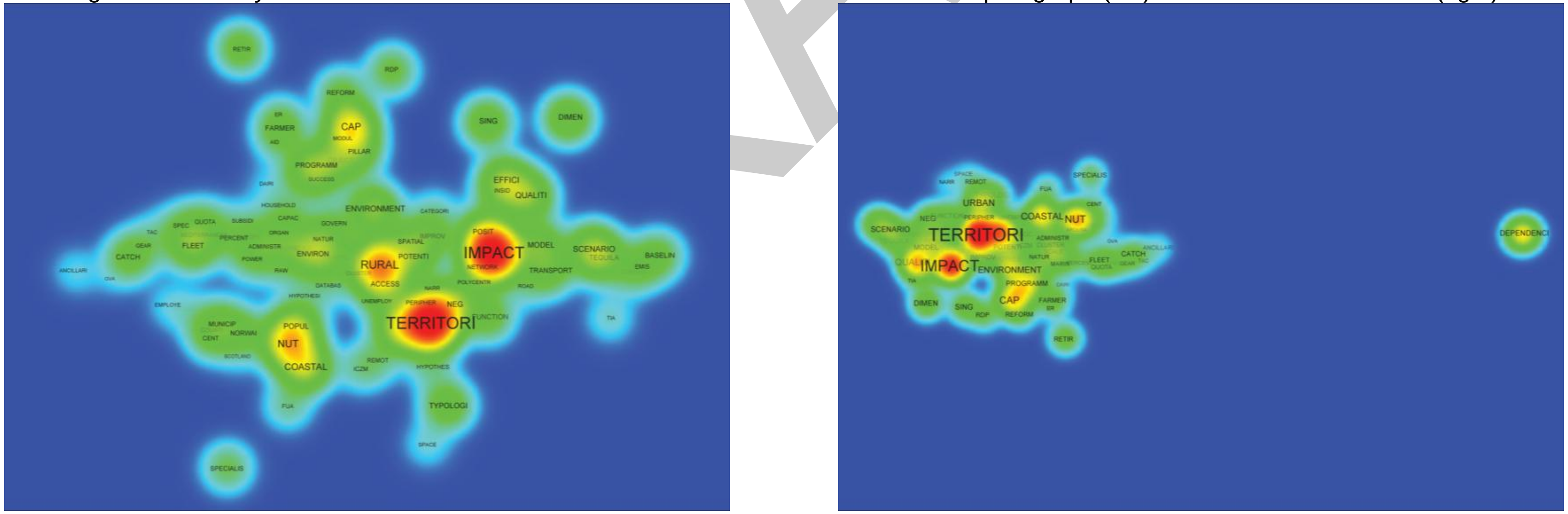


Figure 1d: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a window of 5 words (right)



Note: The word ‘acquacultur’, present in (2c), was added to a exclusion list in (2d)

Figure 2a: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a group of 5 words (right)

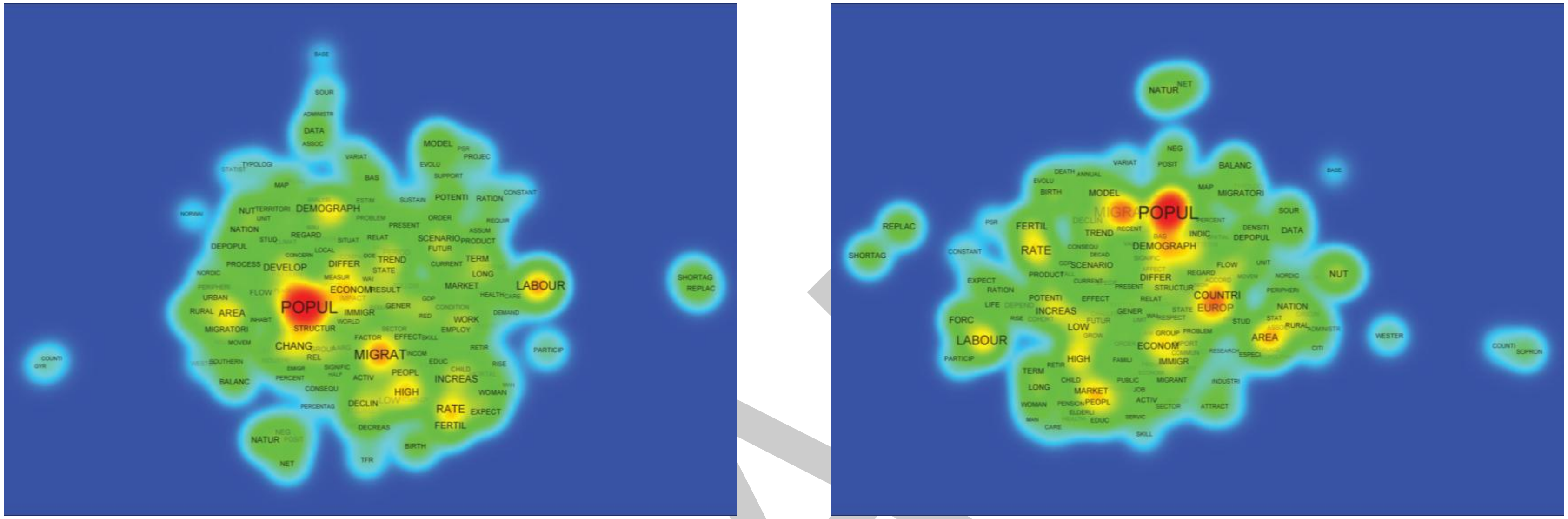
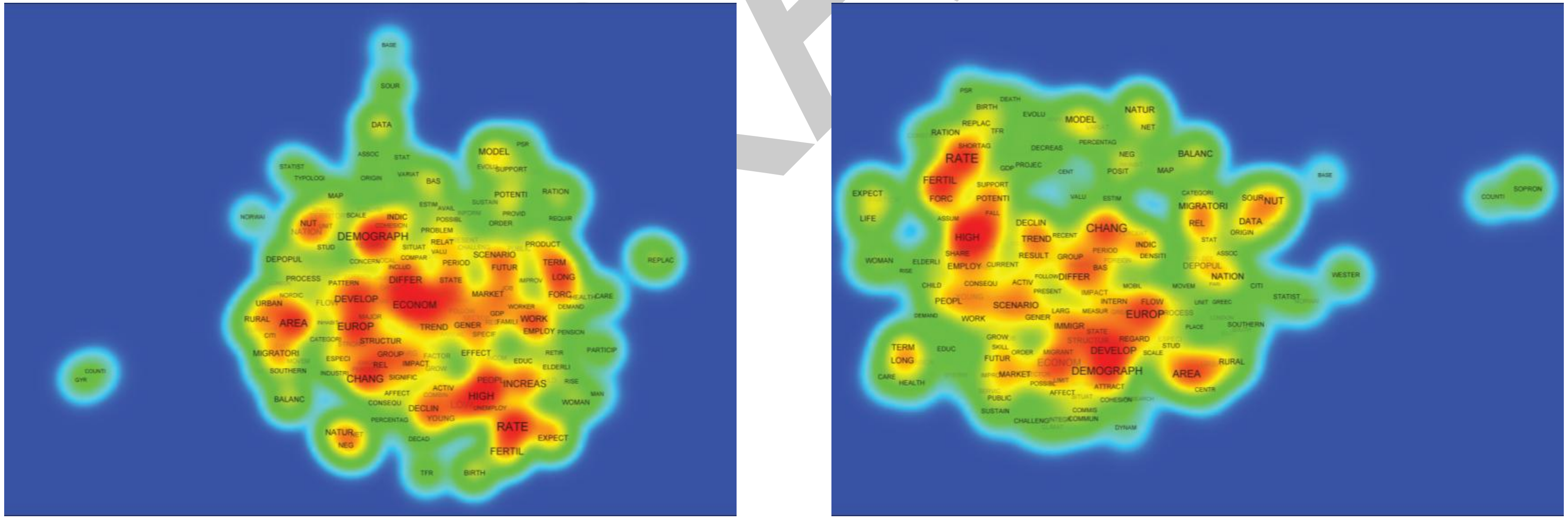


Figure 2b: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a window of 5 words (right)



Note: The words ‘popul’, ‘migrat’, ‘countri’, ‘labour’, present in (2a), were added to a exclusion list in (2b). TF: Term Frequency.

Figure 2c: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a group of 5 words (right)

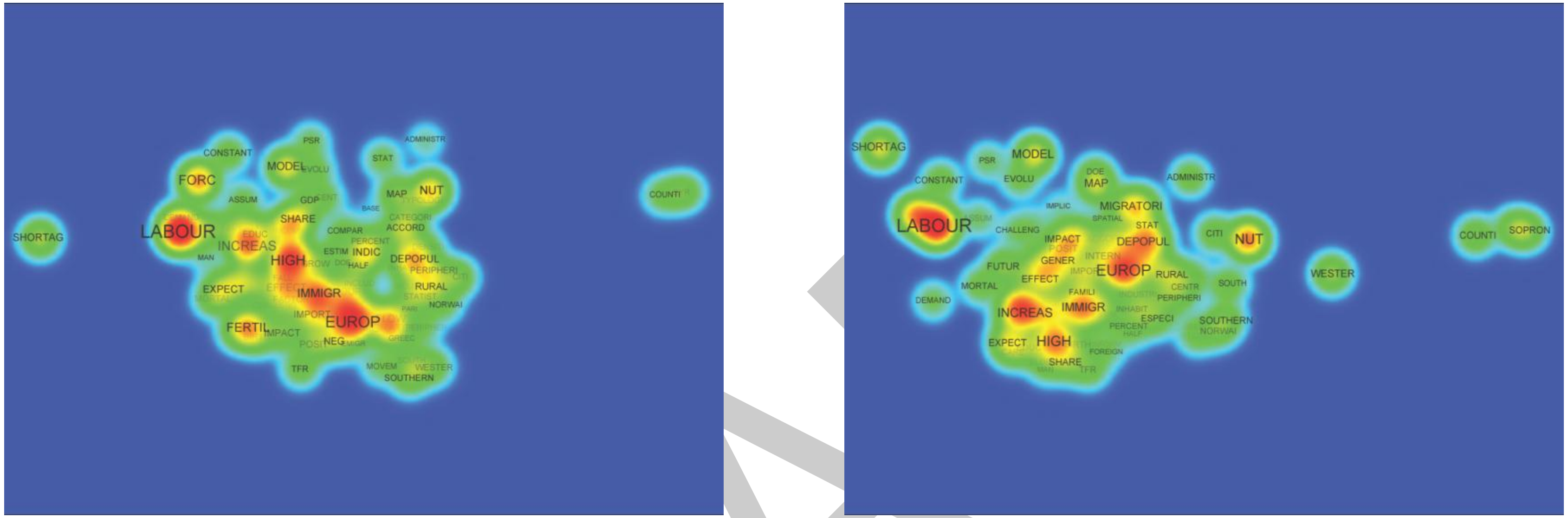
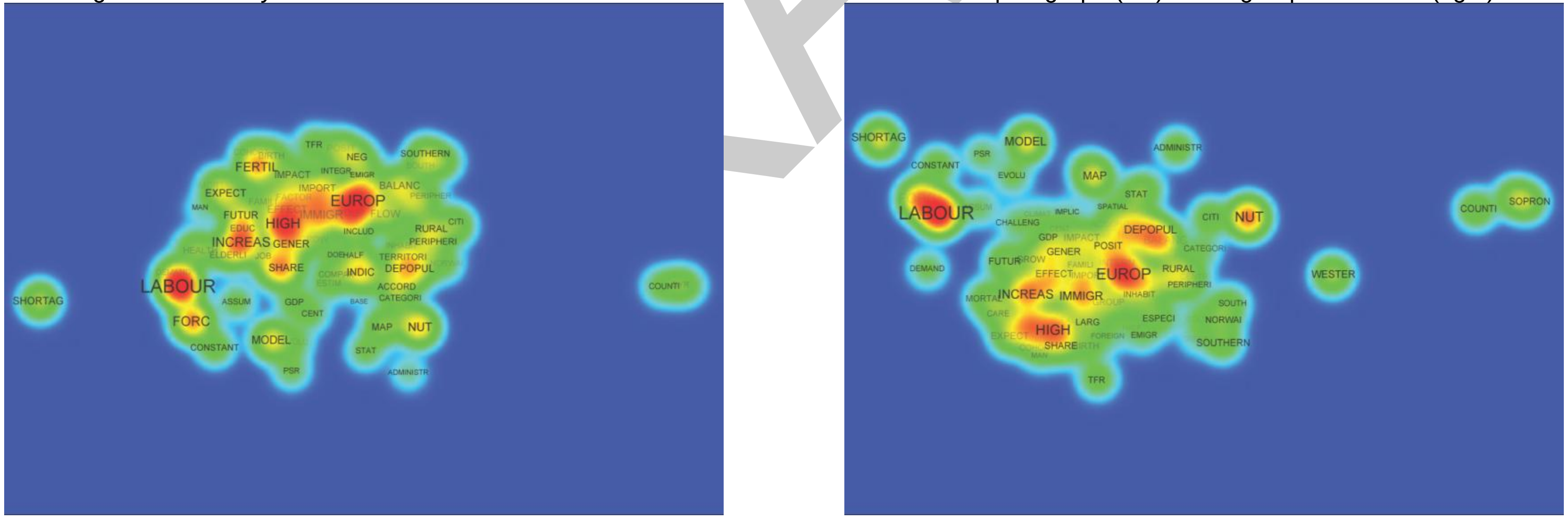


Figure 2d: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a group of 5 words (right)



Note: The word ‘migratori’, present in (2c), was added to a exclusion list in (2d)

Figure 3a: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a group of 5 words (right)

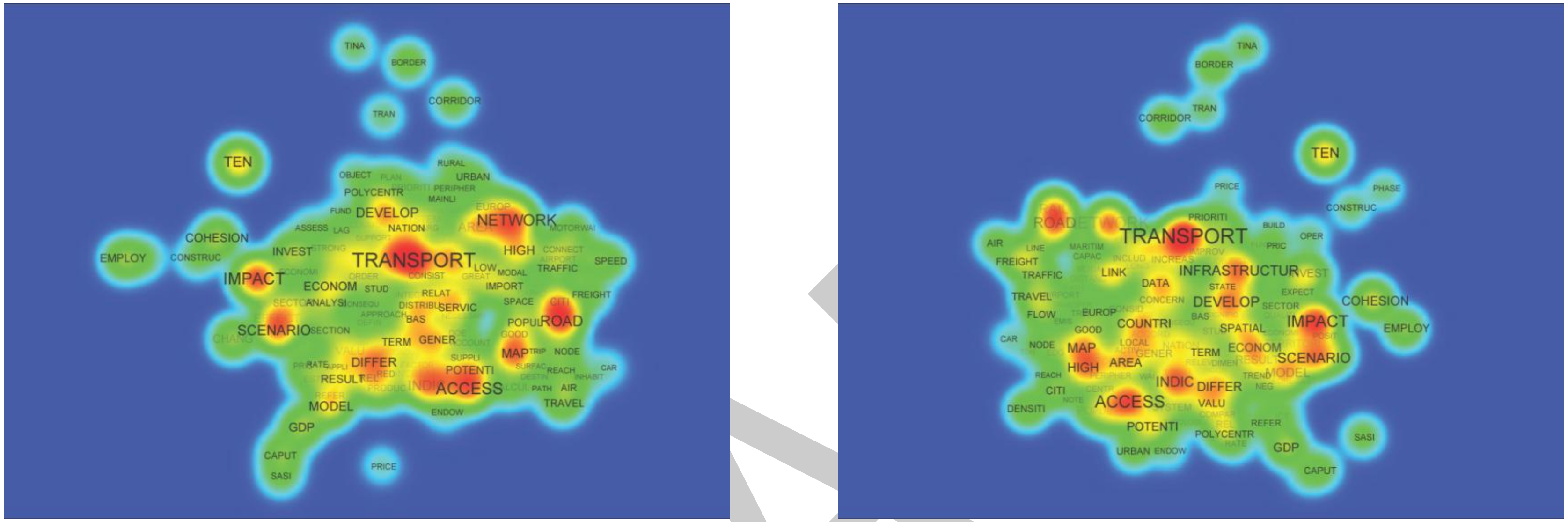
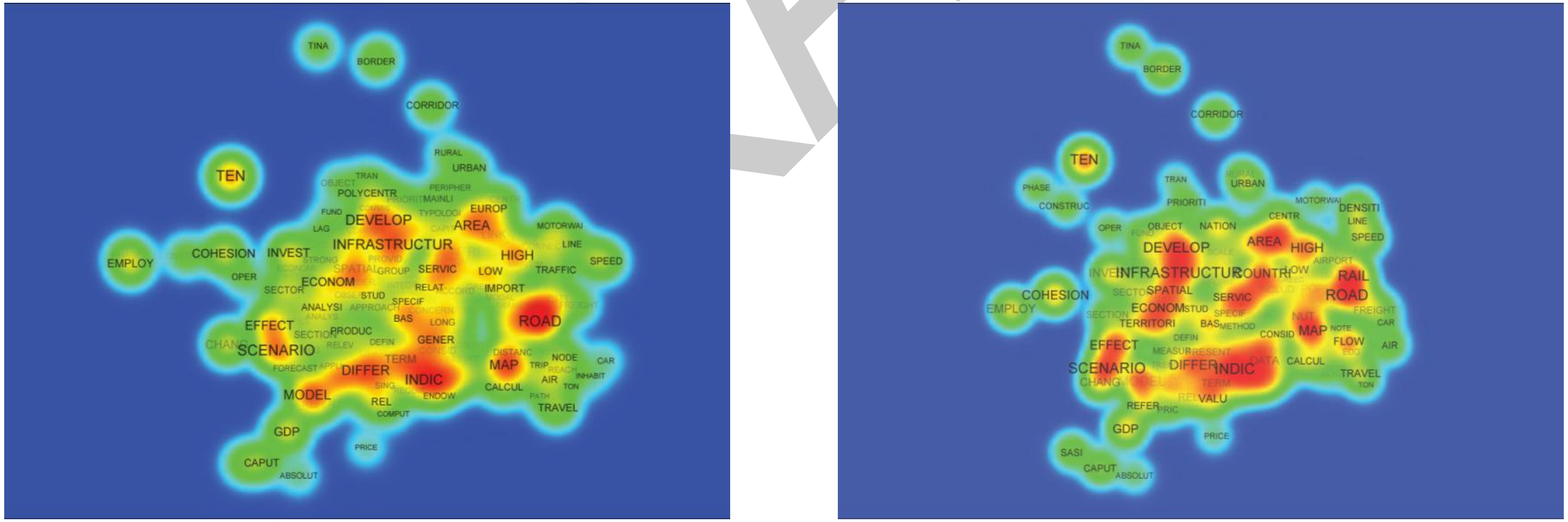


Figure 3b: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a window of 5 words (right)



Note: The words ‘transport’, ‘access’ ‘network’, ‘impact’, present in (3a), were added to a exclusion list in (3b). TF: term Frequency.

Figure 3c: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a group of 5 words (right)

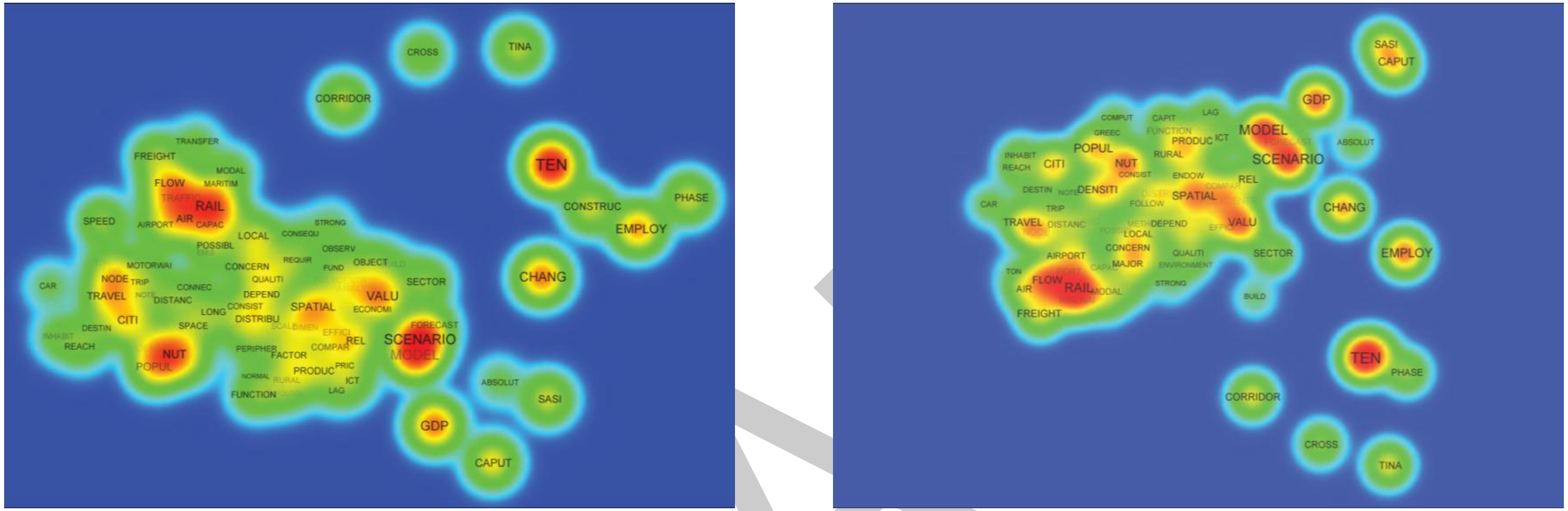
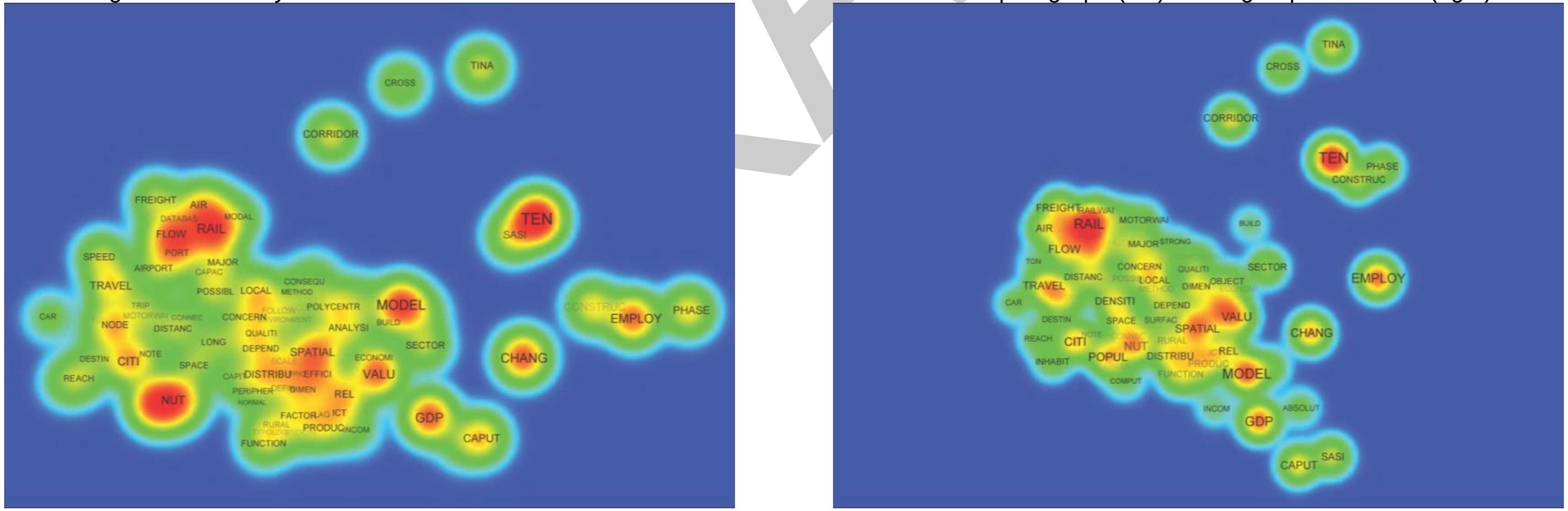


Figure 3d: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a group of 5 words (right)



Note: The word ‘scenario’, present in (3c), was added to a exclusion list in (3d)

Figure 4a: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a group of 5 words (right)

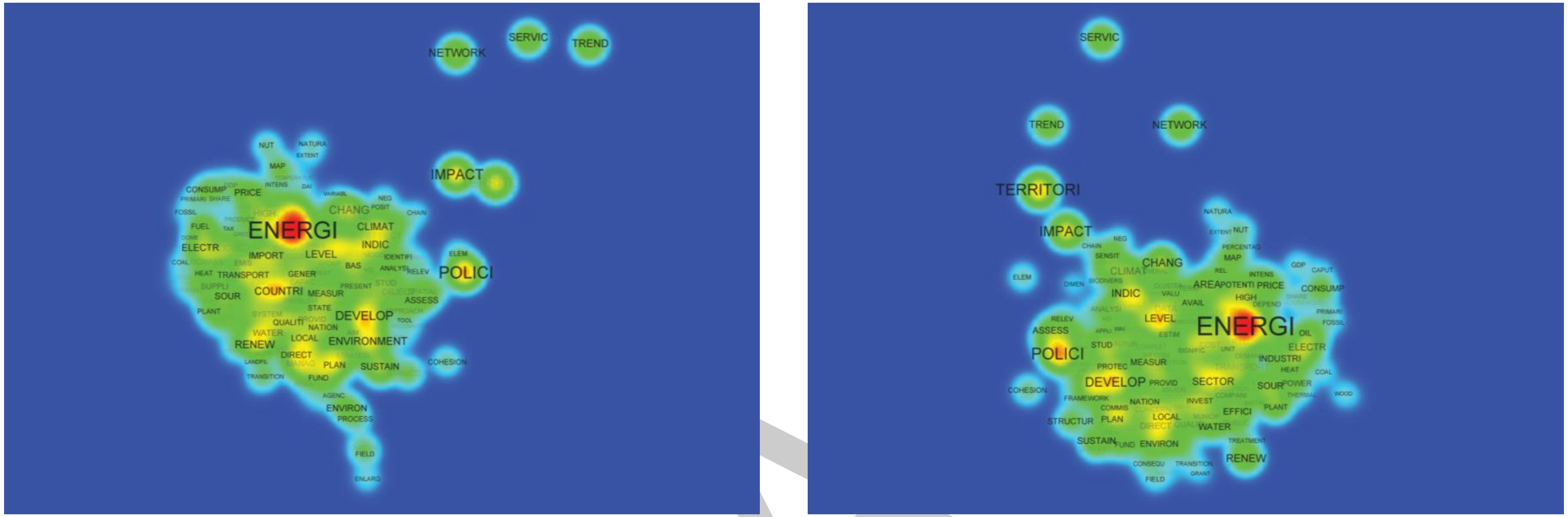
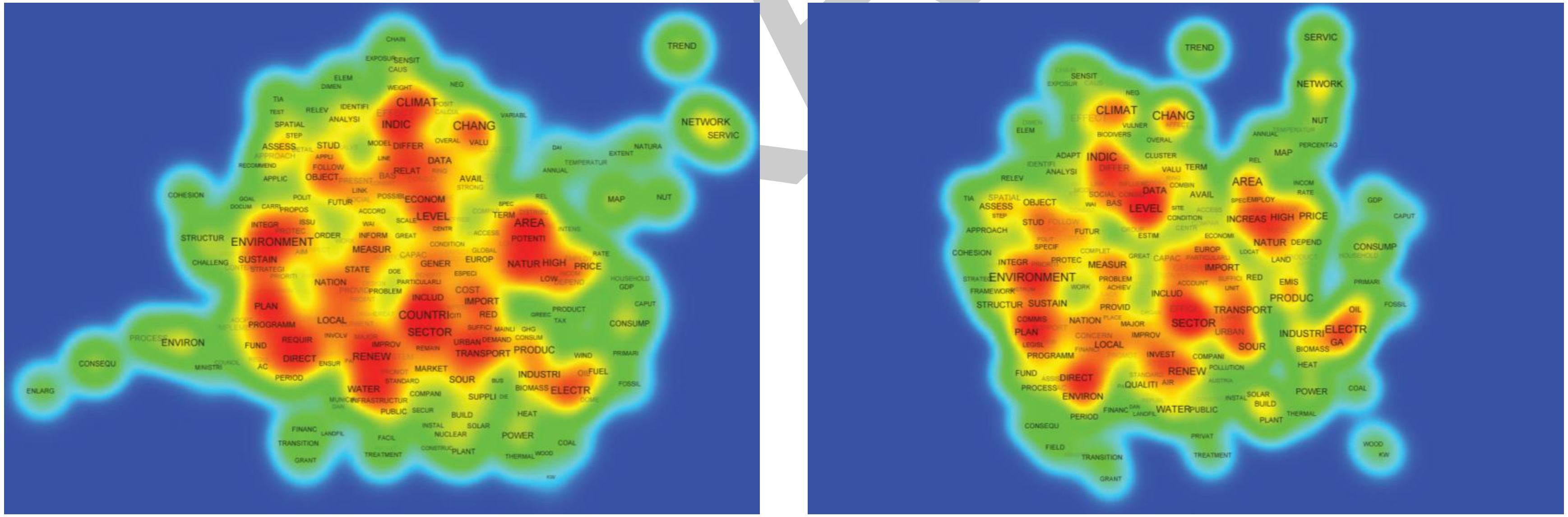


Figure 4b: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a window of 5 words (right)



Note: The words ‘energi’, ‘polici’, ‘territori’, ‘impact’, and ‘develop’, present in (4a), were added to a exclusion list in (4c). TF: Term Frequency.

Figure 4c: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a group of 5 words (right)

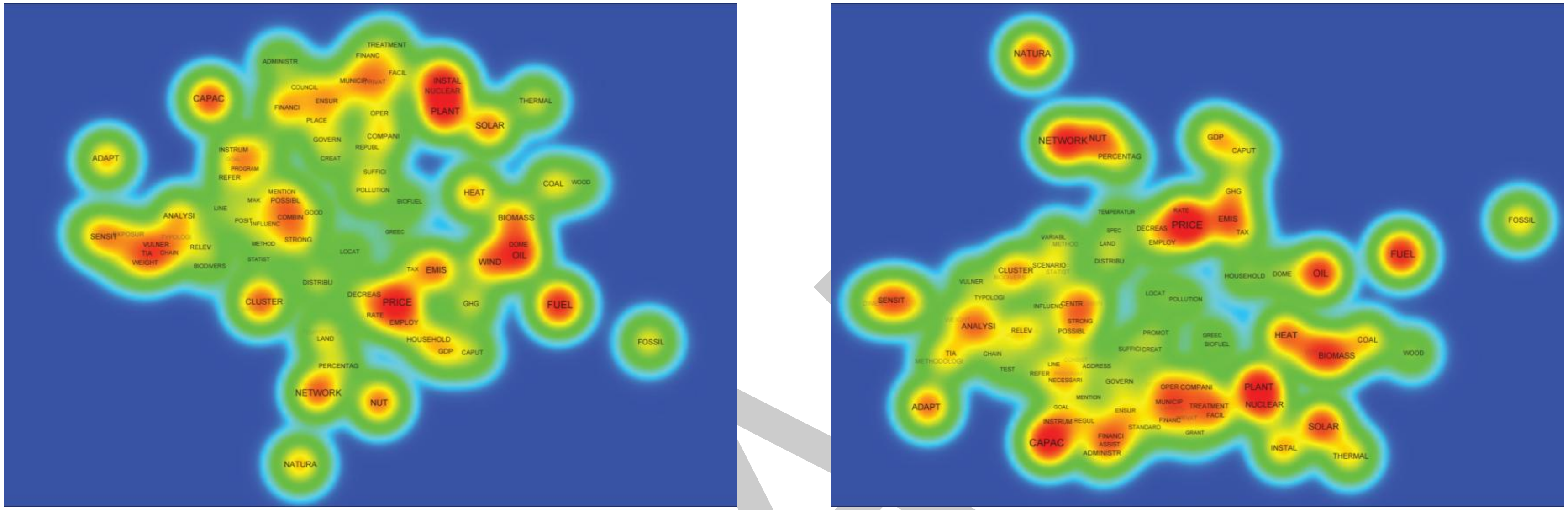
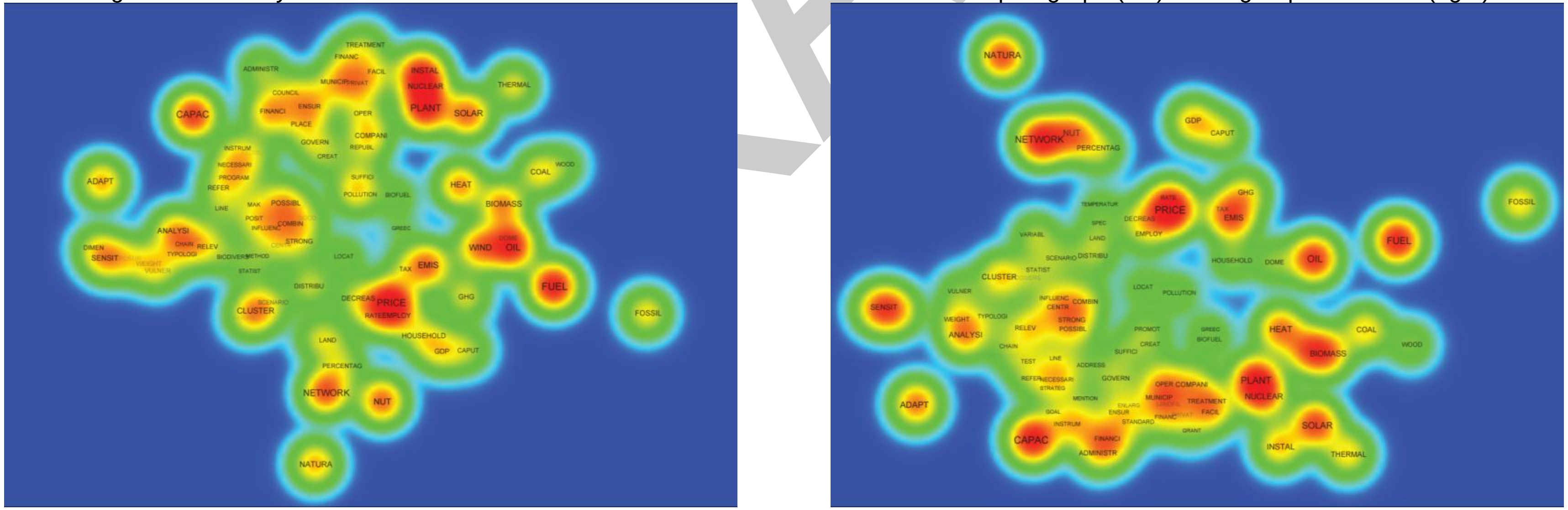


Figure 4d: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a group of 5 words (right)



Note: The word ‘tia’, present in (4c), was added to a exclusion list in (4d)

Figure 5a: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a window of 5 words (right)

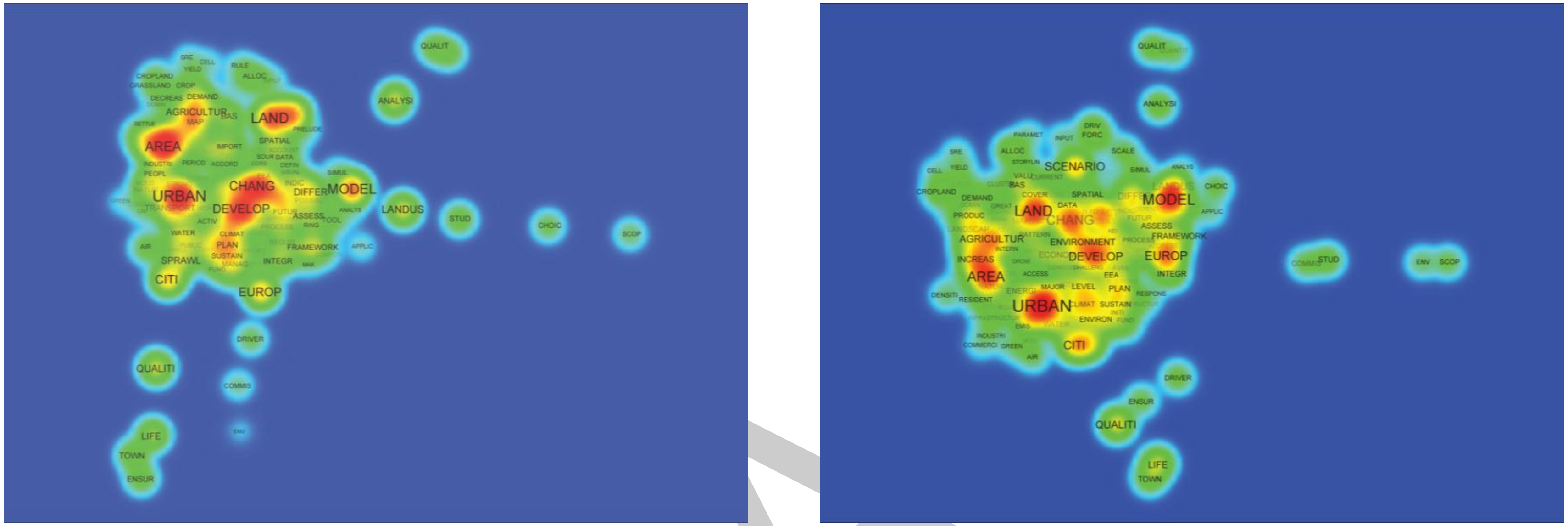
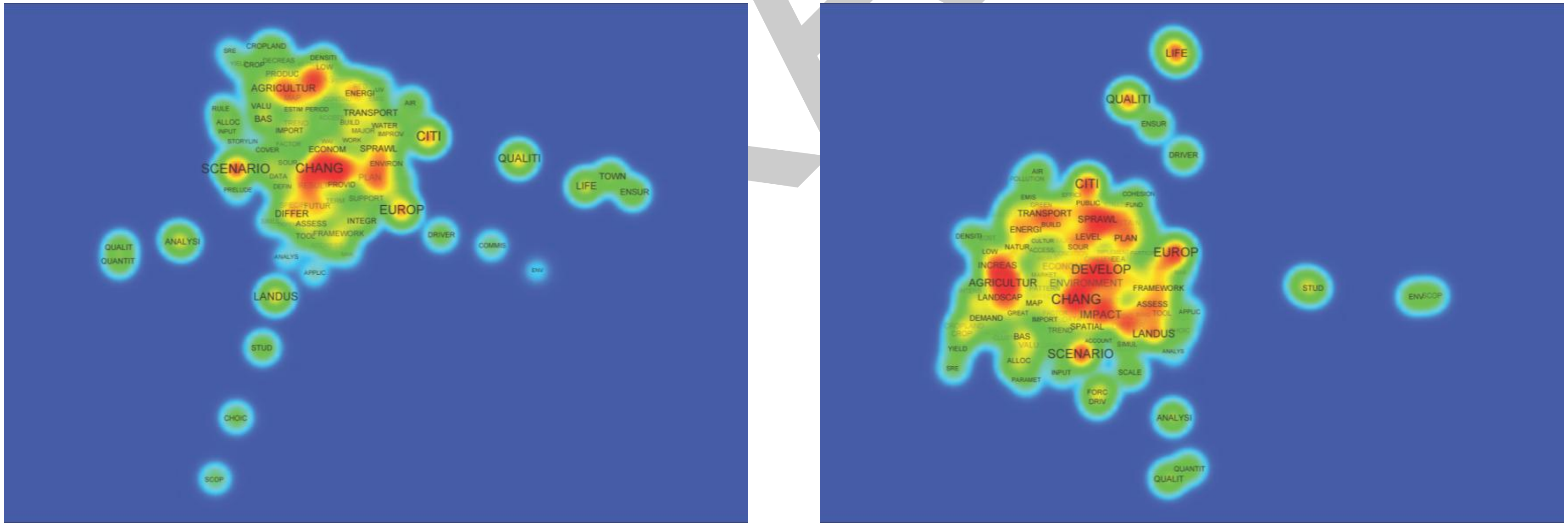


Figure 5b: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a window of 5 words (right)



Note: The words ‘urban’, ‘model’, ‘land’, and ‘area’, present in (5a), were added to a exclusion list in (5b)

Figure 5c: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a group of 5 words (right)

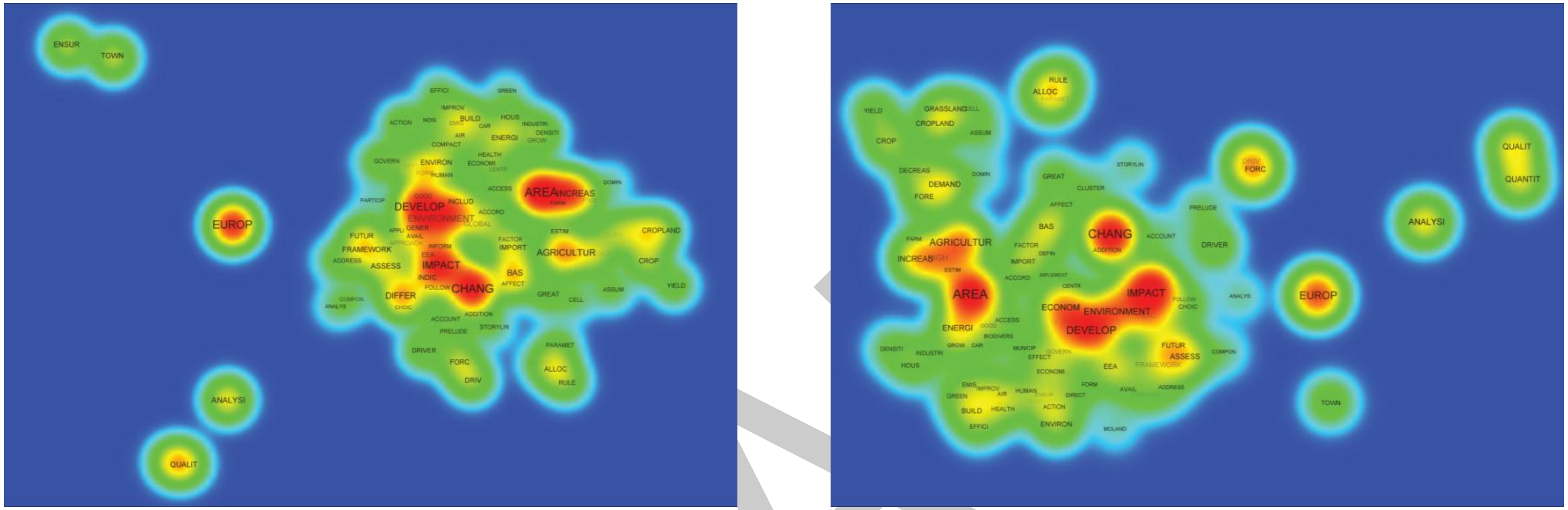
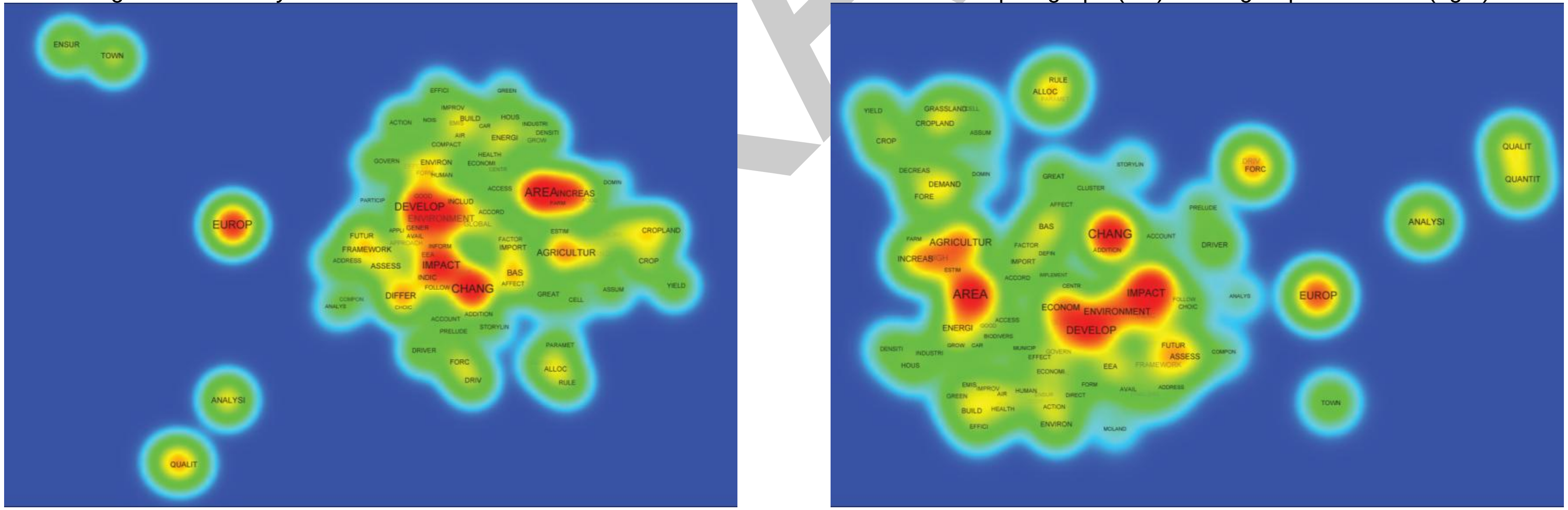


Figure 5d: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a group of 5 words (right)



Note: The word ‘scenario’, present in (5c), was added to a exclusion list in (5d)

Figure 6a: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a group of 5 words (right)

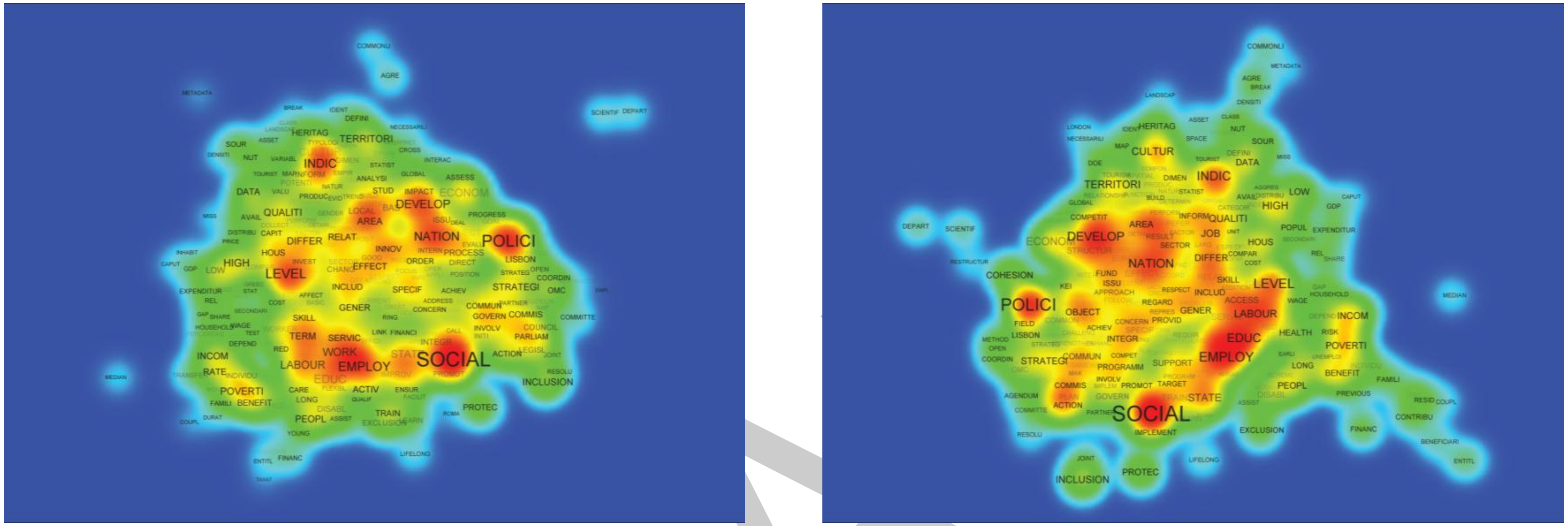
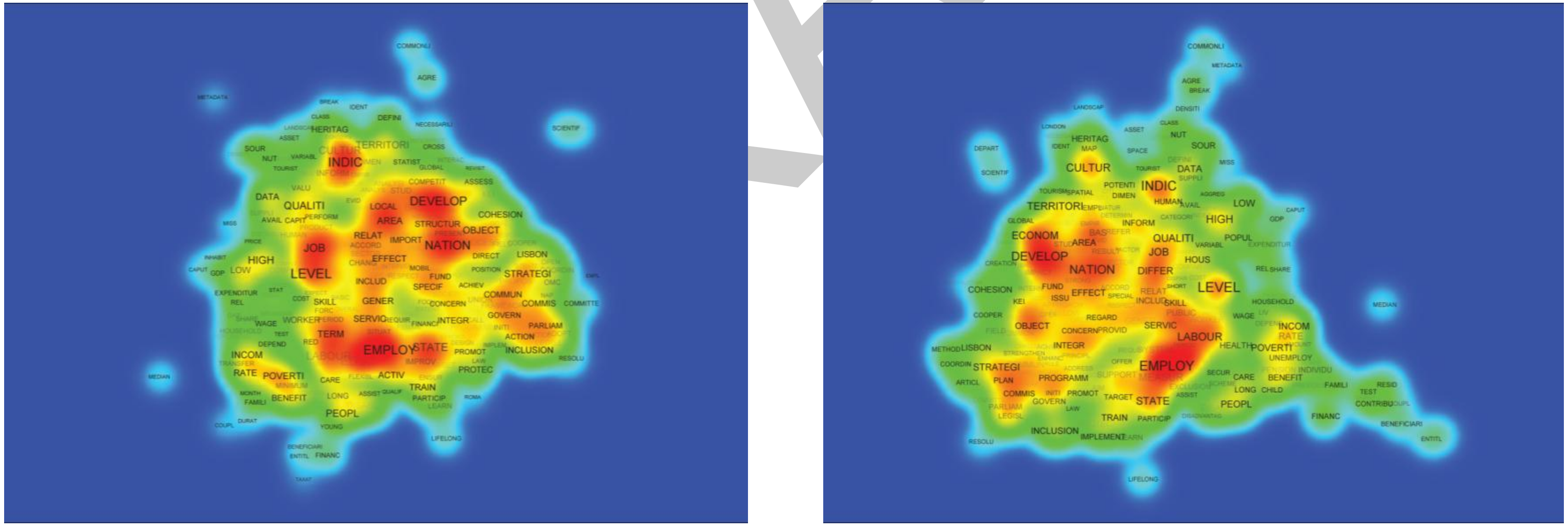


Figure 6b: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a window of 5 words (right)



Note: The words ‘socio’, polici’, present in (6a), were added to a exclusion list in (6b). TF: Term Frequency.

Figure 6c: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a group of 5 words (right)

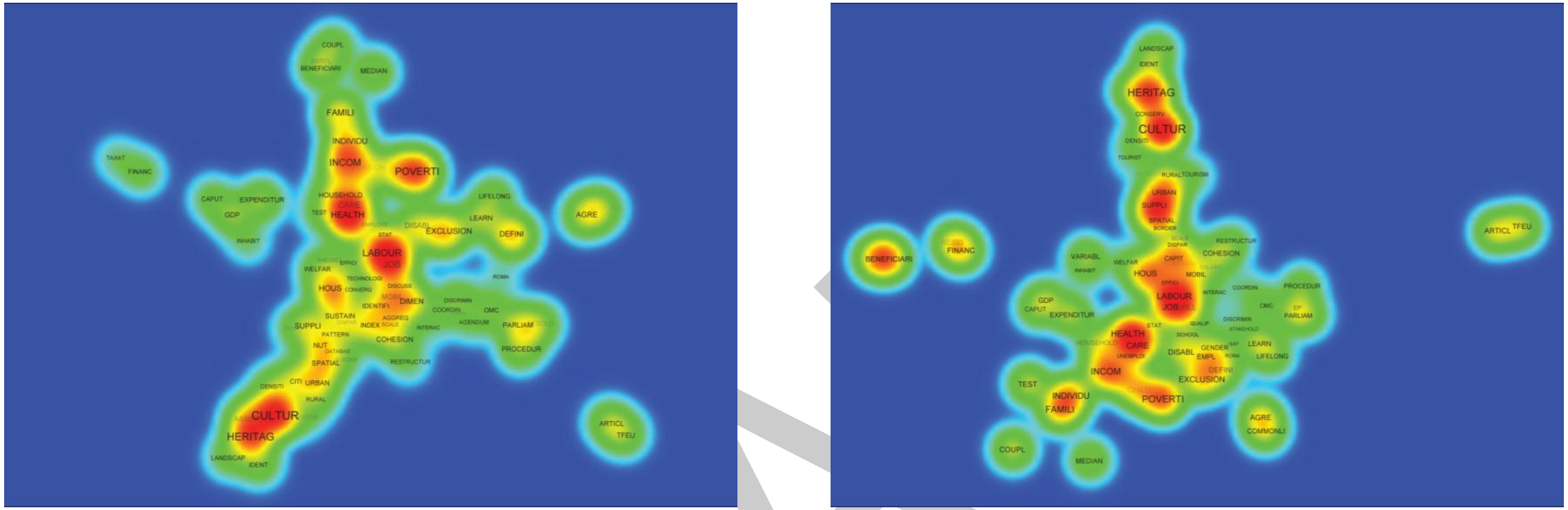
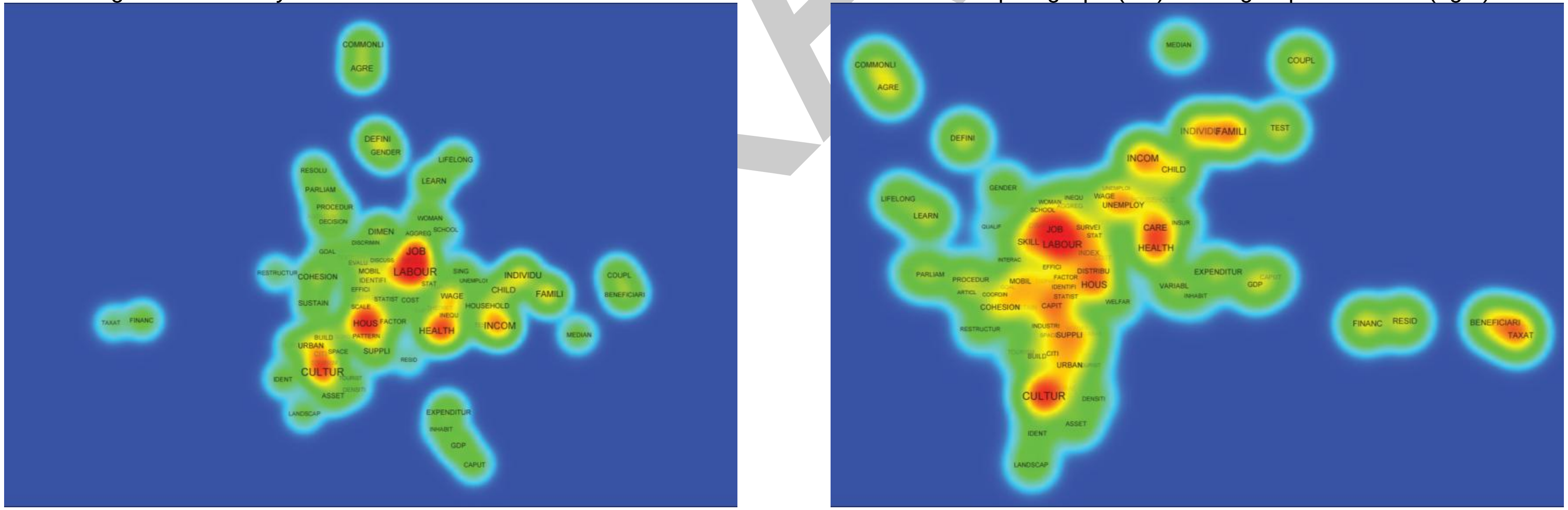


Figure 6d: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a group of 5 words (right)



Note: The words ‘heritag’, ‘empl’, ‘spatial’, pension’, ‘disabl’, and ‘poverti’, present in (6c), was added to a exclusion list in (6d)

Figure 7a: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a group of 5 words (right)

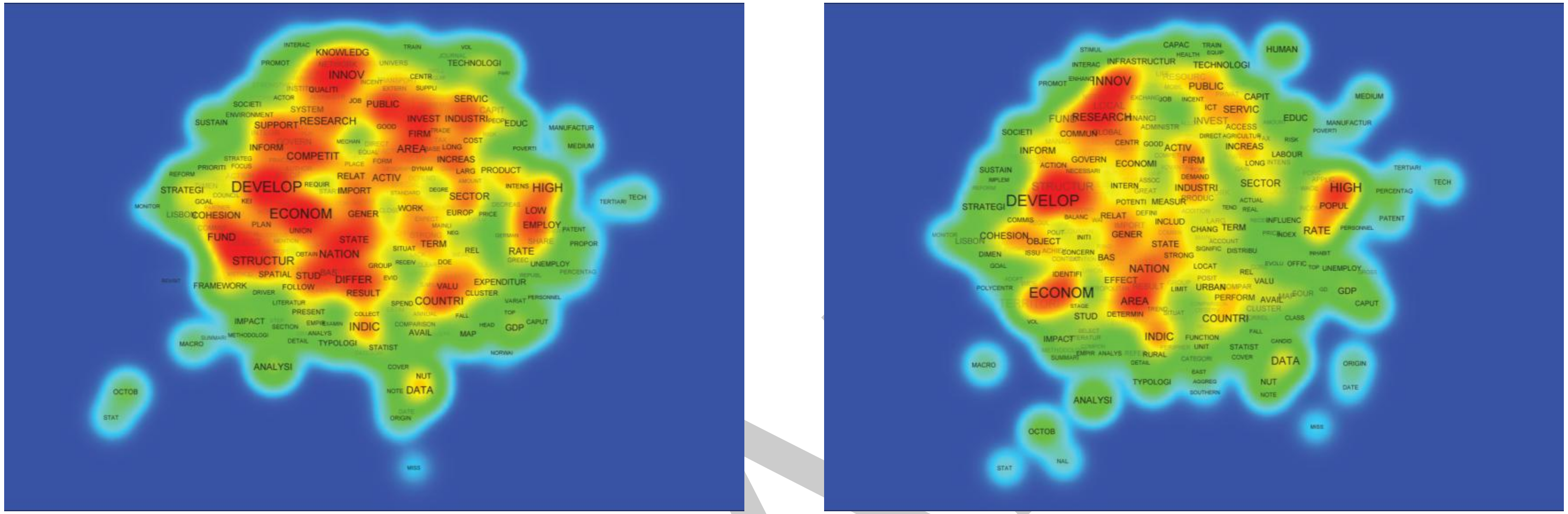
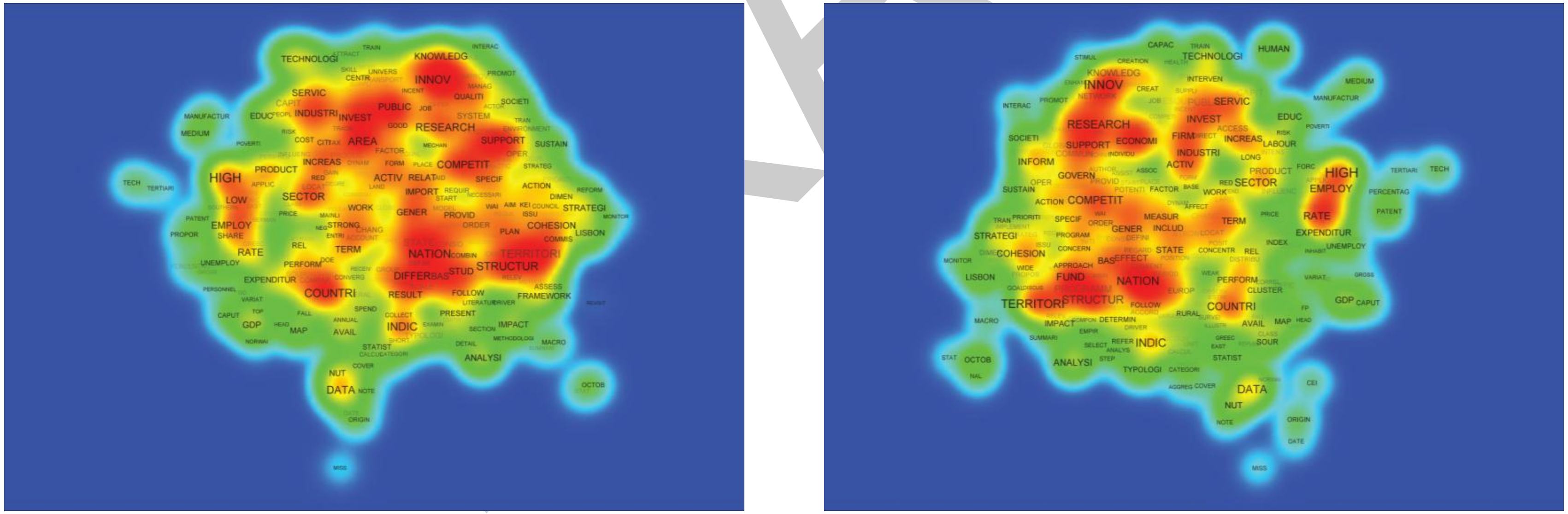


Figure 7b: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a window of 5 words (right)



Note: The words ‘develop, ‘econom’, present in (7a), were added to a exclusion list in (7b). TF: Term Frquency.

Figure 7c: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a group of 5 words (right)

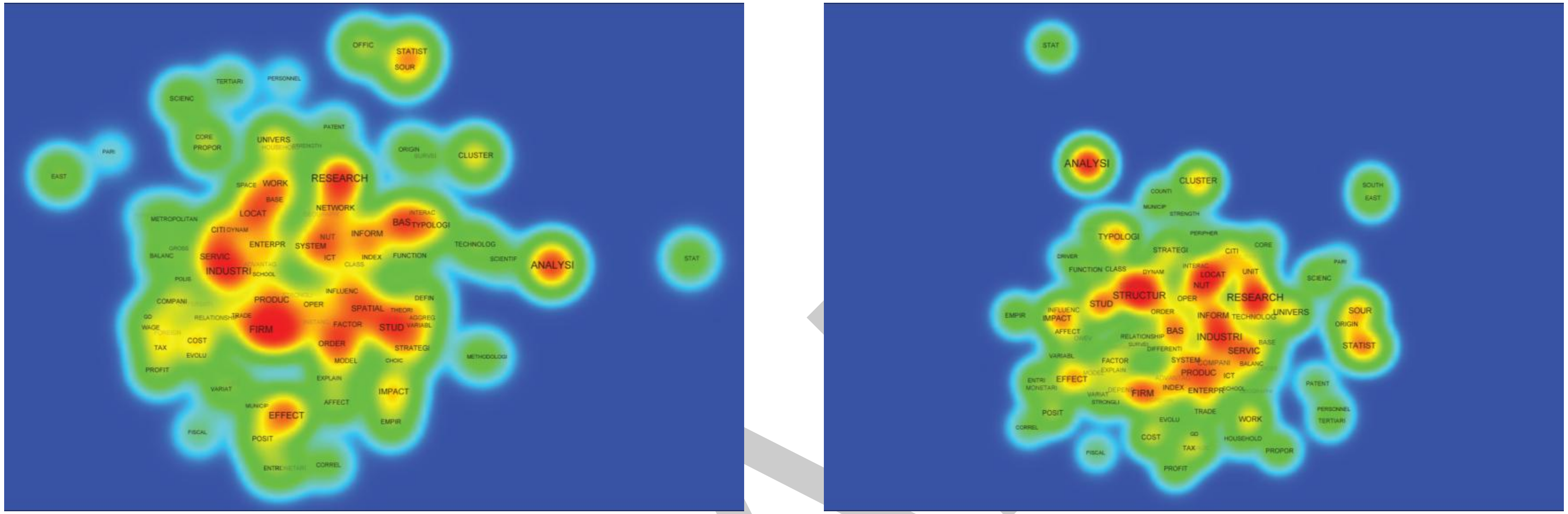
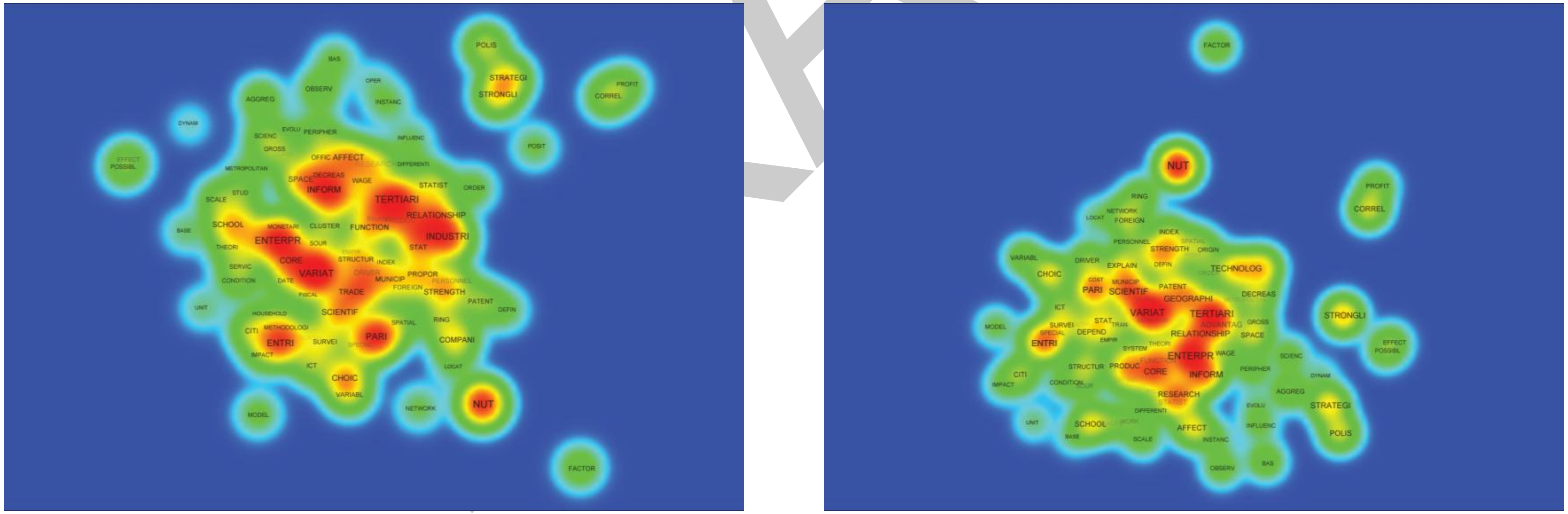


Figure 7d: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a window of 5 words (right)



Note: The words ‘typologi’, ‘firm’, ‘tax’, ‘counti’, and ‘gdp’, present in (7c), were added to a exclusion list in (7d)

DRAFT