AN EXPLORATORY DATA ANALYSIS ON ARCHITECTURAL RESEARCH AGENDA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

BARAN EKINCI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF ARCHITECTURE
IN
ARCHITECTURE

DECEMBER 2019

Approval of the thesis:

AN EXPLORATORY DATA ANALYSIS ON ARCHITECTURAL RESEARCH AGENDA

submitted by **BARAN EKINCI** in partial fulfillment of the requirements for the degree of **Master of Architecture in Architecture Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. F. Cânâ Bilsel
Head of Department, **Architecture**

Assoc. Prof. Dr. İpek Gürsel Dino
Supervisor, **Architecture, METU**

Examining Committee Members:

Prof. Dr. C. Abdi Güzer
Architecture, METU

Assoc. Prof. Dr. İpek Gürsel Dino
Architecture, METU

Prof. Dr. T. Nur Çağlar
Architecture, TOBB - ETÜ

Assoc. Prof. Dr. H. Ela Alanyalı Aral
Architecture, METU

Assist. Prof. Dr. Pelin Yoncacı
Architecture, METU

Date: 09.12.2019

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Baran Ekinci

Signature:

ABSTRACT

AN EXPLORATORY DATA ANALYSIS ON ARCHITECTURAL RESEARCH AGENDA

Ekinci, Baran
Master of Architecture, Architecture
Supervisor: Assoc. Prof. Dr. İpek Gürsel Dino

December 2019, 136 pages

The means of architectural production varies in terms of knowledge generation; however, the intellect product of the architectural domain is mainly written. The written mediums that are dedicated to architectural theories, discussions, research and agendas differ from articles, books, theses, etc., and these textual data increase exponentially by ever year. Thus, conventional qualitative data analysis became incapable to catch the pace of researching these texts and their corresponding knowledge. Therefore, novel methods such as textual data analysis emerged to index or extract useful information from texts. This study will focus on the architectural research conducted under Master of Architecture degrees in Turkey and aims to extract knowledge from the national electronic thesis database through text mining methods to develop a data-driven statistical model. This model will be the basis for the critical reading of the research agendas, their trends and the research characteristics of universities. Besides the analysis, the systematic approach that will be developed during this study is aimed to be a reproducible and scalable method for future studies.

Keywords: Architecture, Architectural Agenda, Text Mining, Topic Modeling, Knowledge Extraction

ÖZ

MİMARİ ARAŞTIRMALAR GÜNDEMİ ÜZERİNE KEŞİFSEL VERİ ANALİZİ ÇALIŞMASI

Ekinci, Baran
Yüksek Lisans, Mimarlık
Tez Danışmanı: Doç. Dr. İpek Gürsel Dino

Aralık 2019, 136 sayfa

Mimari üretim araçları bilgi üretme biçimi olarak değişmesine ragmen, mimari fikir ürünleri çoğunlukla yazılı biçimdedir. Mimari teoriler, tartışmalar, araştırmalar ve gündemine dönük olarak üretilen makaleler, kitap, tezler, vb. her geçen yıla göre sayıca katlanarak artmaktadır. Dolayısıyla, geleneksel nitel veri analiz yöntemleri bu artan sayıdaki belgelerdeki bilgileri araştırmak konusunda yetersiz kalmaktadır. Dolayısıyla, hızlı artış gösteren belgelerdeki bilgiye erişmek ve dizinlemek için metinsel veri analizi yöntemleri ortaya çıkmıştır. Bu çalışma, Türkiye'de yazılmış Mimarlık Yüksek Lisans araştırmalarına odaklanacak ve ulusal tez veritabanından bilgi çıkarımı için metin veri madenciliği yoluyla veri odaklı bir istatistiksel model geliştirmeyi amaçlamaktadır. Bu model, incelenen tezlerdeki mimarlık gündemlerini, yılları içindeki eğilimlerini ve üniversitelerin araştırma yönelimlerini anlamak ve eleştirel bir okuma için taban oluşturacaktır. Analizlerin yanı sıra; bu çalışma sırasında geliştirilen sistematik yaklaşım, gelecek çalışmalar için tekrarlanabilir ve ölçeklenebilir bir yöntem olmayı hedeflemektedir.

Anahtar Kelimeler: Mimarlık, Mimarlık Gündemi, Metin Madenciliği, Konu Modelleme, Bilgi Çıkarımı

To my companion of one: HAR

ACKNOWLEDGEMENTS

First and foremost, my gratitude is immeasurable for my advisor, supervisor and tutor Assoc. Prof. Dr. İpek Gürsel Dino. Her endless patience and infinite support kept me motivated over the years. Her extensive knowledge has driven my research motive and her guidance made this thesis possible. It has been an honor and privilege to be able to work with her.

I would like to thank to the thesis examine committee for their attendance, sincere contributions, and precious feedbacks. Their evaluation of my thesis defense has brought broader perspective to the discussions of this study. Additionally, my utmost appreciation to Prof Dr. Soofia Tahira Özkan, Prof. Dr. Ahmet Can Baykan, Prof. Dr. Güven Arif Sargın and Assoc. Prof. Dr. Ela Alanyalı Aral whose classes I was lucky enough to attend, which greatly contributed to my knowledge and intellect.

My warmest regards and love to my family of three: Erzem, Nazım and Bawer. They have never given up on me and provided me with the best they can.

I offer my heartfelt thanks to my dearest friends and colleagues Ece Yeğen, Burak Ercan, and Tuğçe Çelik for their companionship, encouragement and solidarity during the past years. Besides of their intellectual contributions to this study; I will always remember their friendship, and I hope to be remembered by them.

My sincere gratitude to Soner Güner, who provided the professional working environment and opportunity during this study. Furthermore, I am thankful to Ömer Emre Şavural and Fatih Yavuz for their encouragement and recommendation for my acceptance to this program.

Last but not least, my deepest gratitude to Tuba Engel for her presence and support during the overwhelming stressed times: where there is sorrow, let it be joy; where there is hatred, let it be love; where there is doubt, let it be faith; and where there is despair, let it be snoogums boogums.

TABLE OF CONTENTS

CHAPTERS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

EDA: Exploratory Data Analysis

EM: Expectation-Maximization Algorithm

IDF: Inverse Term Frequency

LDA: Latent Dirichlet Allocation

LSA: Latent Semantic Analysis

LSI: Latent Semantic Indexing

PLSA: Probabilistic Latent Semantic Analysis

STM: Structural Topic Model

TF: Term Frequency

TM: Text Mining

CHAPTER 1

INTRODUCTION

Prior to this study, the initial research motive was to observe, comprehend and mimic the design process of architectural design with computational tools. The grift and ill-defined nature[1] of a design problem requires an intuitive design process and with a unique approach for each designer subject[2]. The research subject was to construct a dataset of the decision making and problem space generation of individual designers within a given design problem and with the aid computational tools regenerate the process with novel stochastic approaches. The motive was to map the problem space and navigate the computer through it with a probabilistic process of decision making, in order to elevate the concept of computer-aided design as a feasible supportive decision-making tool for architectural design.

This initial research motive evolved from the aim of generating and observing the cognitive map of individual intellects; towards comprehending and mapping the collective products of the architectural domain. While these collective products may initially assert as mere praxis; within the domain of architecture, the theory is a discourse that describes the practice and production of architecture and identifies challenges to it[5]. As a furthermore Jacobin statement: there is no architecture without theory [6].

[1] Reitman, *Heuristic Decision Procedures Open Constraints and the Structure of Ill-Defined Problems*, New York: John Wiley & Sons Inc, 1964.

[2] Lawson*, How Designers Think the Design Process Demystified, How Designers Think The Design Process Demystified*, Burlington, MA, Elsevier, 2005.

[5] Nesbitt, *Theorizing a New Agenda for Architecture an Anthology of Architectural Theory 1965-1995*.

[6] Jencks, *The Architecture of the Jumping Universe. A Polemic: How Complexity Science Is Changing Architecture and Culture*.

Architectural theory drives the architectural praxis and the theory clearly distinguishes from a retrospective evaluation: architectural theory differs from the history in terms of being incentive to the praxis and emergence as an agenda settler. This aspect of the theory is even acknowledged by the renowned "starchitects"[7]:

> "[…] Architectural theory unifies and stabilizes architectural practice. In its written, theoretical treatises an architectural practice fixes its premises, values, turns of argument and conclusions. In this explicit form – open to everybody's inspection and reflection – architectural theory exposes itself to criticism and further dialectical evolution. As an invitation to criticism, theory thus becomes an engine for the progressive transformation of Practice. […] "[8]

In the state of theory initiating the architectural praxis, this initiation -or transformation- has different means of production: from drawing to writing. As a general presumption, the main architectural expression or communicator is deducted to sole drawings; however, the architectural theory -thus the knowledge- is represented in texts:

> "[...] architecture does not exist without drawings, in the same way, that architecture does not exist without texts. Buildings have been erected without drawings, but architecture itself goes beyond the mere process of building. The complex cultural, social, and philosophical demands developed slowly over centuries have made architecture a form of knowledge in and of itself. Just as all forms of knowledge use different modes of discourse, so there are key architectural statements that, though not necessarily built, nevertheless inform us about the state of architecture -its concerns and polemics- more precisely than the actual buildings of their time. [...] "[9]

[7] The term "starchitect" refers to architects whose celebrity and critical acclaim have transformed them into idols of the architecture praxis and even further some degree of fame among the general public

[8] Schumacher, *The Autopoiesis of Architecture: A New Framework for Architecture, Volume I*.

[9] Tschumi*, Architecture and Limits I*, Artforum 19 no 4,1980.

Tschumi presents architecture as a “form of knowledge”. Furthermore, criticizes the comprehension of architecture as a “knowledge of form” and ultimately remarks that this knowledge is expressed in a written medium.

Starting from the early 1900s to the 1950s tracking the architectural agenda was more relaxed; they were relatively few publications and conferences, any paradigm changes happened was in slow motion. By the 1960s, the architectural theory and agenda started to rapidly expand and change. Anthological studies like Kate Nesbitt[10] and Michael Hayes[11] complied different agendas by either topically or chronologically.

Furthermore, Charles Jencks mapped the evolution of architectural styles with an empirical method of content selection (Figure 1.1). However, he published multiple maps spanning until 2010 and his empirical approach remained intact with the new maps.

> “[…] In recent years, the architectural intellectual discourse underwent a significant transformation as the historical and historiographical scholarships were influenced by critical theories and methodologies. Architectural history is not any longer considered as a grand-narrative, but rather interpreted as a multiplicity of political conditions of identity created by spatiality and architecture. Nevertheless, while most researches effectively elaborate on the interrelations manifested by space and architecture, they sometimes collapsed into narrow points of view, neglecting to address the multilayered significations of the architectural texts as such. […] ”[12]

Studies as such tracked the agenda by a discursive approach and their timeline ended by the mid-1990s where more topic dedicated studies emerged. With a rapid change in technology and the ways of information sharing, the architectural agenda became

[10] Nesbitt, *Theorizing a New Agenda for Architecture an Anthology of Architectural Theory 1965-1995*.
[11] Hays, *Architecture Theory since 1968*.
[12] Declaration for the *“Architecture and Phenomenology an International Conference at The Technion”*. Israel Institute of Technology Faculty of Architecture and Town Planning. 2007

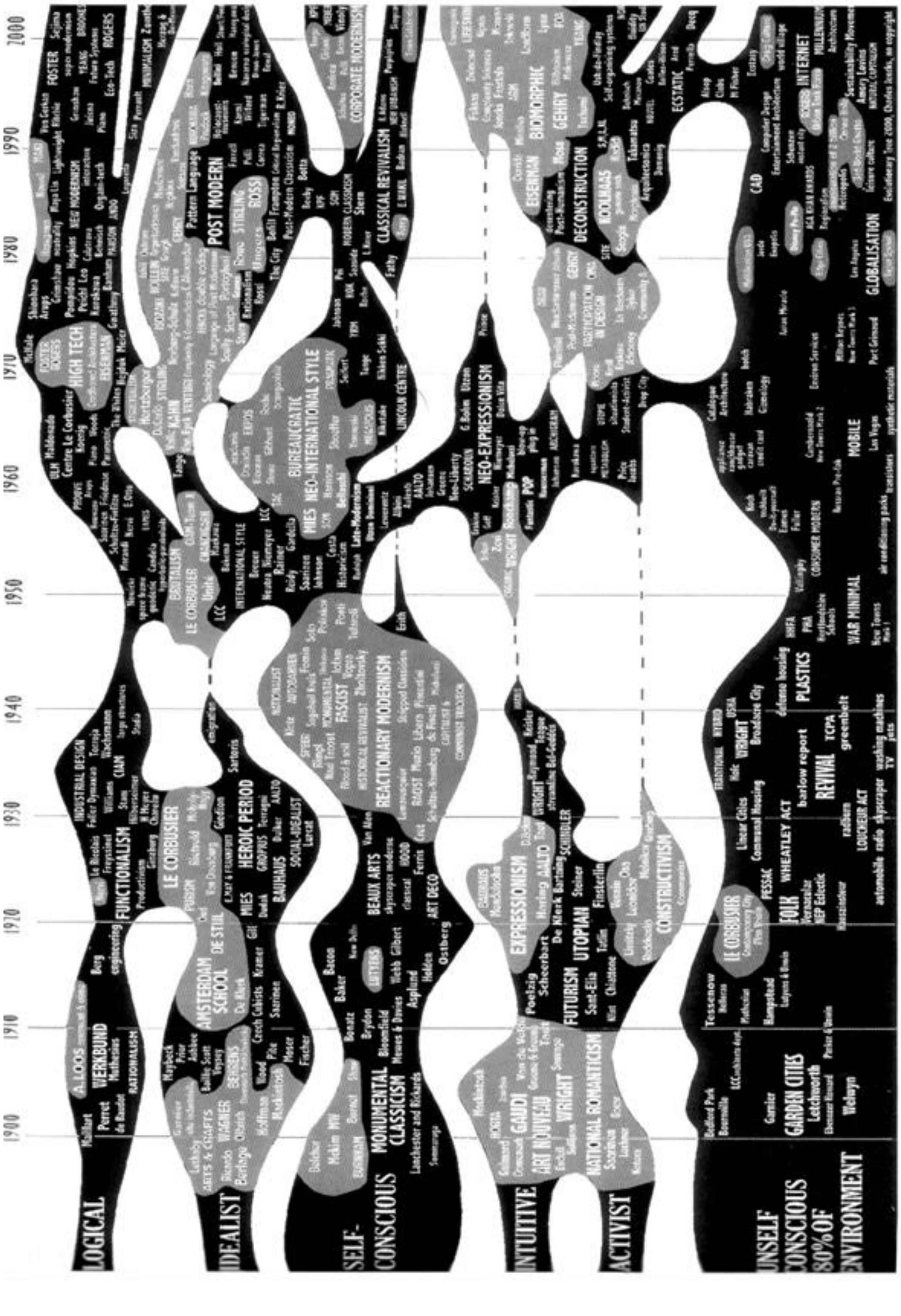


Figure 1.1. The Century is Over, Evolutionary Tree of Twentieth-Century Architecture. 2000. Jencks

hard to track. It is constantly changing due to the vast amount of contributors: published sources, magazines, theses, online articles, personal blogs and so on.

These documents are from different genres; scholar or industry-related and varies by means of aim and scope. Since the architectural domain contains social, economic, technological, philosophical, cultural, historical, etc. topics it makes the track of architectural agenda unpractical by homogenizing these texts with such variant sources. Additionally, on a global scale, these texts are written in various languages and increasing exponentially. Thus, a single-handed study cannot be conducted in this diversity.

> “[..] When I began work in 1993 the writing in the discipline was largely produced as four kinds of things: academic scholarship supposedly neutral on covering and sharing effects in journals like JSAH and Journal of Architectural Education. The trade magazines Record, Progressive Architecture, etc. devoted primarily presenting new works. Architectural criticism in the New York Times, Village Voice, Chicago Tribune, etc. added a fringe of theoretical speculation presented in publications… Almost twenty years later writing in the discipline has greatly diversified breaking out of such categorizations in large part of course because of the colossal platform that the internet presents. Although, architectural books made in 2011 do not seem categorically different from books made in 1991, this is partly an illusion that the discipline of architectural, history in particular, has undergone significant changes [..]” [13]

[13] Saunders, *“What is to be written”*. 2011

1.1. Problem Statement

Within this context, in the era of mass-communication and “big data”, the problem of tracking and exploring the architectural agenda emerges. While conventional -manually compiled- methods deem to fail in terms of scale, reproducibility and time consumption; novel approaches can provide a feasible ground of exploration and evaluation.

This thesis aims to gain insight into the architectural research agenda from the Master of Architecture theses published in Turkey from 2003 to 2018. These theses are to be treated as textual data that in terms of research represents the architectural theory, discourse and above all the agenda of post-2003’s Turkey. Henceforward, the textual data will be evaluated by statistical and computational techniques, and through critical reading, the aim is to extract knowledge by:

1. Defining and categorizing architectural research agendas
2. Detecting novelty and emergence of agendas over time
3. Comparing the agendas of different sources

The statistical techniques are presented as a novel medium to explore and interpret with the vastly increasing architectural textual documents. Therefore, another outcome of this study is to introduce a reproducible and scalable method of qualitative research to the architectural domain.

1.2. Method and Structure

This study utilizes topic modeling algorithms and techniques to classify the architectural research collection: these techniques are commonly used for discovering information from a collection of documents by classifying them in abstract topics. This is achieved by enumerating words that constitute the documents, and then by clustering these words under pre-determined number of topics. Afterwards, the documents are clustered under mixture of topics. The clustering process is a probabilistic process in which the words are clustered in topics according to their probability of co-occurrence and frequency, hence words are gathered in every topic with different probabilities. Likewise, documents are affiliated to topics with different probabilities; the degree of affiliation (probability) enables to interpret a topic regarding to the information it contains through clustered documents and keywords (Figure 1.2).

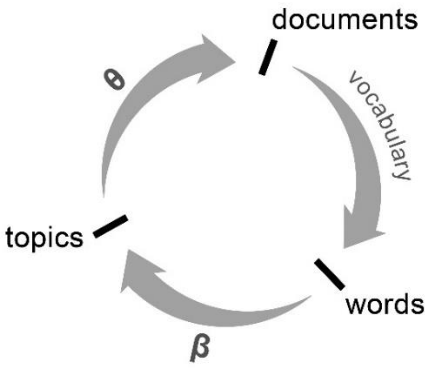


Figure 1.2. Diagram of Topics -Words- Documents relation in Topic Modeling

The first crucial concept regarding the mechanics of this document clustering process is that topics are probability distributions over words and documents are over topics, where a probability distribution is an equation which associates each probable outcome of a random variable with its probability of occurrence. In other words, every topic (probability distribution) is constituted by all the words (random variable) in the dataset, but with different probabilities. Likewise, every document (probability distribution) is associated with all the topics (random variable) by various probabilities.

The second central notion is the process of generating a model of topics from a set of words: in brief, this is achieved through Bayes' theorem [15]. This theorem allows the use of "prior" knowledge of a data to calculate the probability of a relevant event, in other terms a probability of an occasion can be calculate by a depending condition. Based on this theorem, a process of deducing properties regarding a distribution from the data, which is defined as Bayesian inference. The conditional probability distribution allows to generate topics based on desired expectations. The core concept is that the topic-word and document-topic relations is generated through an inference model based on expected (conditional) properties. This subject has immense applications on the field of statistical modeling, especially in machine learning.

In the view of these information, a topic is not an absolute representation of a concept, term or notion; but rather a pattern of co-occurring words which suggests a hidden knowledge. In depth information and other related key concepts regarding topic modeling are presented at Chapter 2.

Although topic modeling does not attain an absolute certainty over the knowledge contained within the document collection due to its statistical mechanics, it provides a certain level of flexibility to interpret the results depending on the researchers' aim and intention. Moreover, this study highly emphasizes the framework of Structural Topic Modelling[16] -introduced by its authors and contributors- to interact and interpret the results.

[15] Bayes, T. and Price, R. "An Essay towards solving a Problem in the Doctrine of Chances". 1763

[16] See Section 2.2.5

The workflow of the methodology is similar to most textual data analytics methods: the gathered text data are to be organized, structured, preprocessed, visualized and processed for human interpretation (Figure 1.3)

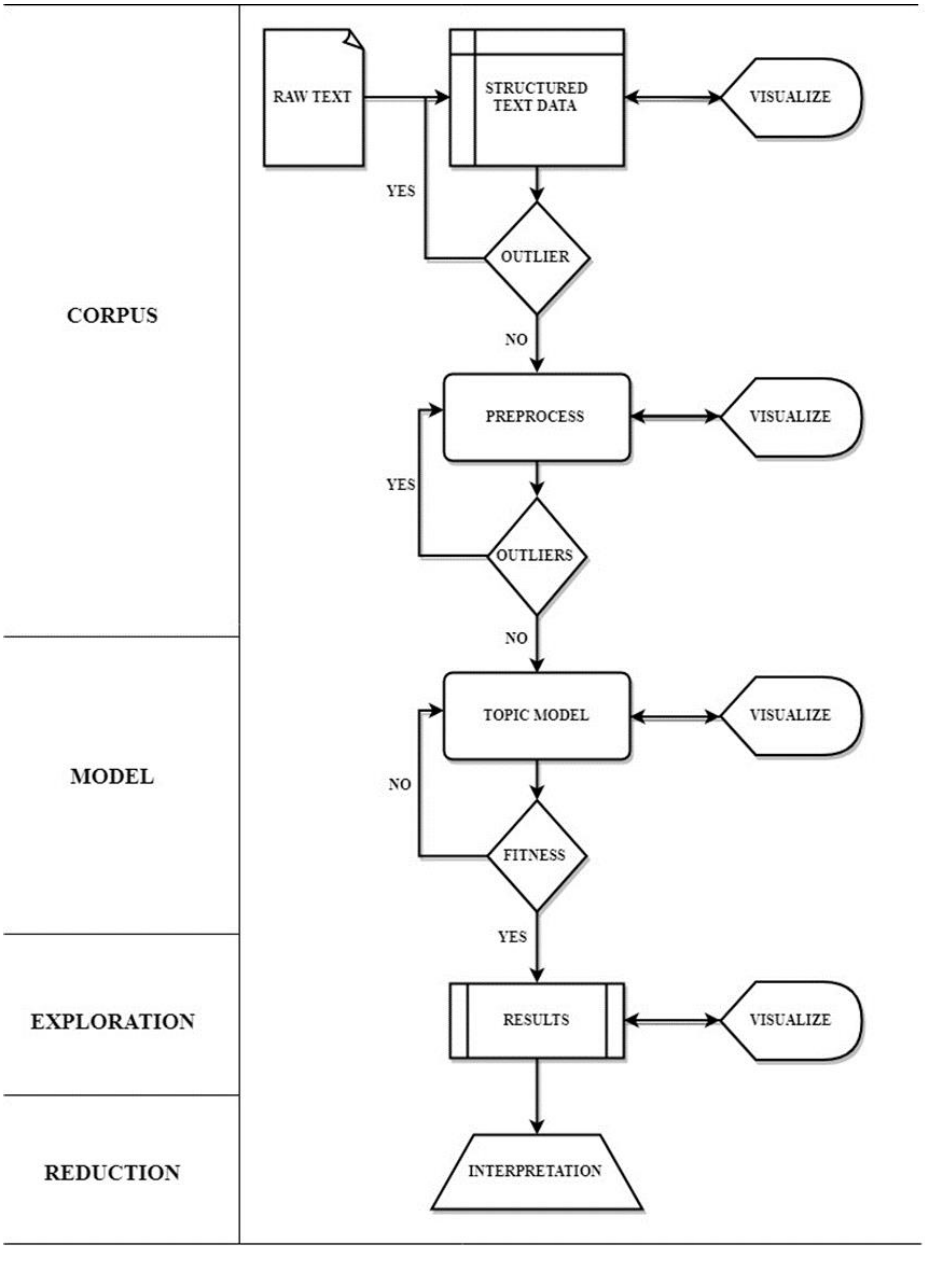


Figure 1.3. Layout of Methodology

This study heavily depends on concepts from different fields such as computer science, statistics and information technologies; hence, Chapter 2 of the thesis is dedicated to the explanation of the terms, concepts and literature reviews that are not related to the architectural domain.

Chapter 3 introduces the attributes, assumptions and features of materials that are subject to this study. Chapter 4 introduces the methodology and the initial results of the study based on the information and techniques discussed in Chapter 2. The continuing Chapter 5 is dedicated to the interpretation of the results within the architectural epistemology. The last chapter remarks the final thoughts, conclusion and the discussion of the future works.

The subject documents and the numerical outputs of the study is presented as a tabular format in the appendices. These outputs both serve as a validation of the results and a dataset for future works.

1.3. Limitation

This study is limited by its to the theses published under the degree of Master of Architecture, that is available to access and submitted to the Higher Education Council (YÖK) of Turkey. Therefore, theses that still have an aim or scope regarding the architectural agenda of Turkey, but not unpublished or not submitted to YÖK are neglected.

CHAPTER 2

BACKGROUND

This study is an interdisciplinary research were architectural domain intersects with different fields: statistics, information technologies and natural language processing which itself is a junction of computer science, artificial intelligence and linguistics. While these are large fields, this chapter is purposed to be an explanatory passage for the subjects, concepts and terms that are essential for the research and that are out of the bounds of architectural knowledge and epistemology.

2.1.1. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a data analysis approach which is a combination of methods (mostly graphical) to gain insight into a dataset. EDA is commonly used to extract significant aspects, detect outliers and anomalies, testing assumptions and to further develop prediction models for hypothesis testing[18]. A typical statistical analysis focuses on one or limited perspective of the data, in general the variable of interest or subject to a hypothesis [19]; EDA on the other hand has a broader scope. It is an attitude towards data analysis that ignores any assumptions about the data or its aspects. EDA is a more distinctive approach of allowance to the data itself to reveal its content. It is not a mere compilation of techniques; rather a strategy of selecting a dataset, evaluation of variables and interpretation of the results.

Tukey introduced EDA[20] as a contradiction to the confirmation -biased- data analysis

[18] Anscombe and Tukey, “The Examination and Analysis of Residuals.”

[19] Chatfield, “Exploratory Data Analysis.”

[20] Tukey, “The Future of Data Analysis.”

which is mainly statistical hypothesis testing. Tukey argues that data analysis is not just testing a pre-defined hypothesis and cannot be reduced to a single set of separate calculations, on the contrary that exploratory data analysis is about observing the data to see what it seems to "express"[21]. By neglecting any pre-determined hypothesis, EDA provides a basis for a subjective observation of the data. To communicate with the data within this manner a graphical portrayal the data should be rendered to: data visualization[22].

The primary notion for data visualization is to present data in a graphical format to facilitate the ease of understanding since it is easier to detect a data pattern from a picture than from a numeric output. Data graphics render the quantities of interest to a visual medium by the mixed usage of a coordinate system with the data represented as points, lines, numbers, symbols, words, shading, color and so on. They are also substantial instruments for evaluating quantitative information, rather than a substitute for statistical tables. Often the most effective way to assess, explore, and summarize an extensive set of input is to look at the visual representation of those inputs. Furthermore, of all the techniques for analyzing statistical information, data visualization is the simplest and the most efficient method[23].

The actual exploring of a dataset can be possible with data visualization: Detection of errors or examining relationships between variables can be possible by observing data plots. Creating a meaningful visualization requires several successful steps: so, with each plot early stages may be addressed again for structuring, organizing or cleaning the data. By observing of the data with various plots of variables and by questioning the data, generation of a hypothesis or re-exploring the data with different observational units to come to different conclusions or problem statements.

[21] Tukey, *Exploratory Data Analysis*.

[22] Friendly, "Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization."

[23] Tufte, *The Visual Display of Quantitative Information*.

2.2. Text Mining

The primary aspiration and motive of "knowledge discovery" is to extract previously unknown information to an appropriate and more definite form of knowledge through data[24]. Since the available electronic (digital) data, in various mediums, is exponentially growing the automated knowledge extraction methods become essential to evaluate the vast amounts of data stored in the information systems [25]. Furthermore, while data mining techniques were initially designed to gather data on structured databases, mostly numeric (quantitative) variables; to process unstructured, commonly textual inputs, a series of dedicated methods and techniques have developed, where the general term referring to the is Text Mining (TM).

TM can be defined as an extension of conventional approaches of unstructured textual data mining with the concern of with various tasks extracting implicit knowledge from large sets of text document collections to explicit knowledge utilizing structuring and visualization[26].

Besides the classical approach of processing and mining raw text, the advent of web-enabled applications requires novel methods for mining and processing, such as the use of linkage, multi-lingual information or the joint mining of text with other kinds of multimedia data such as images or videos [27].

Text mining, or text analytics[28] has various implication fields from: machine learning applications, security applications, biomedical applications, online media applications, business and marketing applications, sentiment analysis applications, academic research applications, digital humanities and computational demographics applications. The scale of the implication fields of TM is trending concerning the

[24] Frawley, Piatetsky-Shapiro, and Matheus, "Knowledge Discovery in Databases: An Overview."
[25] Rajman and Besançon, "Text Mining - Knowledge Extraction from Unstructured Textual Data."
[26] Rajman and Besançon.
[27] Charu C. Aggarwal and Chengxiang Zhai, eds., Mining Text Data, Mining Text Data (Springer-Verlag New York, 2012), https://doi.org/10.1007/978-1-4614-3223-4.
[28] Fayyad, Piatetsky-Shapiro, and Smyth, "From Data Mining to Knowledge Discovery in Databases."

exponential growth of accessible text data. Besides the scholars, almost half of the world population is connected through online internet platforms and they are generating bulks of textual information which can be mined to get insights about different themes, domains, disciplines and topics.

2.2.1. Terminology

TM is a growing field for researchers and analysists alike, and it is a vital technique for this study. Before proceeding with the core concepts and related background information, this section covers the essential terms regarding TM terminology and furthermore.

- Implicit knowledge defines as information, or knowledge, not expressed directly. It is the antonym of explicit knowledge which refers to clearly and fully expressed information.

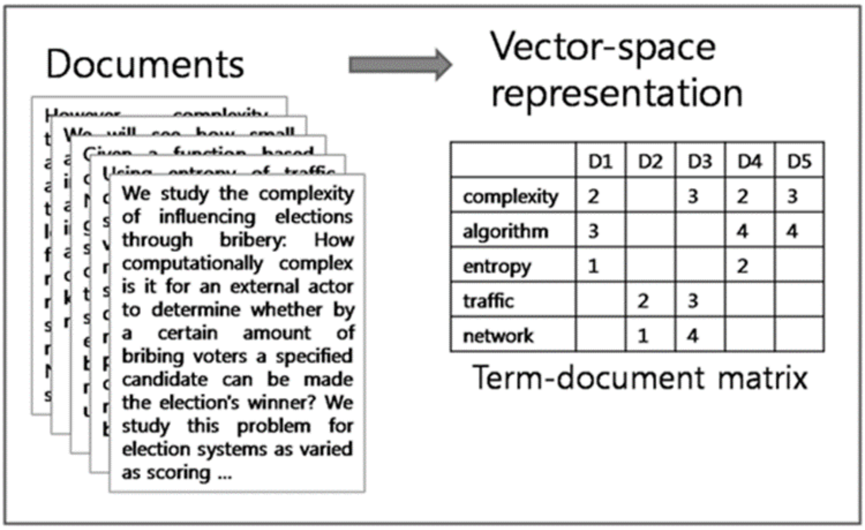
- Semantic by definition means connected with the meanings of words. Languages structure around combined rules known as syntax and semantics is the meaningful output of this structure.

- Latent defines as something present and capable of emerging or developing but not now visible or apparent. Therefore latent semantic refers to a hidden meaning in a language; thus, it expresses implicit knowledge.

- Corpus refers to a collection of documents, texts, or any form of textual data.

- Bag of words (BOW) is a representation of textual data by neglecting its semantic rule of word order. BOW is commonly used to organize and enumerate words in a document or corpus for ease of computer (machine) processing. In general, words in a text are re-ordered according to their frequency and represented in a tabular form.

- Document Term Matrix (DTM) is a vector-space representation of a corpus as a matrix where the rows are documents, columns are words (terms) and cell values are the frequency of each term regarding their document. DTM is an expression of BOG model which serves to calculate metrics of interest, clustering and statistical analysis. (Figure 2.1)



We study the complexity of influencing elections through bribery: How computationally complex is it for an external actor to determine whether by a certain amount of bribing voters a specified candidate can be made the election's winner? We study this problem for election systems as varied as scoring ...

	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Figure 2.1. A sample Document-Term Matrix[29]

- Metadata is the data set that contains information relating to other data.

- A statistical model is a mathematically formalized model that incorporates a dataset to progress by statistical techniques.

- Regression is a statistical measurement of determining the significance of the relationship between two or more variables of a dataset. This determination is achieved by aligning a mathematical function that will fit with most of the

[29] http://www.speedlab.io/en/2017/03/08/creation-semantic-variables-based-document-term-matrix-dtm/

variables. Although they are various types of regression analysis, the most common are linear and non-linear regressions (Figure 2.2).

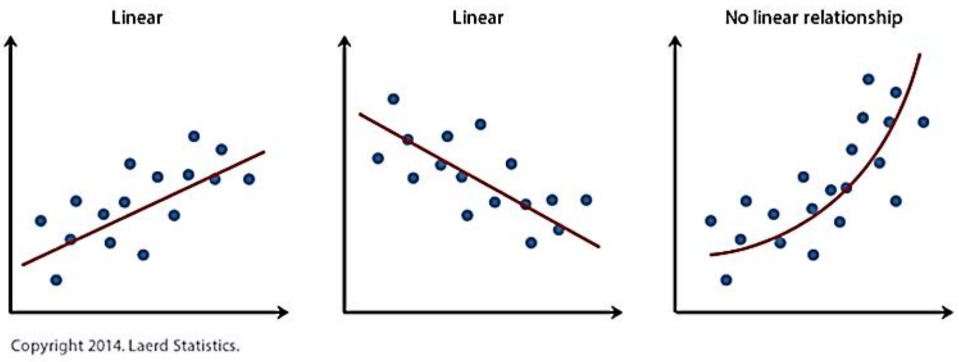


Figure 2.2. Linear and Non-Linear Regression plots. Laerd Statistics 2014 [30]

- Machine learning is the most common application of artificial intelligence in which a so-called “learning” is accomplished through data by algorithms. This learning is eventually a statistical model to generate a regression of the data for further predictions of non-existing scenarios. Machine learning is categorized into two main subjects: Supervised and Unsupervised. The former is a learning process with the “training” of predefined attributes: e.g., facial recognization applications that can identify a specific person from photographies are “trained” with sample photographs so that the algorithm can identify humans from the previously unseen images. Contrarily, unsupervised learning resembles the data and tries to identify a pattern to classify them with any prior training or information: e.g visual recognization applications that detect different colors from photographs by categorizing the contrast between colors.

[30] https://statistics.laerd.com/spss-tutorials/linear-regression-using-spss-statistics.php

2.2.2. Latent Semantic Analysis

The essentials of information retrieval through a human-machine interaction is achieved by natural language queries[31]: where the user submits an inquiry to the computer by keyword inputs and eventually the machine outputs a list of relevant subjects or contents. Typical examples for this kind of queries are search engine platforms such as Yahoo, Bing and Google. While these search engines are improving their information retrieval techniques, conventional techniques are built on basic word matching algorithms where the printed outputs are determined by the relevance order of the content regarding the user inquiry. However, this workflow has several handicaps: mainly because of the entry of ambiguous words, the prevailing lack of accuracy in matching and the difference of personal style in word usage[32].

Latent Semantic Analysis (LSA) was introduced in the late 1990s, to overcome these drawbacks, as an approach for automatic indexing and information retrieval by representing the corpus in a Document-Term Matrix (DTM) space[33].

The next phase to LSA is to apply a linear algebra implementation, Singular Value Decomposition (SVD), which takes the DTM of the corpus and applies a dimension reducing linear projection (Figure 2.3). SVD is a matrix decomposition method for shrinking a matrix to its composing parts apparent eliminating less important contents of the data, thus elevating the associative patterns of the data. While reducing the matrix, terms that did not appear in a document may still be aligned close to that document; if that is consistent with an associative pattern of the data, the positioning the document and term in the matrix can be evaluated as semantic indexing. Thus, documents can be associated with each other through different words according to their neighborhood in the vector-space[34].

[31] Hofmann, "Probabilistic Latent Semantic Indexing."
[32] Hofmann.
[33] Deerwester et al., "Indexing by Latent Semantic Analysis."
[34] Deerwester et al.

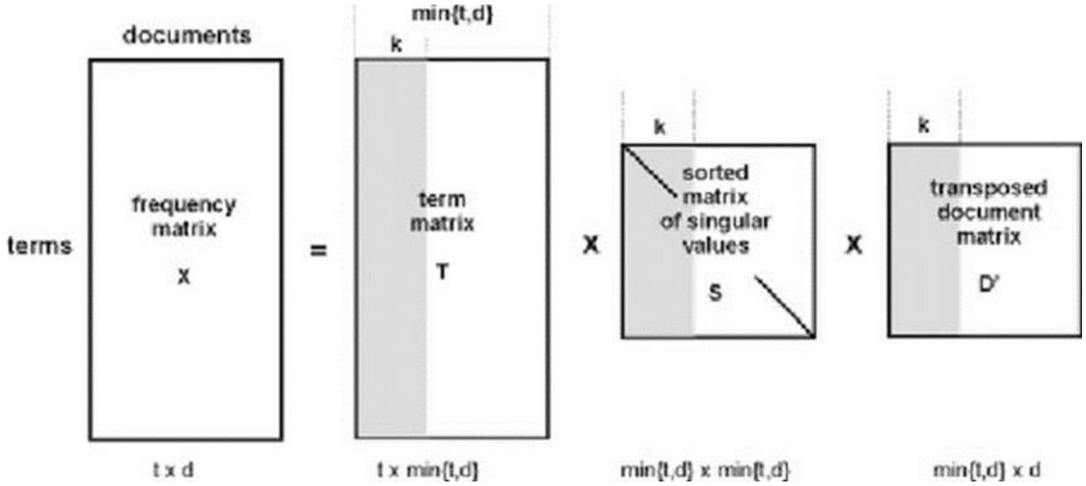


Figure 2.3. A Singular Value Decomposition results in three partial matrices.[35]

LSA method claims that similarities between documents -and queries- can be more accurately estimated in the reduced vector-space representation rather than its original space. The reasoning is that documents that neighbor due to the high co-occurrence of the terms will have a similar representation, even though they have few terms in common. LSA thus performs a type of noise (sparsity) reduction and can detect words that refer to the same topic[36].

[35] W. Lenhard, Bridging the Gap to Natural Language: A review on Intelligent Tutoring Systems based on Latent Semantic Analysis. 2008

[36] Hofmann, "Probabilistic Latent Semantic Indexing."

2.2.3. Probabilistic Latent Semantic Analysis

In 1999 Thomas Hofmann introduced probabilistic latent semantic indexing (PLSI)[37], also known as probabilistic latent semantic analysis (PLSA), as a novel approach to automated document indexing which is based on a statistical model for factor (topic) analysis of textual data.

In general, a probabilistic topic model automatically extracts latent (hidden) topics from document sets. Since the documents are comprised of words, extracting topics is achieved by treating a topic as a probability distribution over words and a document as a mixture of topics: as a result, the topics define a joint probability distribution[38]. Probabilistic topic models are generative models, for every document in the dataset the model generates a latent topic and further evaluates the words by topics to find similarities or pattern in the model [39].

In contrast to LSA, PLSA has a statistical background and it provides a generative data model (Figure 2.4). In a *M* numbered space of documents and where the documents consist *N* number of words, a PLSA model processes by:

1. selecting a document *d* with probability *P(d)*

2. picking a topic *z* with probability *P(z|d)*, where the probability that the topic *z* occur given that document *d* has occurred.

3. generating a word *w* with probability *P(w|z)*, the probability that the topic *w* occur given that document *z* has occurred.

[37] Hofmann.

[38] Blei, "Probabilistic Topic Models."

[39] Wu et al., "Locally Discriminative Topic Modeling."

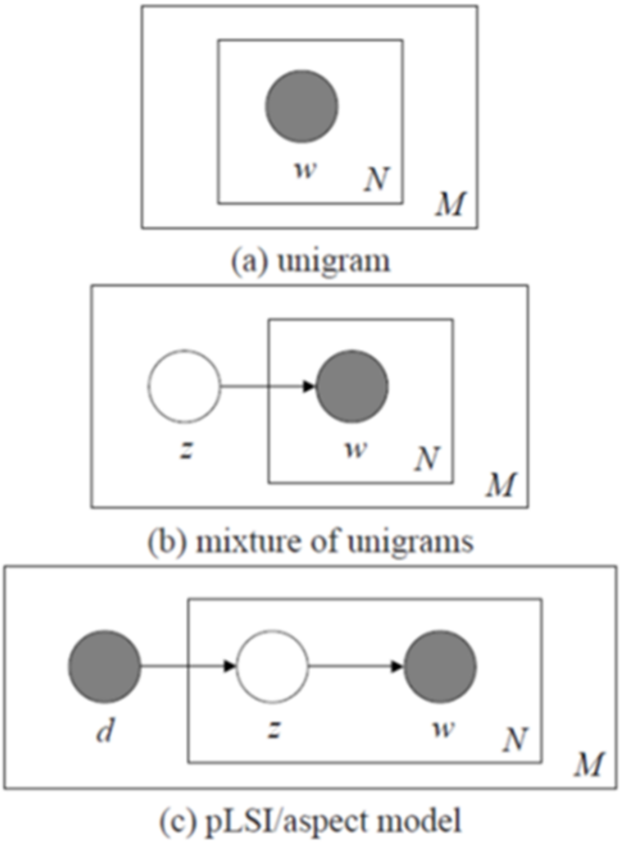


Figure 2.4. The Plate Diagram of PLSA.
d represents the document, z represents the topic and w represents the word. N is the number of words and M is the number of documents

The PLSA improves the initial attempt of modeling method for a mixture of unigrams (bag of words) where each document is associated -therefore generated- from a single topic: on the contrary it assumes a document can contain multiple topics [40]. PLSA, as a two-level hierarchical Bayesian model, associates each document with a probability distribution over topics and subsequently each topic with a probability distribution over words[41].

[40] Blei et al., "Latent Dirichlet Allocation."
[41] Liu et al., "Bayesian Parameter Estimation in LDA."

2.2.4. Latent Dirichlet Allocation

While PLSA is a useful step toward probabilistic modeling of text, it is incomplete in that it provides no probabilistic model at the level of documents: where each document is represented as a list of numbers (the mixing proportions for topics), and there is no generative probabilistic model for these numbers [42]. This leads to two major limitations for PLSA:

1. The number of parameters in the model grows linearly with the size of the corpus, which leads to severe problems with overfitting,

2. It is not clear how to assign a probability to a document outside of the training set.

To overcome these issues a novel approach called Latent Dirichlet Allocation is introduced as a genuinely generative probabilistic model of a corpus [43]. The basic idea is similar to PLSA documents are represented as random mixtures over latent (hidden) topics, where each topic is a distribution over words. However, LDA adds another level of distribution to the two-leveled PLSA: a Dirichlet distribution which acts as a distribution over distributions [44].

The generative process of LDA for each document *w* in a corpus *D*[45] is as follows (Figure 2.5):

1. Choose $N \sim \text{Poisson}(\xi)$.

2. Choose $q \sim \text{Dir}(\alpha)$.

[42] Blei et al., “Latent Dirichlet Allocation.”
[43] Blei et al.
[44] Thomas P. Minka, “Estimating a Dirichlet Distribution.”
[45] Blei et al., “Latent Dirichlet Allocation.”

3. For each of the *N* number of words w_n:

   (a) Choose a topic z_n ~ Multinomial(θ).

Choose a word w_n from p($w_{n\,|}$ z_n, β), a multinomial probability

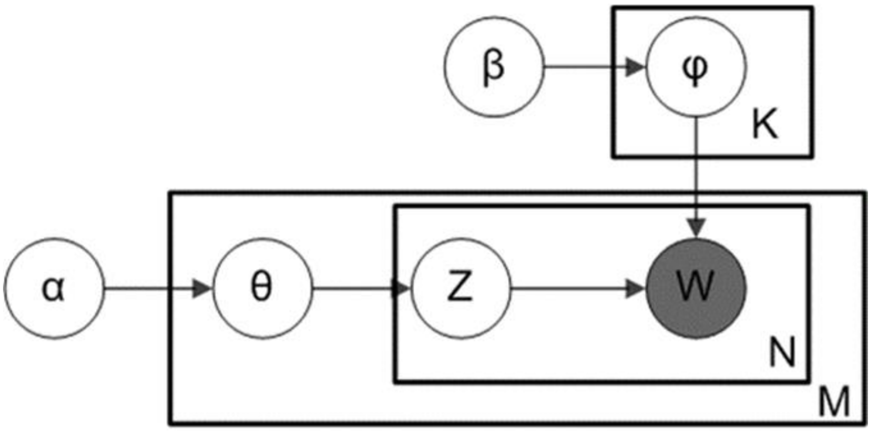


Figure 2.5. The Plate Diagram of LDA
M is number of documents, N is the number words, K is number of topics[46].

LDA has become one of the most influenced, cited and used topic modeling since its introduction. Besides its' core implementation, various other topic modeling methods have derived from its concept.

[46] Blei et al.

2.2.5. Structural Topic Modelling

Structural Topic Modeling (STM) was developed as a combination and extension of the existing models to provide a general way to incorporate corpus structure or document metadata into the standard topic model[47]. Building-off of the tradition of generative topic models, such as the Latent Dirichlet Allocation (LDA) and Correlated Topic Model (CTM)[48] , the STM's key innovation is that it permits users to incorporate metadata, defined as information about each document, into the topic model[49] (Figure 2.6)

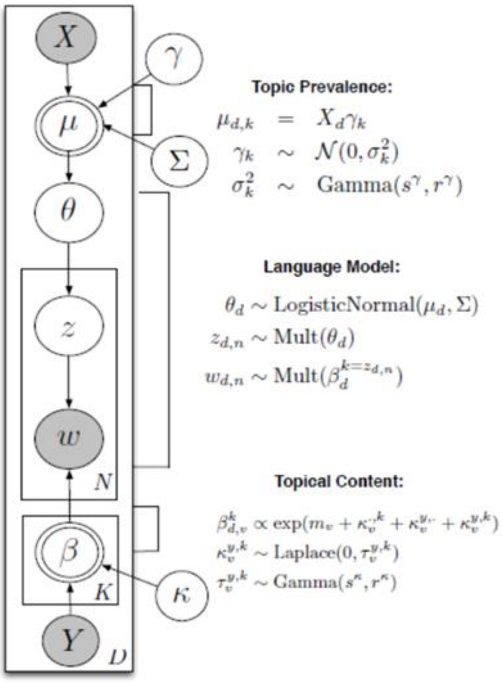

X
γ
μ
Σ
θ
z
w
N
β
K
κ
Y
D
Topic Prevalence:
$\mu_{d,k} = X_d\gamma_k$
$\gamma_k \sim \mathcal{N}(0,\sigma_k^2)$
$\sigma_k^2 \sim \text{Gamma}(s^\gamma, r^\gamma)$
Language Model:
$\theta_d \sim \text{LogisticNormal}(\mu_d, \Sigma)$
$z_{d,n} \sim \text{Mult}(\theta_d)$
$w_{d,n} \sim \text{Mult}(\beta_d^{k=z_{d,n}})$
Topical Content:
$\beta_{d,v}^k \propto \exp(m_v + \kappa_v^{\cdot,k} + \kappa_v^{y,\cdot} + \kappa_v^{y,k})$
$\kappa_v^{y,k} \sim \text{Laplace}(0, \tau_v^{y,k})$
$\tau_v^{y,k} \sim \text{Gamma}(s^\kappa, r^\kappa)$

Figure 2.6. . Plate Diagram for the Structural Topic Model.
"Topic Prevalence" refers to the proportion of document devoted to a given topic and "Topical Content" refers to the rate of word use within a given topic[50].

[47] Roberts et al., "The Structural Topic Model and Applied Social Science."
[48] Lafferty and Blei, "Correlated Topic Models."
[49] Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley, "Stm:R Package for Structural Topic Models."
[50] Roberts et al., "The Structural Topic Model and Applied Social Science."

Like the LDA model the STM is a generative model. The generative process for each document (indexed by *d*) can be summarized as[51]:

1. Draw the document-level attention to each topic from a logistic-normal GLM based on document covariates X_d.
   $\theta_d \,|\, X_{d\gamma}, \Sigma \sim \text{LogisticNormal}(\mu = X_{d\gamma}, \Sigma)$

2. Form the document-specific distribution over words representing each topic (*k*) using the baseline word distribution (*m*), the topic-specific deviation κ_k, the covariate group deviation κ_g and the interaction between the two κ_i.
   $\beta_{d,k} \propto \exp(m + \kappa_k + \kappa_{gd} + \kappa_i = (k_{gd}))$

3. For each word in the document, ($n \in 1, \ldots,$ Nd):

   a. Draw word's topic assignment based on the document-specific distribution over topics.
      $z_{d,n} \,|\, \theta_d \sim \text{Multinomial}(\theta)$

   b. Conditional on the topic chosen, draw an observed word from that topic. $w_{d,n} \,|\, z_{d,n}, \beta_{d,k=z} \sim \text{Multinomial}(\beta_{d,k=z})$

There are three critical differences in the STM as compared to the LDA model[52]:

1. Topics can be correlated

2. Each document has its prior distribution over topics, defined by covariate *X* rather than sharing a global mean

3. Word use within a topic can vary by covariate *U*. These additional covariates provide a way of "structuring" the prior distributions in the topic model,

[51] Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley, "Stm:R Package for Structural Topic Models."
[52] Roberts et al., "Structural Topic Models for Open-Ended Survey Responses."

injecting valuable information into the inference procedure.

In brief, the STM process assigns words from the documents with their level of interest document information (e.g. author, source, date, etc.) to fit to a selected number of topics by an Expectation-Maximization (EM) algorithm[53] (Figure 2.7).

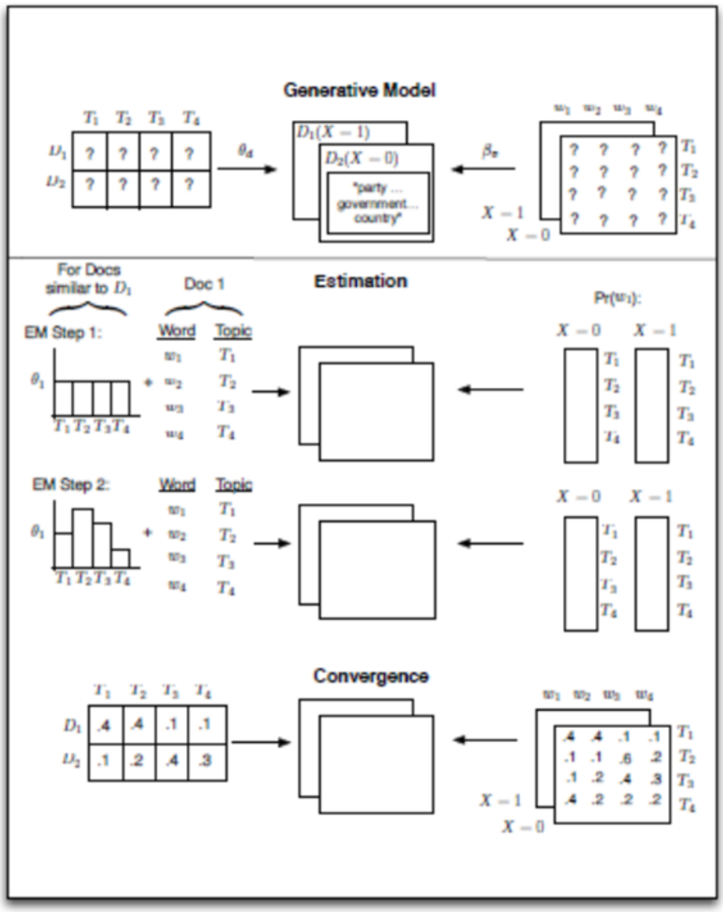


Figure 2.7. Heuristic description of generative process and estimation of STM.[54].

[53] Blei et al., "Latent Dirichlet Allocation."

[54] Roberts ME, Stewart BM, Tingley D. stm: R Package for Structural Topic Models. Journal of Statistical Software. 2017.

2.3. Related Works

Exploratory data analysis, text mining and topic modeling are very sparse fields of interest within the architectural domain. Thus, any related works sharing a similar focus subject is hard to be found. Nevertheless, there are a wide variety of studies with similar methods such as research, applications or textbooks.

K. Chai and X. Xiao (2012)[55] study explores the evolution and future trends of design research through a series of bibliometric studies of the articles released at the Design Studies Journal, as it is one of the leading international and interdisciplinary academic journal focused on developing and understanding of design processes. Employing a bibliometrics and network analysis, the paper analyses citations and co-citations from Design Studies. The authors utilized the citations of every article as a variable to find the trends of citation index in design research studies.

Y. Acar (2017) [56] dissertation focused on the field of urban design and as a case of knowledge discovery of urban design, audits the Turkish academia of the same discipline. The study is conducted by a series of analysis of the documents regarding the urban design that has been produced for the last three decades. The study is a textual analysis with a critical reading of the collection of documents; the author used textual analysis tools to explore and visualize the corpus. The visualizations were used as maps of knowledge and these maps have been evaluated discursively. The author then brought discussions and findings based on these maps. Acar's dissertation is highly overlapping with this study in terms of research motives, however its discursive evaluation gradually differs with the statistical inference of this study.

E. D'Avanzo et al (2008) [57] study presents text mining as a tool that provides human

[55] Chai and Xiao, "Understanding Design Research: A Bibliometric Analysis of Design Studies (1996-2010)."
[56] Acar, "Atlas of Urban Design: Textual Analysis and Mapping of Production of Knowledge in Turkish Context."
[57] D'Avanzo et al., "Where Does Text Mining Meet Knowledge Management? A Case Study."

coherent textual summaries that have been generated from big organizations documents repositories. The aim of the study was to demonstrate the feasible usage of basic text mining tools that ease searching and acquiring relevant information from corporate databases.

M. E. Roberts et al [58] represent their STM method to the analysis of open-ended surveys. The study aims to bring a semi-automated alternative to the commonly used human-coding procedure for analyzing open-end response of surveys or interviews. The article provides a general framework of STM and a benchmark with typical LDA modeling performance. Additionally, the study compares STM results with manual compiled results on a selected open-end survey.

E. Tvinnereim and K. Fløttum (2015) [59] study argues the cruciality of opinions of people for action of decision-makers on climate change. Within this context they have conducted an open-ended survey to a group with different age, gender and education attributes. The survey responses have been categorized under several topics by the application of the STM methodology, which provided a comprehensive evaluation and analysis process.

C. Miller et al [60] study analysis the trends and topics of the Bldg-sim email list, which was established at 1999 and became a major platform for the building simulation proffesionals. The study is conducted by a LSA method with the word counting technique that calculates the weight of words by a mixture of document and corpus scale. By doing so, the researches have generated 6 latent topics for each year and cross-examined them under the categories of systems and softwares, organizations and companies, and the trends of superusers agendum.

[58] Roberts et al., "Structural Topic Models for Open-Ended Survey Responses."

[59] Tvinnereim and Fløttum, "Explaining Topic Prevalence in Answers to Open-Ended Survey Questions about Climate Change."

[60] Miller, Quintana, and Glazer, "Twenty Years of Building Simulation Trends : Text Mining and Topic Modeling of the Bldg-Sim Email List Archive Topic Modeling of the Bldg-Sim Email List Archive."

2.4. Tools

The main digital tool for this study is R programming language for statistical computing. R is a language and environment for statistical computing and graphics[61]. As the user interface the official RStudio[62] will be used.

Although, the base R library contain most of the necessary functions; many external libraries will be used: rvest[63], tidyverse[64], tidytext[65], primary[66], tm[67], dplyr[68], ggplot2[69], purrr[70], quanteda[71], stm[72], corrplot[73]. These libraries all have specialized functionality for different purposes such as web-scraping, data structuring, data manipulating, statistical computation and data visualization. These tools are all open-source and obtain through The Comprehensive R Archive Network (CRAN)[74].

[61] https://www.r-project.org/about.html

[62] https://www.rstudio.com/

[63] Wickham, H. https://cran.r-project.org/web/packages/rvest/index.html

[64] Wickham, H. https://cran.r-project.org/web/packages/tidyverse/index.html

[65] Siege, J. https://cran.r-project.org/web/packages/tidytext/index.html

[66] Wickham, H. https://cran.r-project.org/web/packages/stringr/index.html

[67] Feinerer, I., Hornik, K. https://cran.r-project.org/web/packages/tm/index.html

[68] Wickham, H. https://cran.r-project.org/web/packages/dplyr/index.html

[69] Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takashi, K., Wilke, C., Woo, K. https://cran.r-project.org/web/packages/ggplot2/index.html

[70] Henry, L, Wickham, H. https://cran.r-project.org/web/packages/purrr/index.html

[71] Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A.,Müller,S., Matsuo, A., Lua, J.W., Perry, O.P., Kuha, J., Lauderdale, B., Lowe, W. https://cran.r-project.org/web/packages/quanteda/index.html

[72] Roberts, M., Stewart, B., Tingley, D., Benoit, K. https://cran.r-project.org/web/packages/stm/index.html

[73] Wei, T., Simko, V., https://cran.r-project.org/web/packages/corrplot/index.html

[74] https://cran.r-project.org/

CHAPTER 3

CORPUS

This chapter presents the materials used in this study, their features and assumptions regarding the selection. The Master of Architecture theses are collected from the Thesis Center of the Council of Higher Education [75], under the section of Department of Architecture (Mimarlık Anabilim Dalı) and the field of Architecture (Mimarlık Bilim Dalı).

Regarding the indexing of the document at the data source, two major problems occur. First, entries of the theses prior to 2003 has different indexing for various universities. Hence, some universities either have not any thesis submitted for certain years or related documents are mis indexed and cannot be reached through index searching. Furthermore, most of these entries lack an English abstract, or has no abstract at all. Secondly, although some theses are indexed as “Architecture”; they’re submitted under research programs such as computer engineering, information technologies or various disciplines of fine arts. Therefore, to retain the content quality of the corpus the dataset is manually examined and cleaned from empty documents. Universities with less than 5 documents were also omitted due to their sparsity.

The corpus contains 4370 documents, submitted from 39 universities and spanning from 2003 to 2018 (). The main attributes collected for each document is the unique identifier number (id) given by YÖK, the author name, the submission year, the university that it was submitted, title and the abstract in English. However, the author names were omitted from the attribute set since the individuals are not the subject of interest for this study.

[75] https://tez.yok.gov.tr/UlusalTezMerkezi/giris.jsp

Table 3.1. *Universities and number of thesis*

No	University	Tag	No. of Thesis
1	Anadolu University	anadolu	25
2	Istanbul Arel University	arel	17
3	Balıkesir University	balikesir	12
4	Bahcesehir University	bau	47
5	Beykent University	beykent	50
6	Cukurova University	cukurova	45
7	Dicle University	deu	192
8	Dokuz Eylül University	dicle	45
9	Erciyes University	erciyes	34
10	Gazi University	gazi	348
11	Gebze Technical University	gtu	15
12	Halic University	halic	84
13	Istanbul Aydın University	iau	31
14	Izmir University of Economics	ieu	7
15	Istanbul Kültür University	iku	55
16	Istanbul Technical University	itu	1391
17	Izmir Institute of Technology	iyte	38
18	Kadir Has University	kadir	8
19	Hasan Kalyoncu University	kalyoncu	9
20	Karabük University	karabuk	23
21	Kocaeli University	kocaeli	18
22	KTO Karatay University	kto	11
23	Karadeniz Technical University	ktu	91
24	Maltepe University	maltepe	42
25	Mardin Artuklu University	mau	12
26	Middle East Technical University	metu	502
27	Mimar Sinan Fine Arts University	msgu	199
28	Necmettin Erbakan University	nbu	9
29	Eskisehir Osmangazi University	ogu	22
30	Süleyman Demirel University	sdu	36
31	Selcuk University	selcuk	109
32	University of Econ. & Technology	tobb	8
33	Toros University	toros	13
34	Trakya University	trakya	68
35	Uludağ University	uludag	46
36	Yasar University	yasar	21
37	Yeditepe University	yeditepe	17
38	Yıldız Technical University	ytu	654
39	Sebahattin Zaim University	zaim	16

3.1. Assumptions

The main assumption regarding the materials selected for this study is that the abstract of each thesis clearly express the content with the most significant words, this is an expected quality. Although ideally, an abstract of a study represents these attributes at a higher quality, it is debatable that this is applicable for every case. Some abstracts might be written poorly in terms of quality and might even be misleading regarding knowledge discovery.

The alternate option is to text mine the full-text of each thesis, which leads to severe problems in its own right. One of the major drawbacks is the issue of accessibility and eligibility of full texts: some theses are restricted to access by its author thus shrinks the corpus size and almost all theses are published in a portable text format (PDF) which limits the text mining ability. While a PDF extended document can be converted to a raw text file (TXT), which is the ideal state of mining it; a full-text thesis contains figures, footnotes, formulas, tables and references which require filtering and furthermore preprocessing actions. If not, these contents can dominate the dataset without any semantic characteristics. In addition to the immense computational power and workforce that will be required, every thesis presents existing and commonly accepted information in multiple chapters or sections. These sections comprise less significant knowledge discovery and can lead to erroneous results. Eventually, there is a tradeoff regarding the selection of the abstract or the full-text[76], and with consideration of all these aspects, the abstracts of each thesis are preferred for representing the knowledge that its corresponding research contains.

[76] Westergaard D, Stærfeldt HH, Tønsberg C, Jensen LJ, Brunak S (2018) A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. PLOS Computational Biology 14(2): e1005962. https://doi.org/10.1371/journal.pcbi.1005962

3.2. Overview of the Corpus

Regarding the distribution of the documents over the years, the drastic increase of the thesis submittal, thus the increase of textual data is evident (Figure 3.1). The rapid increases in 2004 and 2010 are caused by the student amnesty offered by the government to students.

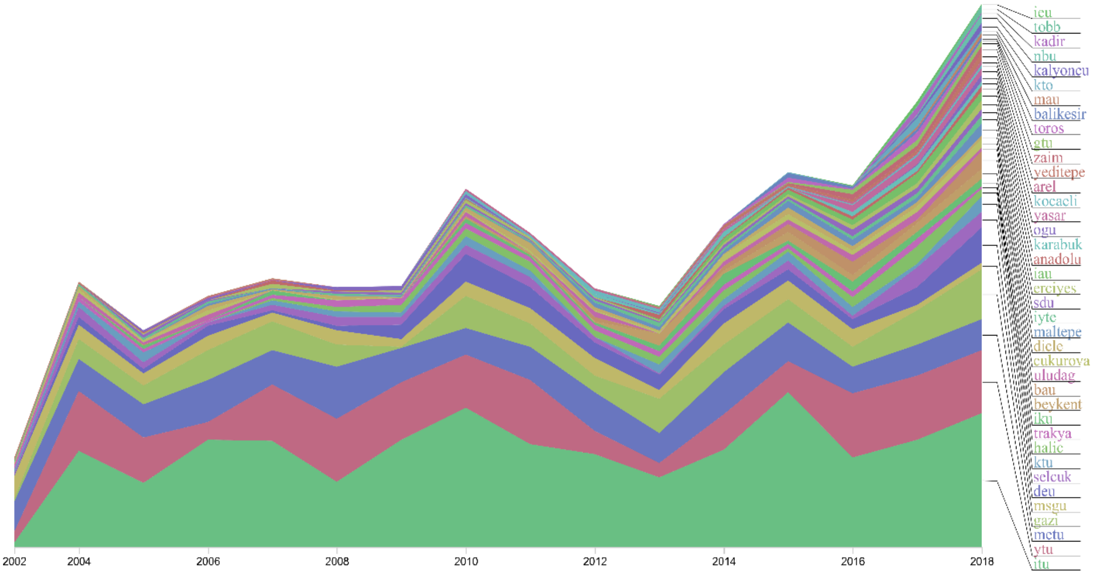


Figure 3.1. Stacked Chart of Corpus over Years. Universities ordered by number of documents.

Additionally, there is also a steady growth of the corpus from 2013 due to the increasing amount of private funded universities and the newly established departments of architecture and research programs. This pattern also fits with the overall statistics for all fields of research provided by YÖK [77].

This initial overview of the data set indicates to the two aspects of the corpus: a group of universities in terms of consistency and proportion; and a breaking point on the timeline regarding the corpus size expansion.

[77] https://tez.yok.gov.tr/UlusalTezMerkezi/IstatistikiBilgiler?islem=1

Despite the increasing amount of novel departments, more than %65 of the corpus is produced by 5 universities: ITU, YTU, METU, GAZI and MSGU. These universities have a long history of the department of architecture with various degrees. Especially ITU, contributes to %31 of the corpus alone, is always the proportionally most dominant university. These 5 universities are all public funded and located in Ankara and Istanbul, most populated and economically improved cities of Turkey. They have an extended historical consistency of architectural research programs which spans for more than 40 years and they have been the pioneers of architectural education and discussions of Turkey.

The following high proportional universities of CUKUROVA, DEU, DICLE, KTU, SELCUK and TRAKYA are all public funded and located at cities other than Ankara and Istanbul. These universities have relatively old programs, where all of them has submitted documents from 2003. However, their proportion of document contribution is notably lower than the first group, only %12 of the corpus contains documents from these universities.

Besides theses 11 universities, the remaining 28 has similar characteristics in terms of time span and contribution proportion. Most of these universities started their architectural research programs after 2010, where their immediate impact can be observed at Figure 3.1. With their contributions the corpus size has increased drastically over the last five years of the timeline. However, beside their resemblance of timing, they differ in terms of funding and location. While the public funded 12 universities are mostly located at mid-level populated cities, the private funded are located at major cities. The location and funding differences may impact their research agendas due to external conditions of education structure, facilities and the culture of the integrated city.

This categorization will be used to determine the effects of different factors regarding the conducting of architectural research. Therefore, all universities and texts related to them are gathered under four groups (Table 3.2).

Table 3.2. Universities gathered under Groups

GROUP	Universities
GROUP-A	gazi, itu, metu, msgu, ytu
GROUP-B	cukurova, deu, dicle, ktu, selcuk, trakya
GROUP-C	anadolu, balikesir, erciyes, gtu, iyte, karabuk, kocaeli, mau, nbu, ogu, sdu, uludag
GROUP-D	arel, bau, beykent, halic, iau, ieu, iku, kadir, kalyoncu, kto, maltepe, tobb, toros, yasar, yeditepe, zaim

3.3. Structuring

The initial step towards the data analysis is to structure the documents from their raw text extension. In this initial phase, the documents are tabulated to evaluate their content and quality briefly. By quality, the documents required to be in English language and contain valid attributes. The data is structured in tabular format: which is a standard way of mapping the meaning of a dataset to its structure[78]. Thus, in this tabular form:

- Each variable forms a column.
- Each observation forms a row.

A straightforward construction of the structure is presented at Table 3.3, where the observations (rows) are the theses, and the variables are ID, title, year, university, abstract, and binary attributes for every group; where the variable is either included (1) or not (0) to observed group (Table 3.1).

[78] Wickham, "Tidy Data."

Table 3.3. Section of the Structured Corpus in tabular format

ID	*Year*	*Uni.*	*Title*	*Abstrat*
93094	2000	metu	Architecture and metaphor: An inquiry into the virtues of metaphorical expressions in architecture	This thesis explores into the conceptual merits of 'metaphor' as a mode of expression within a framework of its relationship with architecture[…]
93693	2000	ytu	The Effects of the development of computer technology on architectural material technology	this research work's purpose is to point out the effects of the development of computer technology on material concept which is expressed on architectural work[..]
……….	……	……	………………………………………….	…………………………………………
543100	2018	iyte	Analyses of four urban squares in Izmir according to the leading urban design literature	The squares are the significant elements of the urban public spaces that provide the users different experience in the city, activate their city life [..]
543960	2018	itu	Changing process of housing typologies in Istanbul	the housing that emerges to the built environment as a result of people's need for housing is a constantly changing concept. istanbul is the city where the [..]

3.4. Preprocessing

To further proceed with the structural topic modeling of the corpus a preparation of the data is required: it is a must for the transformation of the texts from a -human-language to a machine-readable and enumerable format. This enumerable format will consist of the vocabulary for the STM.

To extract meaningful knowledge through a computational process, the vocabulary must be eliminated from irrelevant and insignificant contents. This elimination is necessary to improve the abstract quality of the input and eventually affect the quality of the output. While this phase is a linear process, constant observations of the data are required for quality control. These observations can be done through the visualization or the numeric outputs.

3.4.1. Tokenization

Tokenization is the process of unitizing a string set, in this case, the texts of documents. Regarding the unit of interest; a token can be a word, an adjacent sequence of words (n-gram), a phrase or even a whole paragraph. While phrases or paragraphs are rarely used as tokens, n-grams are especially used for query searches regarding proper nouns (e.g. authors, institutions, organizations) or noun phrases (special terms). In this study, a token refers to a single word where it is separated from each other by either whitespace, punctuation marks or line breaks.

Furthermore, in order to enumerate properly, the following steps have applied to the text variable of the data:

1. Tokens are lowercased. Therefore, same words written in uppercase or with capital letters are considered as the same token: e.g. the words Architecture, ARCHITECTURE and -typo- aRchitecTure are all transformed to architecture.

2. All punctuations are removed. Hence, words such as it’s and its’ are considered as its.
3. All alphanumerical and numeric character containing words are removed: e.g. 1968 and big1.
4. Words with two or less characters are removed: e.g. it, in, a, on, etc.

The term data trimming refers to the process of feature selection by the manipulation of the dataset by either removing or aggregating insignificant, erroneous, extreme (outlier) or abnormal values. While initial trimming was done during the preliminary tokenization, further steps are required: these steps are to remove stop-words, statistically insignificant words or the combination of similar knowledge value words by stemming.

3.4.2. Stop-words

In general, stop-words refer to the most common words in a language [79]: for English, the word The is the most common word in all linguistic cases. Although there is no definite metric to claim any word as a stop-word, studies like Zipf’s Law [80] that examines the frequency of word usage in a language. While mostly observed in the English language, this law reveals that the most common word in regular usage of language occurs twice as much as the second most frequent word and triple from the third word, and so on. Eventually, a word in any rank is proportionally less occurring to the first word: e.g., the 50th most frequent word in a text is 1/50 times less frequent regarding the most frequent one. Zipf’s Law can be encountered on various occasions: from the speech of an infant less than three years old to a Charles Dickens novel.

[79] Aggarwal and Zhai, *Min. Text Data*.
[80] Named after George Kingsley Zipf

Although it is still deemed to be an empirical law, the globally agreed stop-words of the English languages are validated with this law. In this study, the pre-defined stop-words list created by Dr. Martin Porter [81] is used to improve to content quality of the vocabulary and to increase the computational efficiency.

Additionally; a custom list of stop-words is defined to eliminate words that frequently appear in theses abstracts, but insignificant in terms of knowledge discovery. This list contains words that usually address to the thesis itself, define its structure or common verbs to describe the research process and method. These words are:

- abstract, aim, analyze, chapter, conclusion, discuss, evaluate, examine, example, explain, find, focus, introduction, investigate, mention, method, page, research, scope, study, subject, thesis.

Although the significance of any token (word) is a disputable subject, descriptions related to the research methods of the subject documents does not come into focus of this study. Therefore, by eliminating these manually selected stop-words, the textual information is further relaxed from insignificant contents

3.4.3. Stemming

Stemming is the process of reducing inflected words to their base (root) form. In this process, different forms of the same words are unified under a single word. The stem form of words does not necessarily have to be identical to the root of the word where it can be enough to map the related words to the same stem, even though it is a not morphologically valid root or word.

This study utilizes Porter's stemming algorithm[82] to aggregate tokens in the model vocabulary. It is one of the most preferred algorithm for the English language that stems words by stripping suffixes: i.e. the words economy, economic, economics and

[81] http://snowball.tartarus.org/credits.html

[82] Porter, "An Algorithm for Suffix Stripping."

economical can all be stemmed as econom. By stemming tokens, the vocabulary is gradually reduced, but has an improved representation and quality of the information to be processed. Stemming, also further improves the computational performance of the model generating process.

3.4.4. Trimming

As a final process of vocabulary construction, the stemmed words will be filtered regarding their occurrence on document level. This filtration can be variant for different cases, but in this study words that appear less than five per thousand (0.05%) and more than ninety percent (90%) of the documents are determined as insignificant and therefore removed from the vocabulary. With all the preprocessing phases complete, a set of 2014 words has been generated.

3.5. Data

By applying the preprocessing steps to the corpus, it is now enumerated under a Document-Term Matrix. This DTM will be the data basis for the topic model with a metadata as a tabular form (Table 3.3). As a crucial remark, although the ordering of the terms or documents in the DTM does not affect the statistical outputs; however, to preserve the relation of the data and the metadata, the DTM will not be altered in terms of arrangement of rows or columns and in terms of sparsity reduced which is a typical method at LSA.

Table 3.4. The preprocessing stages applied on a sample text.[83]

Raw Text	This study tries to discuss the relations between the critical culture of architecture and the concept of virtuality, which gradually emerged with the technologies of virtuality, and are intensifying in effect by the end of the 20th century. This thesis departs from the idea that technology, especially technologies of virtuality, is a cultural construct and affects architecture in various ways. The definition of the term 'virtuality' is narrowed to imply any kind of experience generated by technological constructs, that transports the self out of physical reality. In the third chapter, representations in the popular cultural narrations are studied in order to derive concepts that will help to construe the changing notions in architecture and lay the theoretical framework, under three categories: time, space and culture. Consequently, an analysis is made in the context of urban environment and dynamics, with the same structure, and the ways in which these categories are reconceptualized in the contemporary cities are discussed. As a result, this study tries to examine the shifting notions in the field of architecture with the involvement of virtuality.
Tokens (lowercase + stopwords)	affects analysis architecture architecture architecture architecture categories categories century changing cities concept concepts construct constructs construe contemporary context critical cultural cultural culture culture definition departs derive dynamics effect emerged environment experience field framework generated gradually idea intensifying involvement narrations narrowed notions notions physical popular reality reconceptualized relations representations self shifting space structure studied technological technologies technologies technology theoretical time transports tries urban virtuality virtuality virtuality virtuality virtuality
Tokens (stem + trim)	analysi architectur categori centuri chang citi concept construct contemporari context critic cultur definit depart deriv dynam emerg environ experi field framework generat gradual involv narrat notion physic popular realiti relat represent result shift space structur studi technolog term theoret time transport urban virtual

[83] Gürsel, İpek. "The Limits of virtuality in space: The transformative relations between immateriality and architecture". 2002

CHAPTER 4

TOPIC MODELING

The corpus of 4370 documents and the corresponding vocabulary of 2014 words are stored under a DTM and prepared to be processed as topic model. Regarding the concept of statistical topic models, topics are defined as probability distributions over a vocabulary of words that represent semantic themes. With the removal of insignificant content, a so-called clean vocabulary has been generated from the dataset on prior phases. The initial tokens generating a document has been reconstructed by this vocabulary.

The core concept of STM is to distribute the vocabulary, as an DTM, over user chosen number of topics, while also keeping each documents relation with the metadata to estimate effects of the user defined attributes. These attributes are the covariates of the model, which in statistics refers to a variable that changes in predictable way (i.e. regression) and can be predict of the outcome. In this case, the documents and terms are the independent variables and the model outputs are dependent variables. Therefore, since the two independent variables lack to provide further prediction over the dependent variables, the covariates become the regressors to observe further quantities of interest.

Beside this core concept, STM is an extension of LDA which also introduces different model initialization and fitness techniques. In this study, the workflow of the STM authors -and contributors- are emphasized regarding these subject.[84] This workflow is reduced to three steps: inputs, processing and outputs (Figure 4.1). In this chapter, the input selection and arrangement are broadly discussed, and the outputs are presented.

[84] Roberts et al., “Structural Topic Models for Open-Ended Survey Responses.”

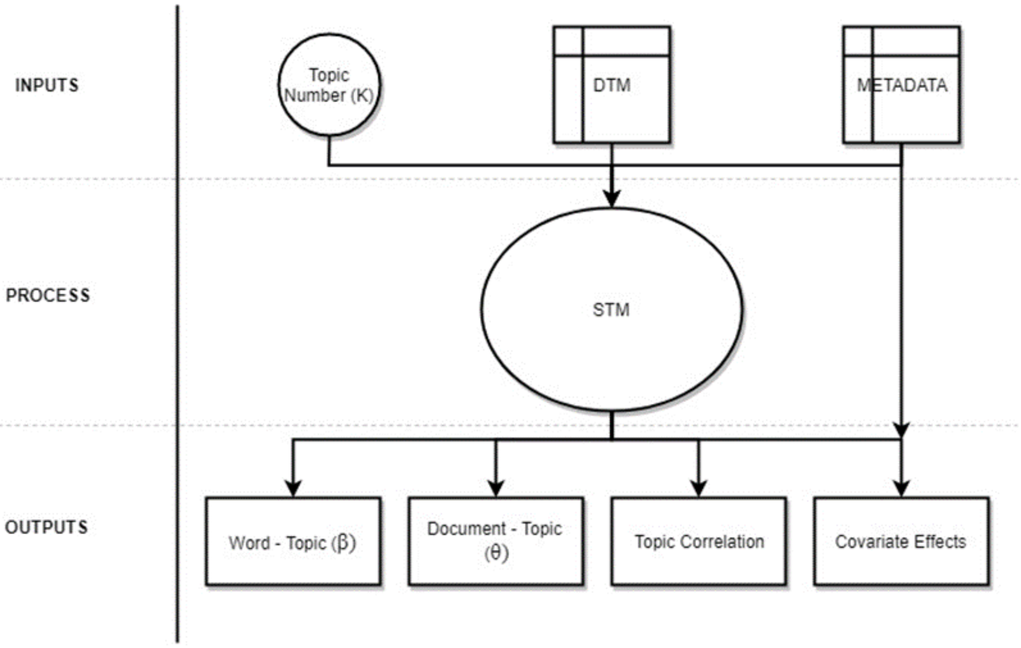


Figure 4.1. Structural Topic Modeling workflow

4.1. Inputs

The steps of constructing a structural topic model are:

1. Introducing the data and metadata
2. Introducing covariates through metadata
3. Selection of number of topics (K)

The first two steps are covered during the earlier phases of corpus organizing and structuring. As a key component for STM, the topic number (K) is a subject of detailed investigation and exploration.

4.1.1. Topic Number Selection

In terms of model selection, the authors of the STM initially state that any number of topics (without exceeding the size of vocabulary) can be intuitively selected depending on the nature of document content and the aim of the research [85]. Although they are various studies concerning the topic number selection elevating different metrics and approaches, the framework emphasized by the authors of STM will be followed to further proceed with the selection.

This selection also designates the quality of each topic and overall the model. This quality is based on two metrics [86]:

1. A topic containing high-probability words that have a high co-occurrence among their source documents.
2. A topic, that words with the highest probabilities in the topic are expected to appear less or none in other topics.

In accordance with these quality aspects, the topic number will be decided.

4.1.1.1. Exclusivity

Exclusivity is measured by the words with high probability under a certain topic have relatively low probabilities under other topics. An exclusive topic is most likely to contain a latent concept that drastically differs from the other topics.

Although, exclusivity is a desired quality for a topic; quantitating it alone has a misleading feature. By generating words solely based on their probabilities under a certain topic, neglects their co-occurrence and semantics regarding their corresponding text. High probability words of a topic must also constitute a meaningful a subset rather than a probability-wise exclusivity.

[85] Roberts et al.

[86] Roberts et al.

4.1.1.2. Semantic Coherence

Semantic coherence measures the topic itself, regardless of the affiliation of the contained words towards other topics. It is an internal value that aids to evaluate the consistency and semantic integrity of a topic, where the co-occurrence frequency of the words are calculated and high frequency is a desirable outcome.

With these metrics comes a tradeoff of topic selection: although a topic that is both cohesive and exclusive assures high quality, a decision should be made regarding one aspect having a greater impact than the other.

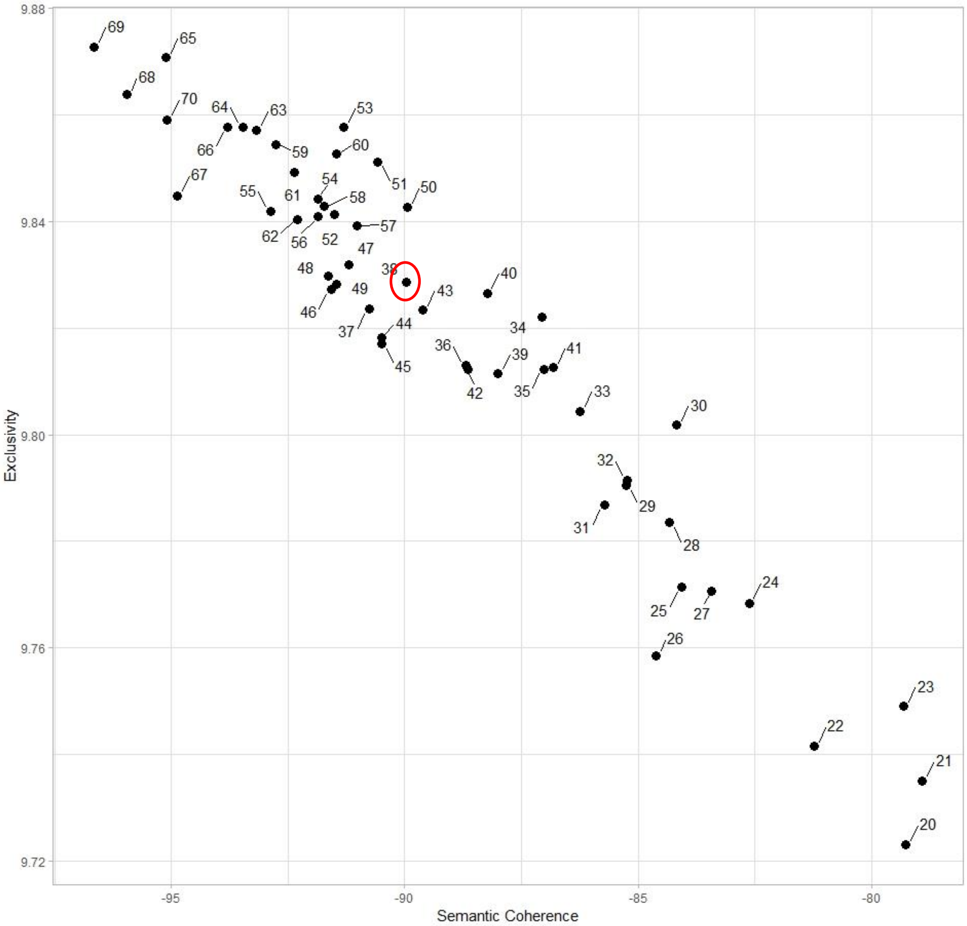


Figure 4.2. Semantic Coherence – Exclusivity score dispersion for topic models with number of topics from 20 to 70. Selected topic model with 38 topics is encircled with red.

According to these two metrics, a model generation is initiated to decide the topic number. 50 models have been converged with topics spanning from 20 to 70 (Figure 4.2).

The results indicate that an optimum model quality is between 34 to 50 topics. Further screening of the models between these values demonstrates that objectively the highest quality model contains 38 topics.

4.1.2. Covariates

The metadata attributes are years as a continuous numeric variable (2003 to 2018) and the binary variables of groups (GROUP-A, GROUP-B, GROUP-C and GROUP-D). These attributes constitute the covariates of the topics of the STM model. These covariates will be the regression variables to calculate the topic-proportion expectations for interested document characteristics.

4.2. Outputs

The quantities of interest for the STM is as follows [87]:

1. Proportion of words over topics (β)
2. Proportion of topics over documents (θ)
3. Document covariate effects
4. Topic correlations

With 43 iterations a STM with 38 topics, 4370 documents and a 2014-word vocabulary is converged. Furthermore, the metadata content is introduced to estimate the expected proportional effects of the covariates.

[87] Roberts et al.

4.2.1. Topics

The crucial output of the model is the topics with meaningful clustering of words and documents. Initially, the so-called topic quality is checked through an exclusivity and semantic-coherence scoring (Figure 4.3). The results demonstrate the dispersion of topics; although, there are some topics with less scores on each metric, i.e. Topic-38 is too exclusive and less coherent while Topic-31 is less exclusive and more coherent, the overall dispersion of the topics is aligned at optimum scale.

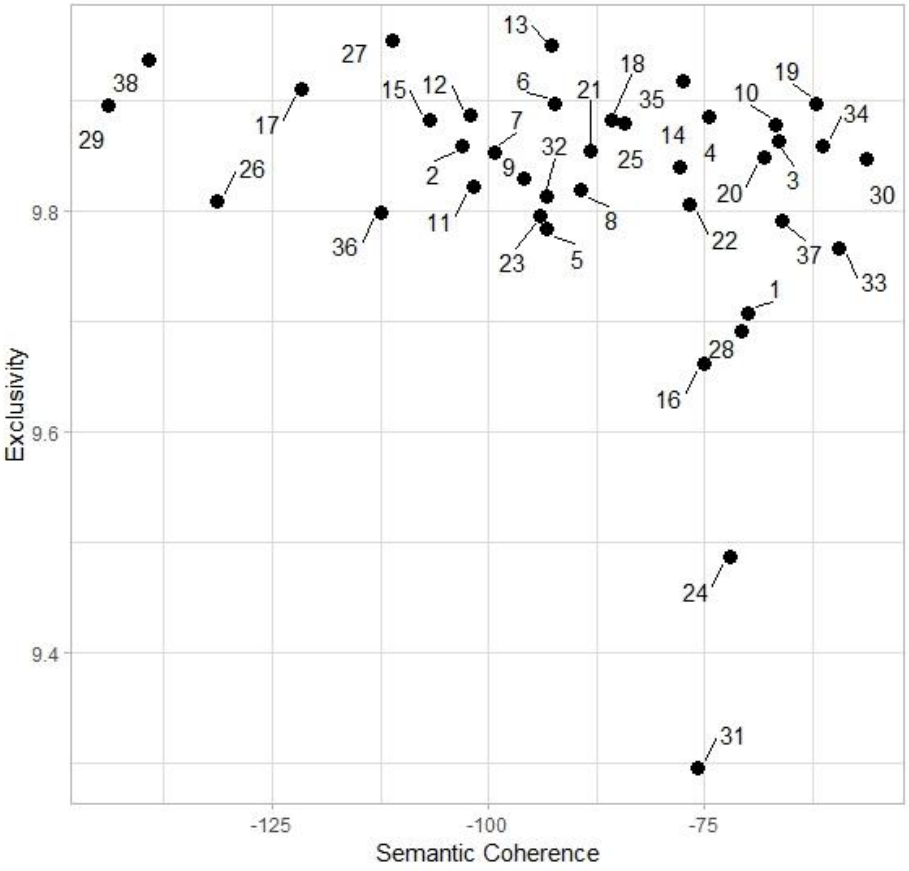


Figure 4.3. Exclusivity-Semantic Coherence for each topic of the STM.

Furthermore, the topics were manually evaluated in-depth according to their highest word probabilities (β) and document affiliations (θ). At this point, the model is deemed satisfactory regarding to the revealed latent topics of corpus and according to the consistency of these topics in terms of representing their relation with their affiliated documents (Table 4.1).

Table 4.1. Table of Topics. Each topic is represented with top 10 highest probability keywords and top 5 highest probability documents.

Topic	*Keywords (top 10)*	*Related Thesis Titles (top 5)*
1	model, data, base, bim, develop, system, propos, user, softwar, technolog	¬ Human crowd simulation as an evaluation tool for generative temporary site organization ¬ The integration of computational design and building information modelling the research of new possibilties with dynamo ¬ Electronic procurement systems and building information modeling integration in construction sector: A case study ¬ A novel computational approach in quantity take-off to support early design estimations ¬ Adaptable network generator (ANG): A generative system proposal with respect to temporal space design
2	architectur, architect, field, cinema, futur, scienc, film, techniqu, subject, fiction	¬ Intersections: Architecture and photography in Victorian Britain ¬ Architecture in 3D animation films ¬ Reconstruction of architectural image in science fiction cinema: A case study on New York ¬ The cinematic represantation of architectural space described in literary works ¬ Evaluation of future artificial environments through science fiction and its influence field

Table 4.1 continued

3	tradit, settlement, region, cultur, plan, villag, rural, local, characterist, locat	¬ Analayzing the traditional settlement of Bozcaada with conservation issues ¬ Development and conservation of cultural properties in rural areas of Eastern Black Sea region: A case study in Karacakaya Village ¬ Traditional residence architecture in Eski Foça city texture and usage problems ¬ The plan of the house of traditional Rize Ikizdere typology investigation ¬ Comparison of two different type of urban fabrics in Osmaneli registered urban historic site and rehabilitation proposals
4	design, process, approach, technolog, tool, digit, develop, product, flexibl, knowledg	¬ Interaction between design research with digital design and production technologies in architecture ¬ A research on architectural design techniques and manufacturing processes in the digital age ¬ Design and manufacturing process in digital architecture ¬ Document inventory system for Mardin ¬ Digital design tools: A research in the context of their roles in design process
5	turkey, countri, regul, standard, public, competit, plan, intern, law, develop	¬ The evaluation of tender laws applied at building trade in Turkey ¬ The tendering procedures of multilateral banks and organizations who finance international construction projects ¬ Public Procurement Authority's working principle's analysis's and public procurement system in Turkey ¬ Closeout procedures in construction administration and recommendations to KIK judgements ¬ Building regulations and granting of a construction permit with the necessary documents and compliance control

Table 4.1 continued

6	build, offic, function, rise, factor, exist, develop, intellig, scale, usag	¬ Fire safety preventions for mixed-use high-rise buildings ¬ Multi-story office buildings and planning open office/Harmancı Giz Plaza, Sabancı Center, Kanyon and Nida Kule samples on methods of planning open office ¬ The standards of users in office design and the evaluation of the usage of technology ¬ Investigation of the relationship between architectural form and structural form in high rise building design ¬ Fire phenomenon and fire protection in high-rise buildings
7	light, visual, interior, design, daylight, illumin, colour, artifici, percept, level	¬ Relationship between facade colours and lighting ¬ Determination of the reflector form dependent on wanted luminaire luminous intensity distribution ¬ Architecture and concept of color ¬ An examination of the shadow quality in lighting ¬ A study on the computer programmes used in lighting design
8	acoust, hall, paramet, perform, music, design, measur, theatr, condit, reflect	¬ The examination of architectural design elemnts in arena type of halls with computer simulation ¬ Analysing the effects of different source positions at the stage part of auditoriums on the room acoustic parameters, simulated in odeon ¬ Effects of architectural design on acoustical performance in halls: Examples of rectangular, fan and diamond shapes ¬ Analysis of the horseshoe planned halls functioned for concert and opera in terms of acoustic design ¬ Acoustical evaluation of four concert halls in İstanbul

Table 4.1 continued

9	particip, communic, individu, factor, level, organ, survey, questionnair, person, satisfact	¬ Workplace happiness of wage worker architects and the determinants of job satisfaction at different career stages ¬ Organizational commitment and job satisfaction in Turkish construction sector ¬ A research on organizational safety culture and mindfulness in construction sector ¬ A study about the leadership behaviours in Turkish construction industry ¬ Mobbing in the construction and architectural services sectors
10	materi, applic, product, element, roof, construct, select, perform, detail, properti	¬ Classification of timber roof trusses according to geometric opportunities and use of material ¬ Investigation of performance properties of composites building materials and their usage possibilities in architecture ¬ The comparative analysis of polymeric sheets used in roofs ¬ The research of the usage of wood and wood-based products in buildings as a cladding material ¬ The Species, properties of glass material and the classification of it's usage for constructional purposes
11	citi, ident, imag, memori, layer, istanbul, landscap, map, trace, element	¬ Night; explorations through dualities and narratives ¬ How the migrant constructs and reproduces space through memory: Collage, montage and translation ¬ Two nomads in the city ¬ The city as a communication interface: Tracing and mapping ¬ Metaphor as a mediator in discovering the urban experience of Karaköy

Table 4.1 continued

12	educ, school, univers, student, children, environ, learn, campus, develop, primari	¬ The investigation of desing of basic education buildings for determent users according to inclusion education ¬ Pre-school education centers: Area study in Lara-Arapsuyu, Dokuma/Çallı/Bayındır districts of Antalya ¬ On the effect of design of montessorian educational models on the embassy school of application review ¬ The name of the thesis interactive education over internet: Graphics and technical drawing course ¬ University campus settlements and spaces for common use
13	facad, wall, system, glass, perform, curtain, window, compon, panel, extern	¬ Comparision of glass curtain walls in terms of components ¬ Modüler cephe tasarımı için öneri ¬ Examination of the relationship between the thermal performance and durability of aluminium curtain walls ¬ Classification of tranparent wall systems and alanysis of their construciton and usage performances ¬ The classification and evaluation of the double-skin curtain wall systems
14	period, modern, centuri, ottoman, turkish, empir, time, style, era, republ	¬ Bab-i Mesihat; the Shaikh al-Islam institution in 19th century ¬ Reading Ankara 19 Mayıs Stadium (Ankara National Stad): Space, social life and ideology relationship in the Early Republican Period ¬ Istanbul muwaqqitkhanas in Ottoman Period ¬ Deterritorialized actors of Pera: Vallauri family, Edouard Lebon, Alexandre Vallauri and M. Vedad Tek ¬ The Evaluation of the residental transformation of the city Ankara in the Early Republican Era through the Novel Ankara

Table 4.1 continued

15	concret, sampl, wood, materi, properti, damag, brick, reinforc, mechan, mortar	¬ Materials properties of contemporary solid bricks and their assessment in reference to the historic bricks ¬ The effects of environmental conditions on the adhesion and durability of cement based mortars ¬ An experimental assessment of textile and wire reinforced horasan mortar strengthening of birick walls in historical buildings ¬ Properties of portland cement ¬ The investigation of the influence of the polymer based latex addition on the properties of the cement based mortars
16	street, district, istanbul, region, citi, histor, develop, town, center, popul	¬ Spatial analysis of grand bazaars: Tabriz and Istanbul ¬ Examining the characteristics of urban space in malls designed with square concept in Turkey ¬ Shopping malls as a center for social interaction Nigeria case ¬ A historic survey of commerce-trade functions, open-air malls and lifestyle centers ¬ The change of the traditional commerce places in Antakya
17	form, shape, pattern, generat, rule, produc, geometri, tradit, geometr, vernacular	¬ Rule for generating Islamic star patterns with shape grammer methods ¬ A study of shape grammar as a generative method, in Anatolian Seljuk geometric patterns (shape grammar as a pattern generation method) ¬ Generative modular system proposal to be used in spatial organization by getting inspiration from organization logic available in nature ¬ Computing the making of Seljuk geometric patterns ¬ A shape grammar model to generate Islamic geometric pattern

Table 4.1 continued

18	project, construct, manag, process, risk, compani, sector, qualiti, cost, contract	¬ A research to determine factors causing cost and time overrun in Turkish construction projects ¬ Payments and payments procedures in construction management as adviser project delivery system ¬ The partnership risks of international joint ventures formed by Turkish contractors and their management strategies ¬ Implementation of dispute review boards in construction projects ¬ Contextualizing and considering the contracts and disagreements between partners in the construction joint venture project process
19	life, peopl, live, human, time, environ, social, communiti, societi, reason	¬ Form - concept relation; space, metaphor, aesthetics ¬ Architectural place design for physically handicapped people ¬ From Architecture to Infrastructure: The Reduction of a Profession ¬ Mobility in architecture: An examination on nomad gypsies and circus life ¬ Discussion of relationship of common interest, common environment and community through practices of living together
20	urban, transform, citi, social, process, project, plan, develop, econom, squar	¬ Social and spatial impects of urban transformation: Fikirtepe transformation zone ¬ User-oriented sustainability in the urban transformation projects; examples of Turkey and abroad ¬ The role of urban regeneration projects to improve the quality of architecture and urban life: The example of Bursa-Central Garage district ¬ Identity problem in urban space the Aksehir example ¬ Transformation projects in urban development: Definition of process and actors, Zeytinburnu case

Table 4.1 continued

21	hous, resid, mass, user, apart, residenti, type, determin, famili, dwell	¬ Urbanization and housing deficit: A comparative analysis of mass housing in lagos and Istanbul ¬ A study on factors affecting the housing prices: Case of Fatih ¬ The analysis of the planning parameters for the multi-storey type housing; the defining of these criteras at İzmir ¬ The development–alteration of mass housing in Edirne from to present ¬ The analysis of property valuation methods: The Evka 3 region
22	space, experi, bodi, movement, interact, virtual, creat, time, spatial, represent	¬ Cinematographic production: Embodiment of space and time experience ¬ Transformativity of body and space: The flesh of space ¬ Experimential montage as a method of representing architecture ¬ Cinematographic notions as a method for space design, representation, experimentation ¬ Effects of space used in multi-player games on enjoyment, immersion, and presence
23	water, wast, air, wind, natur, earth, increas, qualiti, construct, indoor	¬ Ventilation in steep sloped roofs ¬ The negative effects of air conditioning systems on indoor air quality ¬ An implementation proposal for domestic water use and conservation in Kırklareli ¬ Indoor air pollutants and risk management ¬ Applications of water and heat isolation in buildings and investigation of these applications on underground structures
24	mosqu, church, locat, built, structur, centuri, construct, plan, floor, origin	¬ A Byzantine Monastery on Mount Mycale (Dilek/Samsun) in Aydin: Kurşunlu Monastery ¬ Building activities in Istanbul of Sokollu Mehmed Pasha ¬ Rumi mehmet paşa türbesi ve sarnıç-imaret yapısı restorasyon projesi ¬ Beyşehir Eşrefoğlu complex restoration problems ¬ Hipped roof mosques and masjids designed by Mimar Sinan

Table 4.1 continued

25	nois, sound, measur, level, element, calcul, insul, map, valu, effect	¬ Urban soundscape research in the route of Gezi Parkı-Tunel Square ¬ Evaluation of airborn sound insulation in building elements according to EN and ISO standards ¬ Evaluation of D100 highway in terms of noise ¬ Research and review of noise from outdoor entertainment venues ¬ Investigation of barrier effectiveness on noise control in urban scale
26	tourism, hotel, accommod, tourist, develop, khan, istanbul, coastal, facil, histor	¬ Analysing the common used areas in the urban hotels: The region of Talimhane ¬ Cross-border toursim experience: The case of Meric delta ¬ Evaluation of boutique hotel concept as an intersection of architecture and tourism: Izmir city center boutique hotels ¬ Analysis of orientation of bedroom masses in accommodation buildings in Istanbul ¬ Analysis of the architectural plans of hotels constructed in Antalya province between 1990 and 2015
27	art, museum, exhibit, histori, ancient, artist, archaeolog, piec, present, excav	¬ The architectural stone works from Llarisa (Buruncuk) preserved in İzmir Archaeology Museum ¬ The characterization and protection of the Babylonion Ishtar gate glazed bricks ¬ The archaic stone architectural pieces from Larisa (Buruncuk) ¬ Representations of temples on roman period coins from nicomedia ¬ Interpretation and presentation of archaeological sites: The case of magnesia on the Meander
28	energi, effici, heat, climat, consumpt, thermal, perform, design, comfort, build	¬ Heat pump applications in different climatic zones from the energy conservation point of view ¬ Comparison of colling loads of buildings in hot climates (case study on Antalya and Diyarbakır) ¬ An evaluation study on the effect of solar control devices onto cooling energy loads in buildings ¬ Energy-efficient refurbishment building envelope strategies of existing multi-storey residential buildings for degree day regions of Turkey ¬ A study for thermal performance evaluation of building envelope in terms of summer comfort conditions in hot-dry climate

Table 4.1 continued

29	transport, station, park, airport, railway, speed, develop, passeng, termin, pedestrian	¬ A GA based approach to location selection and dimensioning of automated parking facilities ¬ Design considerations for modern railway stations; comparing Berlin, Beijing and Ankara ¬ Study of high speed train station buildings and integration with the city over Bozuyuk, Bilecik and Ankara station buildings ¬ A model for spatial integration of metro station entrances with the city: The case of Istanbul metro stations ¬ Investigation of functional relationships in the design of airport terminal buildings
30	space, spatial, public, social, relat, interact, cultur, physic, organ, function	¬ Traditional Kayseri houses in contex of space syntax and visibility graph analysis ¬ The research of the today's situation of dead-end street function in Anatolian traditional housing tissue ¬ Semantic and syntactic analysis of the relation between user behavior and space in case of interfaces in social interaction spaces of architecture schools in Istanbul ¬ Analysis of urban coffeehouses in the context of public space theories ¬ Perceptional differences depending on social diversity and their effects on use of public space
31	product, practic, critic, discours, context, relat, theori, understand, architectur, transform	¬ Architecture, revolution and temporality: The Soviet avant-garde and the politics of modernism ¬ An ontological inquiry on potentials of criticality and conceptualization of program in architecture ¬ Architecture as an intermediate ground ¬ New Babylon: Discrepancies of utopia and possibility of situationist architectures ¬ (Post-) critical mimarının ötesinde: Realist bir perspektif
32	user, hospit, disast, shelter, health, unit, design, requir, emerg, servic	¬ Emergency department design with user participation ¬ Emergency department design ¬ Creating quality house matrix for emergency shelter tent design ¬ Lighting of emergency service units of hospital; A case study of Hatay-Samandag Public Hospital ¬ Post-occupancy evaluation about quality in military accommodation facilities:case of Sihhiye Orduevi and Gazi Orduevi

Table 4.1 continued

33	industri, product, develop, chang, technolog, global, econom, process, increas, cultur	¬ Branding in architectural offices ¬ Investigation of innovative approaches in high-rise buildings ¬ A model for sustainable urban: Slow city movement ¬ Contributions of architecture to the corporate world in the context of globalization: Headquarters examples ¬ The reflections of popular culture on tourism structures : Accommodation buildings post 1980
34	histor, conserv, cultur, protect, heritag, restor, propos, preserv, function, structur	¬ The restoration proposal of the Karabek Konak in Corum ¬ Restoration Proposal Of The 'Balçiçek House' in Ayaş District, Ankara-Turkey ¬ The restoration proposal of the Kapaklikaya Mansion in Sivrihisar ¬ Restoration project of Kadioğlu 'Mansion' in Niğde ¬ The restoration proposal of residences number 23-23A, located in Sivrihisar Mavikadın Street
35	sustain, environment, ecolog, green, system, resourc, environ, assess, criteria, certif	¬ An evaluation on applications of sustainable approaches in condominiums ¬ A research on the measurement of the consistency of leed green building assesment system criteria in terms of differing design scales, conceptual hierarchy and resource use ¬ Analysis of buildings with leed and breem green building certificate in Turkey ¬ Green office design criteria and comparisons in the scope of BREEAM and LEED assessment certificates ¬ Importance of environmental impact categories in rating systems
36	structur, construct, system, earthquak, steel, timber, develop, bridg, form, load	¬ Worldwide developments of precast frame structures and innovative methods in earthquake resistant design ¬ Prefabricated reinforced structural frames used in the production of industrial buildings and the problems occured in these systems following the 1998 Adana-Ceyhan earthquake ¬ The effect of structural system configuration on the isolated system response ¬ Steel house tecnologies and load bearing system proposals for earthquake regions ¬ Irregularities of Structures According to New Turkish Earthquake Code 1998 and investigations of architectural solutions

Table 4.1 continued

37	concept, approach, form, natur, mean, relationship, context, relat, express, cultur	¬ Form and nature relationship in the context of ecology in architectural design ¬ A research about the metaphor 'form follows flow' in ecological design ¬ Architectural hybridity: Re-definition of “local” through “new” ¬ Mimesis to nature; Art Nouveau architecture ¬ The concept of context in architecture and metaphoric based approaches
38	center, shop, mall, store, konya, bazaar, mix, trade, locat, consum	¬ Shopping malls as a center for social interaction Nigeria case ¬ The examination of sales density of stores in shopping malls by space syntax method ¬ An evaluation of the formation of shopping spaces ¬ The classification of spatial fictions of shopping centers within the context of examples ¬ Transformation of space behavior relation: Dissertation of shopping centers using the method of space syntax

The affiliated words and documents of topics provide a level of knowledge discovery in terms of revealing the latent concepts of the corpus. Beside of the discovery of the topics based on topic-word proportions (β), the prevalence of the topics in the model can be calculate through their document-topic proportions (θ).

As discussed on previous chapters, since documents are distributed over mixture of topics the sum of the θ values of a document is equal to 1. The higher θ value for a document, as a probability between 0 and 1, represents high level of affiliation of towards a topic. On the contrary, the sums of θ values associated with a topic can be higher than 1. Therefore, to normalize the prevalence (proportion) of the topics, the mean of the related θ values for each topic is calculated. With these values the expected topic proportions is presented at Figure 4.4.

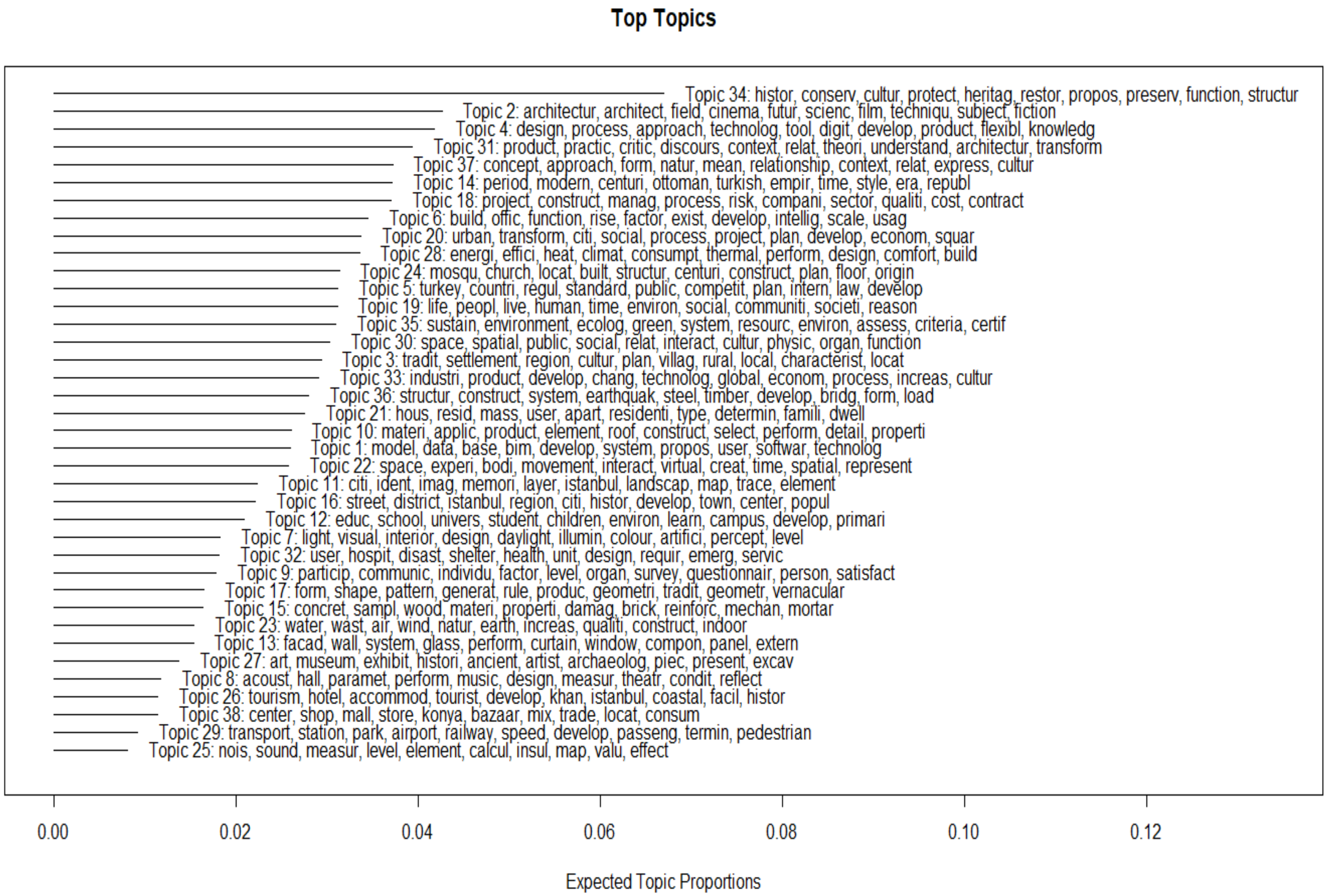


Figure 4.4. Expected Topic Proportions with top 7 keywords

4.2.2. Topic Correlation

One of the beneficial outputs of the STM is the correlation between topics. The correlation between topics are calculated through the co-occurrence of the affiliated documents. Positive correlation between topics indicates that documents are occurring (with high θ) in both topics. While negative correlation indicates the opposite (Figure 4.5).

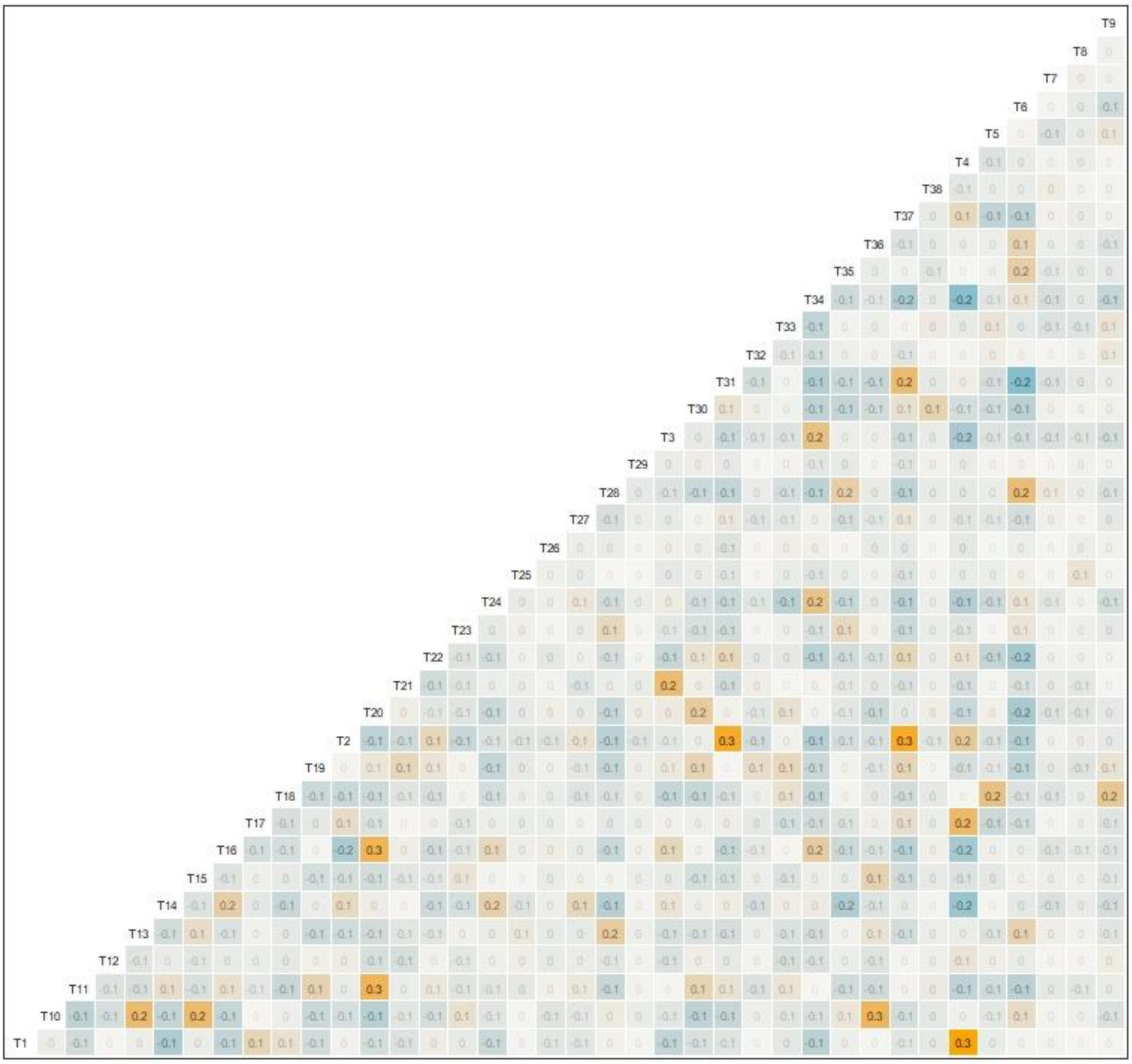


Figure 4.5. Topic Correlation Matrix

4.2.3. Covariate Effects

Regarding the pre-determined covariates, two estimations has been calculated for the effects of covariates over topics: a continuous effect of topics over years (Figure 4.6) and over introduced groups of universities (Figure 4.7). The effects are calculated through the framework presented by the authors of STM [88]:

> "This function performs a regression where topic-proportions are the outcome variable. This allows us to conditional expectation of topic prevalence given document characteristics. Use of the method of composition allows us to incorporate our estimation uncertainty in the dependent variable. Mechanically this means we draw a set of topic proportions from the variational posterior, compute our coefficients, then repeat. To compute quantities of interest we simulate within each batch of coefficients and then average over all our results."

Regarding the effects of the groups, as the dispersion of the corpus indicates, for some universities there is a discontinuity of document submission. Therefore, sudden peaks appear on several topics due to the drastically less number of documents of the related university. The expected proportions over the years for each subject university is present at Appendix B.

[88] Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley, "Stm:R Package for Structural Topic Models."

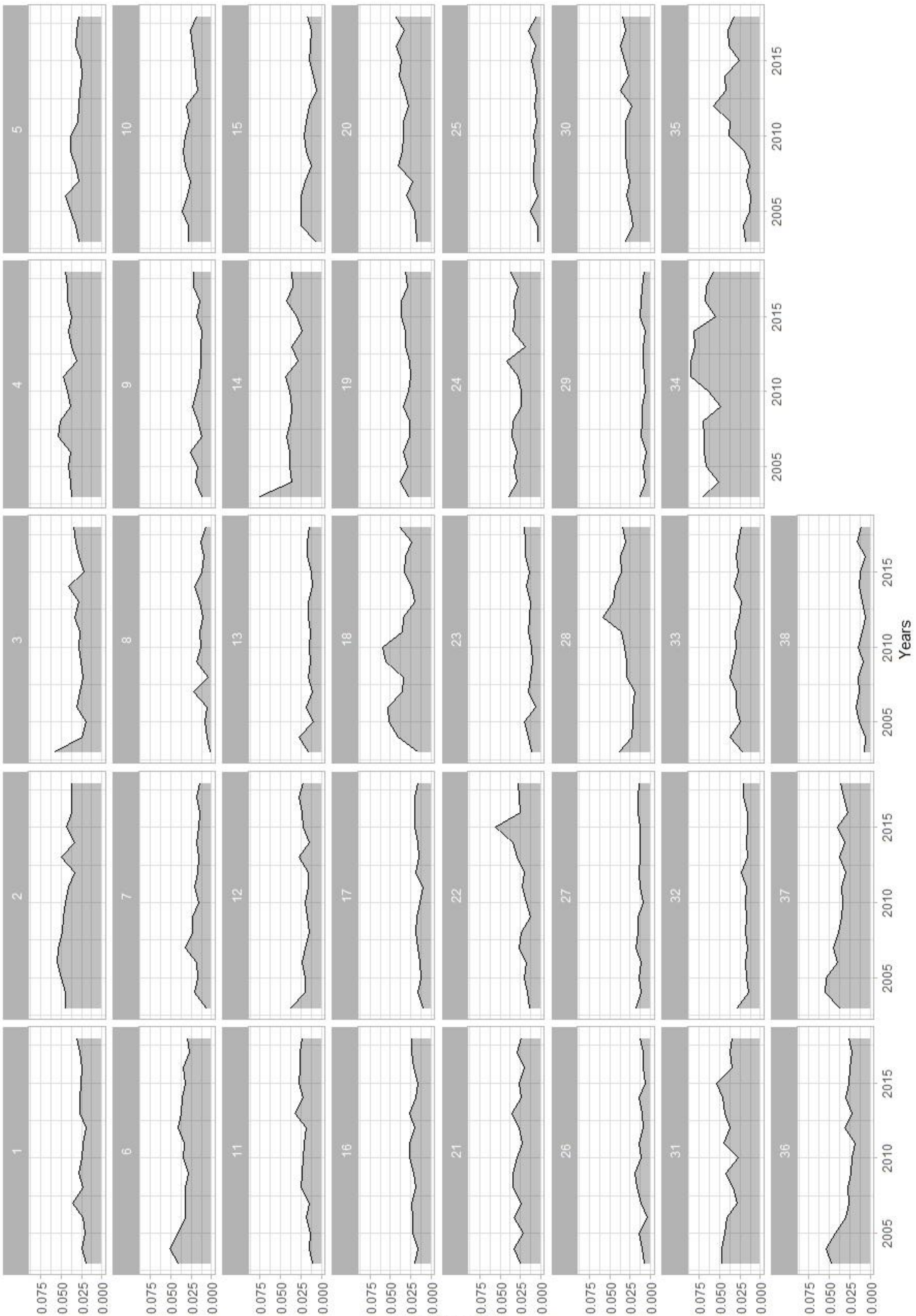


Figure 4.6. Expected topic proportions over years

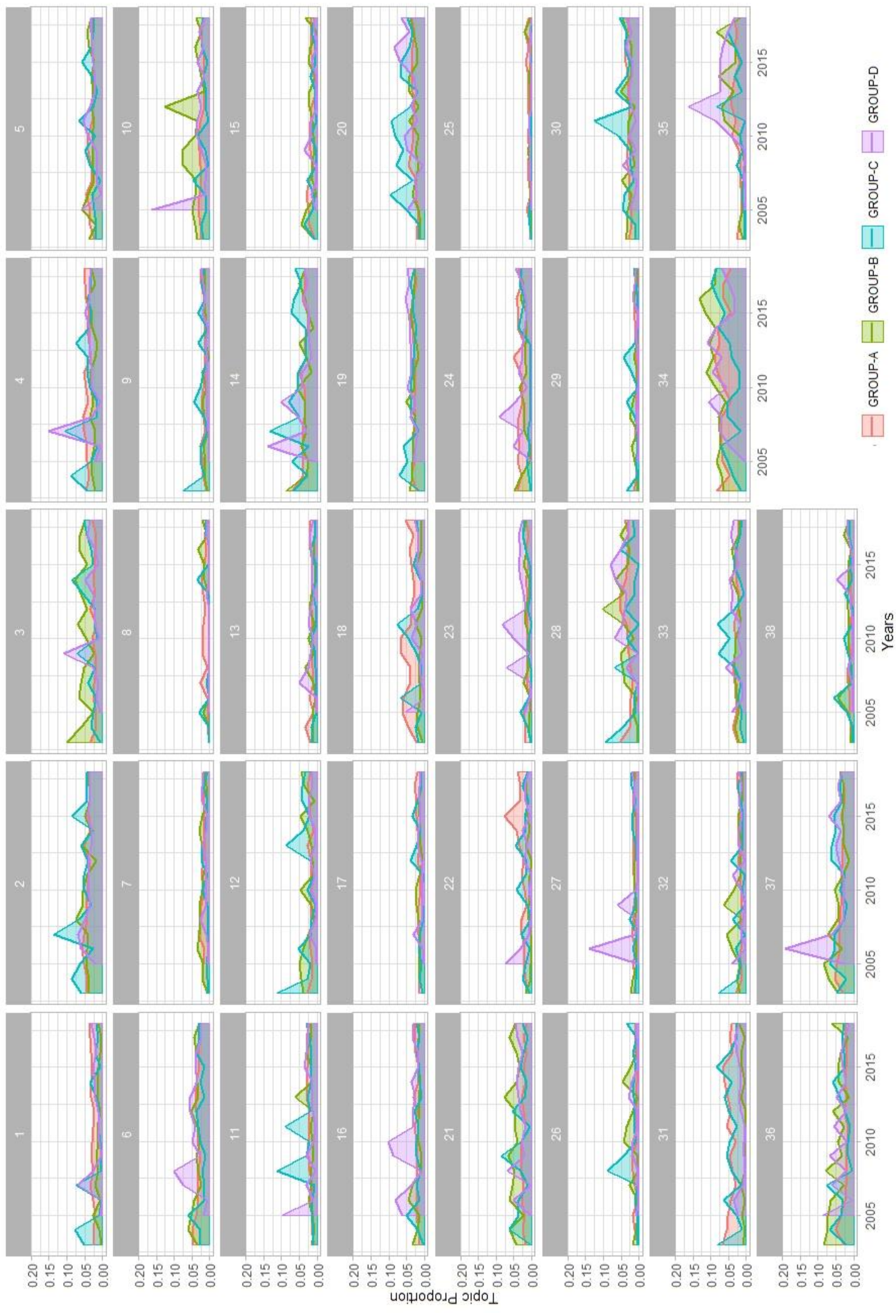


Figure 4.7. Expected topic proportions of Groups over years

4.3. Concluding Remarks on Methodology

Upon this stage, the corpus has been structured, processed and clustered under a structural topic model. Although the overall process has been broadly explained on last two chapters, the decision-making process during the pre-processing and model selection is still a subject of debate.

First, there have been three crucial steps at the pre-processing of the texts that lead to build the model vocabulary:

1. Removal of universal and custom stopwords.
2. Filtering words that appear less than 0.5% of the documents.
3. Stemming of the words.

These steps determine the content quality of the corpus by eliminating insignificant words from the corpus. However, it is worth mentioning that pre-processing methodology have been deemed to have little effect if not no effect at all on the topic model quality [89]. These arguments have been taken to consideration and the pre-processing filters have been decided by observing various model iterations and results under different filtering conditions. Although records of the differing conditions and their effects have been recorded, they been excluded from this study to not overextend the discussions on the methodology.

In brief, not removing the stopwords 3.4.2), especially the custom list, had an negative impact on the model in terms of knowledge discovery. Where topics regarding the methods of the corpus has dominated the results. Additionally, filtering by word frequency has resulted with more coherent topics and less computational power. Therefore, the pre-processing phase of this study is a post-processing treatment of the vocabulary, where the decisions on word filtering has been taken by observing the results instead of predetermination.

[89] Schofield et al., "Understanding Text Pre-Processing for Latent Dirichlet Allocation."

The second crucial step is the decision of the topic number. As the authors of the STM highlight:

> “The analyst must also choose the number of topics. There is no “right” answer to this choice. Varying the number of topics varies the level of granularity of the view into the data. Therefore, the choice will be de- pendent both on the nature of the documents under study and the goals of the analysis.”

Although the selected model with 38 number of topics has relatively high topic quality regarding the semantic coherence and exclusivity metrics (Figure 4.2) , there are models with better scores; such as models with 40 and 50 number of topics. However, these metrics have not been taken under account as an ultimate model selection parameter but rather a guide to narrow the model selection process. Therefore, the model with 38 number topics has been selected after comparing its topics and contained information with various other models.

In conclusion, the methodology of this study can be refutable by others in terms of these two aspects decision-making. Nevertheless, the continuing chapter represents how the results of the selected model overlaps with the architectural research environment in Turkey, therefore it is deemed to be satisfactory under the terms and conditions of this study. It is worth stating that researchers with different research agendas can alter the previous decision-making process to procure a suitable model for their case.

CHAPTER 5

EXPLORATION AND DISCUSSIONS

This chapter covers further visualization and exploration of the results, and discussions regarding the topics, their trends and attributes per sources. In the course of this chapter, the results are discussed by an exploratory basis, without a bias towards the data or any pre-determined expectations from the results.

5.1. Architectural Research Topics

Prior to exploring the results represented on the previous chapter, a more comprehensible visualization of the model outputs (Figure 4.4 and Figure 4.5) is rendered as a network graph (Figure 5.1). This graph contains the quantitative information regarding the topic proportions and relations, where the relation is the positive correlation between topics (higher than 0.1). By eliminating the low correlation edges, this chart demonstrates a legible relationship between the topics. Furthermore, topics are clustered regarding their correlation and contained knowledge to highlight their disciplinary proximity. The addition of these clusters is a manual intervention to the chart in order to draw attention to the major architectural research programs and their relative connections as “Architecture” (design and theory), “Urban”, “Cultural Heritage, Conservation and Restoration”, “Building Science” and “Project Management”. Although this intervention is debatably in terms of representing an epistemological divergence between disciplines, the clustered topics and their highly affiliated documents are submitted under different programs. Therefore, there is a distinct difference in terms of research focus and agenda.

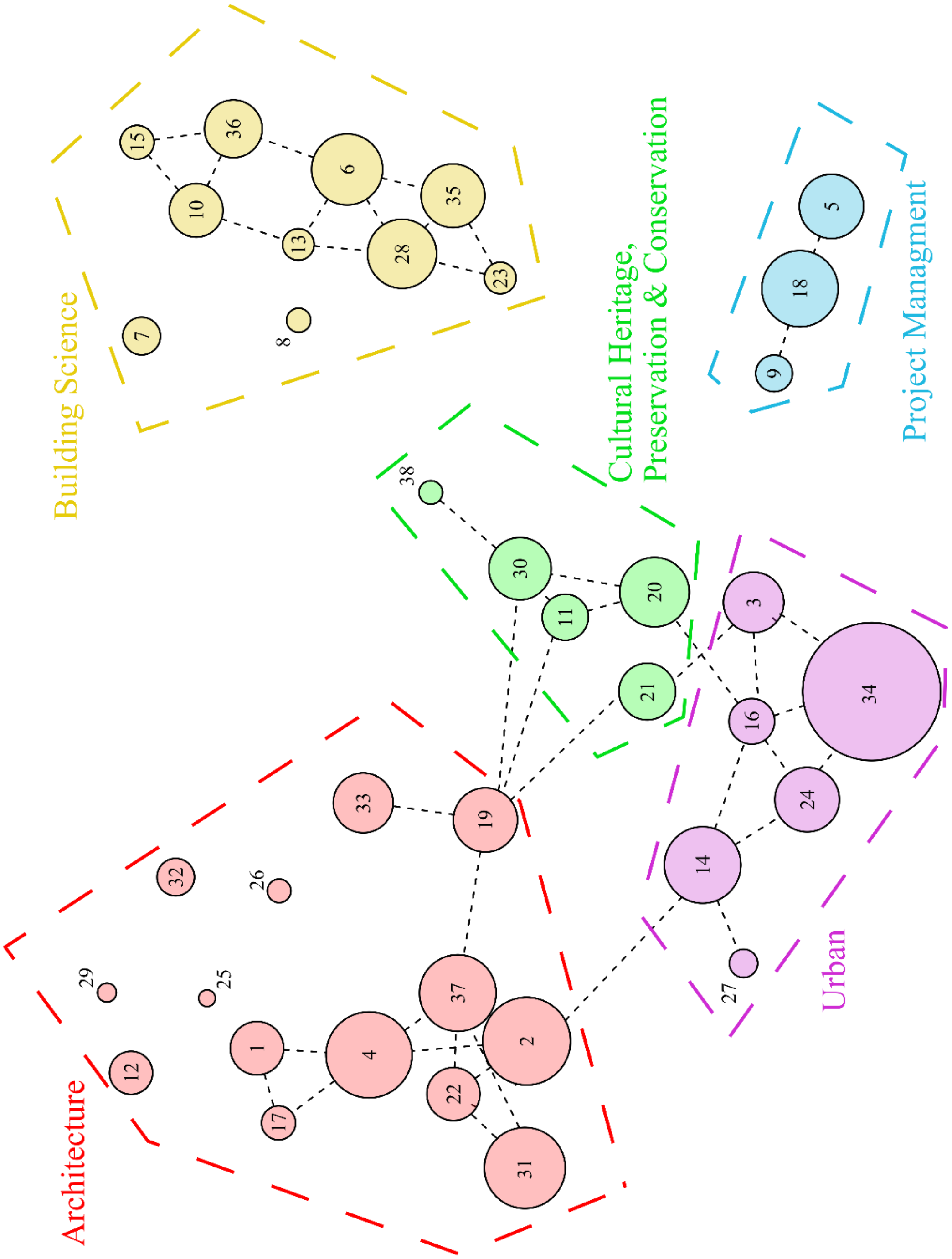


Figure 5.1. Topic Relations.
Node size indicate topic proportions and edge t indicate the positive topic correlation. Dashed polygons are disciplinary clusters distilled by colors.

Initially, the result regarding the topic proportions (Figure 4.4) indicate the proportional dominance is decreasing from the general architectural research focus; such as common subjects of different disciplines cultural heritage and preservation, architectural design and theory, urban theory, project management, building science, towards more specialized subjects such as building and public space inquires, acoustics, building materials and systems. In a general sense, the layout of these results is expected in the overall architectural research environment.

However, one of the relatively unexpected output of the model is the significantly high proportion of the Topic-34; which represents theses with research focus towards architectural preservation and conservation. Although, there is a drastic proportional difference between Topic-34 and the following topic (Topic-2); a first-hand implication that this topic is the most prevailing research subject among Turkeys' architectural research agenda can be misleading. To further investigate this unexpected outcome; first, the quality metrics of the topic is taken for account (Figure 4.3). According to the exclusivity – semantic coherence chart, the topic has an optimum balance between these metrics and therefore it has clustered documents that resemble in terms of knowledge contained and that has less affiliation with other topics. This can be interpreted as that the topic represents documents that are so similar to each other and at the same time differ from most of the other topics and thus it has a concrete structure of information. Additionally, when the highly affiliated documents are checked, the topic contains research regarding the preservation of selected case studies (Table 4.1). Although these theses have different case studies, the descriptions of the preservation and conservation principles (the abstracts of the theses) are mostly alike in terms of word usage and expression. Therefore, the related documents are not proportionally dispersed among other topics. The proportional dominance of this topic can also be observed among different model iterations () discussed on the previous chapter, where the same topic remains at the top rank regardless of models with different number of topics.

Even though the proportional prevalence of the Topic-34 is investigated it can still be accepted as the most prominent research subject. However, due its quantitative nature this topic represents the disciplinary characteristics of the "Cultural Heritage, Conservation and Restoration" research program in terms of main research focus. Hence, its proportion might inaccurately represent the overall dominance of its corresponding discipline. To overcome this possible mispresentation the cumulative proportions of each discipline is represented at Table 5.1.

Table 5.1. Topics clustered under Disciplines and their cumulative proportions

Discipline	**Topics**	**Proportion (%)**
Architecture	1, 2, 4, 12, 17, 19, 22, 26, 29, 31, 32, 33, 37	34.9
Building Science	6, 7, 8, 10, 13, 15, 23, 25, 28, 35, 36	23.8
Cultural Heritage, Conservation and Restoration	3, 14, 16, 24, 27, 34	20.1
Urban	11, 20, 21, 30, 38	12.5
Project Management	5, 9, 18	8.7

The proportional distribution of the disciplines grants an overall comprehension of the leading research programs and further clarifies the effect of Topic-34 and its parent discipline. Eventually, it is revealed that the leading research program is architectural theory and design, which is represented by almost one third of the topics.

Topics clustered under architectural theory and design are compose the fourth of the top five topics. Although these topics are commonly and broadly discussed subjects within the architectural domain; an out of common result is that studies regarding the representation of architecture on fiction and art (Topic-2), such as cinema, literature, and so on, is prevailing among these topics. This topic is yet to be followed by most expected research agendas on architectural design process and tools (Topic-4), then by architectural theory and criticism (Topic-31) and formerly by conceptual design and form agenda (Topic-37). Besides these dominating topics of the discipline, a rather interesting subject is also generated by the STM: Topic-19. First of all, this topic shows the same quality characteristics of priorly discussed Topic-34; it has an optimum value between the exclusivity and semantic-coherence metrics (Figure 4.3). While it has the most desired topic quality, the high probability keywords and documents related to this topic makes it comparatively difficult to comprehend the latent knowledge it represents. In a general sense, the topic reveals a subject regarding to human, environment and space relationship in a broader scope. While the high probability of the affiliated documents of this topics, indicates that the language usage and word occurrence on the abstracts of these documents displays resemblance despite the theses titles slightly diverge in terms of knowledge expression. This matter was discussed as a limitation of the prior chapters [91]. Furthermore, the topic is an intersection between urban and architectural disciplinary discussions due to this homogenous nature.

Albeit studies related to various cities are conducted; Istanbul, the most populated city of Turkey, has emerged as a keyword of Topic-11 under the discussion of urban identity and image with relation with almost all of the urban related research themes. While categorized under a different discipline; Istanbul again appears as a high probability keyword of Topic-16, since the city is both highly populated and has an ancient history. A similar case of city keyword appears on the relatively low proportion Topic-38. Konya, a mid-range populated city but with two contributing universities (SELCUK and KTO), has a high probability on this subject of shopping and trade centers. Compared to the case of Istanbul, this distinction of Konya under a topic is caused by the frequent usage of the word by the two local universities.

[91] see Section 3.1. Assumptions

Building science is highly represented by the studies on high-rise buildings (Topic-6), especially on office buildings and their systems; followed by other attractive subjects of energy efficiency (Topic-28) and sustainability (Topic-35). With the remaining related topics; the general layout of the subjects of this discipline has quite less unexpected outcomes, while the topics are strongly representing the research focus and agenda of the discipline.

5.2. Research Trends and Emergence

The yearly effects of the topics introduced at Figure 4.6 is re-visualized to highlight their corresponding disciplines (Figure 5.2), since the disciplinary clusters also represent the correlation and proximity between the topics. The initial observation of the proportional distributions of topics overlap their overall excepted proportions presented at Figure 4.4; however, some topics display increasing and decreasing trends over the timeline.

Trends regarding the architectural design and theory, the highest emergence rate is with the topic associated with experience and perception of real and virtual spaces (Topic-22). The drastic proportional increase of this topic is a standalone situation where this trend even differs from the trends of the highly correlated topics (Topic-2, Topic-31 and Topic-37). While there are no valid signs to explain this phenomena, the sudden increase and then the decrease around between the years 2012 and 2016 could be interpreted as temporary research focus.

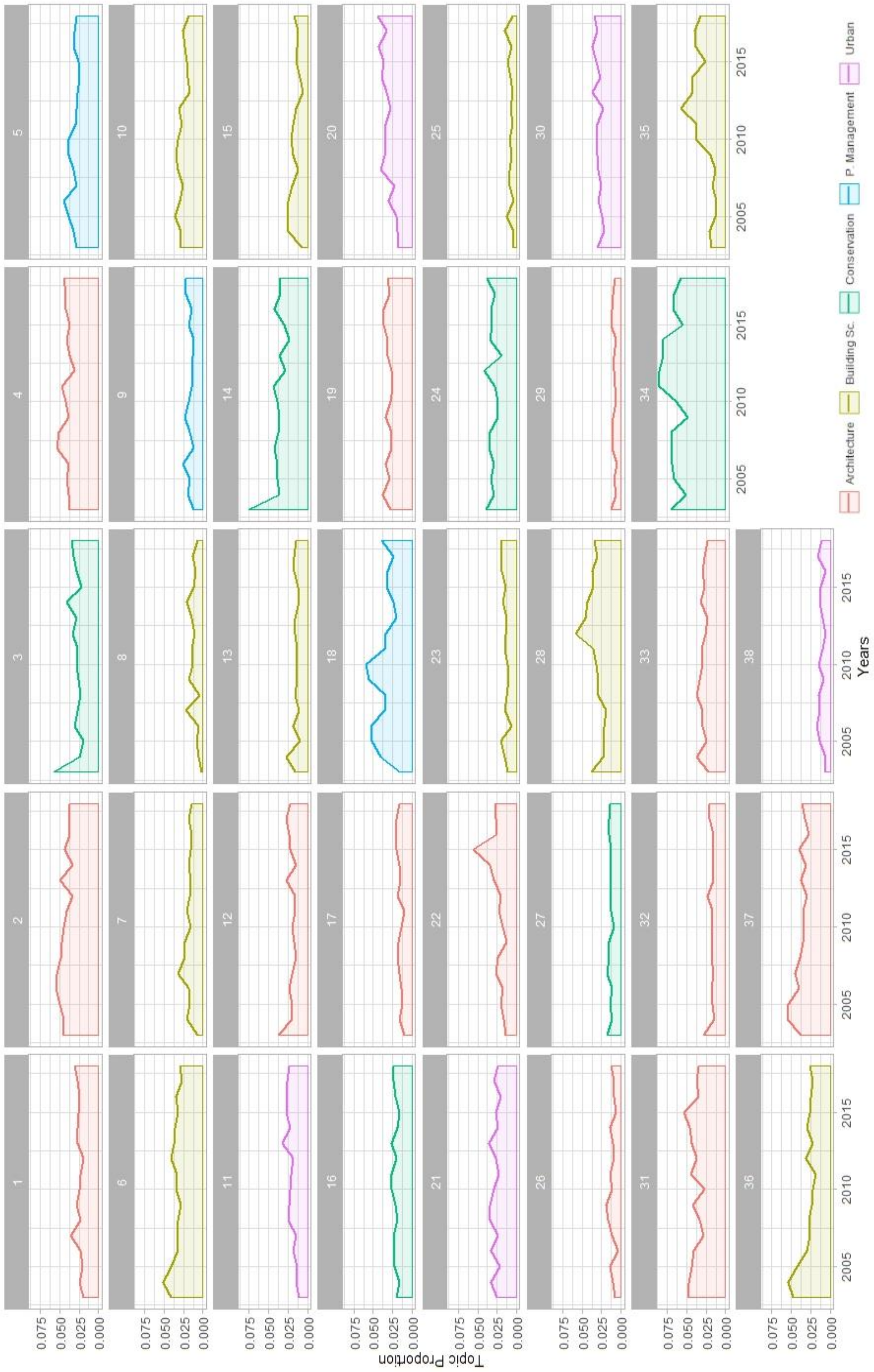


Figure 5.2. Topic proportions over Years. Colored by Disciplines.

Energy efficiency (Topic-28) and sustainability (Topic-35), both highly correlated topics, display a simultaneous increase until 2011 and afterwards a slight decrease. For the case of Topic-35, it is highly affiliated with theses focused on building energy certification and sustainable design. While these agendas have an increased interest worldwide, energy certification became a subject of interest in Turkey on the same time frame with the escalation of the construction industry and high-end projects with energy efficiency considerations. Hence a critical reading of these results indicates that there is a certain level of affection towards the industry from the architectural institutions.

A comparable discussion can also be carried out for Topic-20 as well; where the topic gathers documents regarding to urban transformation and renewal. Although this topic has a mild increase over time; within the context of Turkeys' urban development policies, the trend of the topic indicates a level of research interest to the contextual political effects and results.

As one of the most industry related discipline, project management and its associated topics represent a different tendency. While Topic-5, Topic-9 and Topic-18 show a correlated decline of interest after 2010; especially Topic-18 is gradually decreasing ever since. This topic, representing the agenda of project and construction management, is the prominent agenda of its related discipline. This decrease could be interpreted under two circumstances: firstly, considering the nature of this discipline, it intersects with administration, law and management rather than the architectural domain. Therefore, architectural faculties might leave this field to other faculties or research programs. Secondly, since this discipline is heavily affiliated with the construction industry, the focus of the institutes might be shifted from a research program to a graduate program without thesis; where the main aim is to train qualified workforce to the industry.

5.3. Institutional Factors in Research Tendency

Universities, under four groups, were introduced as covariates to the STM and the effects of these groups over the years is represented at Figure 4.7. As discussed on the prior chapter, the immediate peaks are caused by the sparsity of documents on specific years. This sparsity is mostly observed on GROUP-D, which contains the most amount of universities but the least number of theses. Although the documents submitted by this group starts form 2005, until 2010 there is considerably low number of documents. Therefore; prior to any discussion, these sudden leaps on trends can be considered as outliers and deemed to be insignificant regarding to the knowledge discovery. This consideration is also applicable for GROUP-C, since it also has a discontinuous array of documents until 2010.

With this consideration, the significant effect of GROUP-D can be observed to topics regarding sustainability (Topic -35), the urban culture discussions related to Istanbul (Topic-16), waste management (Topic-23), and high-rise buildings (Topic-6). According to their effects on these topics, the research focus shifts from one to another without any reliable correlation or clarifying data. A noteworthy assessment of these effects can be the nature of the universities that form this group: most of these universities are located at Istanbul and they are all private funded. Since Istanbul is the most densely populated and the economically leading city of Turkey, being located there boosts the research tendency towards the agenda of the city. Apart from that, this group does not have any determined agenda, but leans towards mainstream discussions.

The GROUP-C, which is constituted by public funded universities and relatively newly established architectural research programs, displays dominance mostly over urban studies. The groups effects on Topic-20, urban transformation, and Topic-30, public spaces, can be interpreted to the lack of city and regional planning research degrees and therefore the architectural research program initiated these studies. Additionally, this group align with GROUP-A, the numerically vast and oldest research conducting group, under the Topic-31 which is core theme of architectural theory and criticism. This proximity is somewhat unexpected considering that the GROUP-C is formed by universities located at less populated cities than the GROUP-A universities but nevertheless share a common interest towards this epistemologically rich topic. Another unanticipated result with respect to GROUP-C, is their proportional effect on topics with high relation on urban identity and cultural discussions related to Istanbul (Topic-11 and Topic-16), while none of the constitutive universities of this group is located there. This situation can be interpreted as an attempt to align with the mainstream case studies, such as Istanbul.

Regarding the effects of the GROUP-B universities, there is a distinct tendency towards architectural preservation and conservation that can be observed on their proportions over Topic-3 and Topic-34. This group can be attributed as the most dedicated universities towards this discipline as per their proportional dominance. Additionally, the discipline of building science is also greatly contributed by this group according to their prevalence on Topic-10 and Topic-36; where the former contains the studies on construction materials and applications, and the latter is associated with research based on construction systems. Although the group demonstrate interest to other disciplines as well, their research characteristics can be identified to these two disciplines.

As the major document contributor and with a long history of architectural research programs, GROUP-A has a feeble effect on the overall topic layout regardless of its scale. The group is proportionally leading a handful of topics, and only prevailing with a mediocre difference. The relatively high effects are present at Topic-18, project management, and Topic-22, experience of space, which are the most prominent topics of the group and completely irrelevant in terms of topic correlation and disciplinary proximity. Furthermore, a somehow distinct proportional difference appears on subject of computer aided design (Topic-1) and acoustical design (Topic-8). At this stage it is rather difficult to define a characteristic of the research agenda of this group; therefore, to further investigate this low prevalence over the topics, the group is

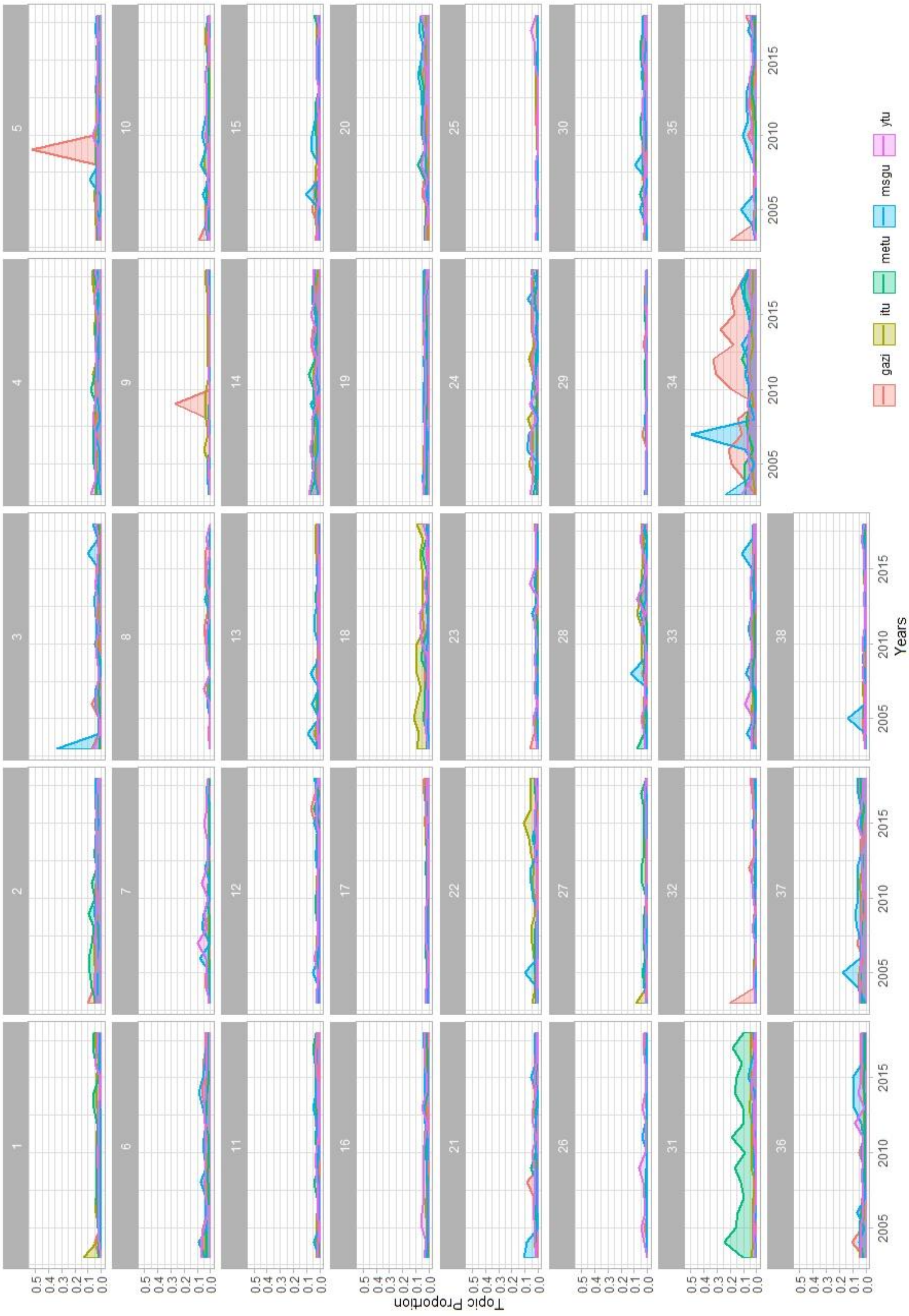


Figure 5.3. Topic Proportions over Years. Colored by GROUP-A universities: GAZI, ITU, METU, MSGU, YTU.

disintegrated into its constituting universities. The effect of the five corresponding universities of GROUP-A is re-calculated based on the previously introduced workflow and the results are introduced at Figure 5.3.

The results of these GROUP-A universities effects over topics yet again indicate the presence of outliers: which is evident on the GAZI and MSGU universities whom has a discontinuous array of documents between 2005 and 2008 (Figure 3.1). This sparsity is caused by missing theses from the database. In the case for GAZI it is much visible at Topic-34, where there is a sudden decrease at 2009 and simultaneously a sudden proportional peak on Topic-5 and Topic-9. A comparable situation is also observed for MSGU: the reduced number of documents at 2006 causes an unexpected proportional increase on Topic-34 while other topics show decreased proportion. Nonetheless, these abnormalities are neglected from of the discussions and the results are further investigated.

Initially, two major circumstances are revealed. First, the dominant effect of METU over the research focus towards architectural theory and criticism (Topic-31). As the results indicate the effect of METU over this topic is higher than the rest combined and without its contribution on this subject, this topic may not be emerged with such high proportion. Second, the effect of GAZI on studies of preservation and conservation, although decreasing since 2015 still considerably high. In a sense, with reference to these results, the research characteristics of these two universities (METU and GAZI) are revealed.

The effect of MSGU affirms an out of common comprehension of this university's' research focus. As a renowned fine arts university, it has a prevailing effect on topics related to building science. This can be observed on the proportional extent over the studies on sustainability (Topic-36), construction materials (Topic-15) building systems (Topic-36) and façade design (Topic-13). While its effects on these subjects are not excessive, compared to the technical universities its dominance is rather unexpected.

The remaining two universities ITU and YTU, combined overwhelmingly comprise %46.8 of the corpus, display a very weak characteristic of research focus. While ITU has a dominant agenda towards the non-correlated studies on project management (Topic-18) and space and experience (Topic-22), which its sudden effects reveal the discussions on the previous section, YTU has no remarkable distinction over the topic proportions. Compared to their corpus scale, this feeble effect is rather caused by their numerous and extended research programs under the architectural faculty. Hence, the divergence within these universities, their research focus and aspect cannot be strictly identified by this model. Their proportional distributions over the topics overlaps with the overall layout, since they are conducting studies under all topics simultaneously.

CHAPTER 6

CONCLUSION

This thesis tackles the problem of what are the architectural research agendas in the context of Turkey and how can this knowledge be extracted regarding the exponential accumulation of textual data. The problem of extracting the research agendas, with the limitations of data selection, has been achieved through means of statistical modeling and computation. This model further allowed to comprehend the trends and emergence of these agendas.

The exploratory phase of this study presented the topics as agendas by interpreting the model outputs that are explored through visualizations. By complying with the framework of exploratory data analysis, this exploration was done without any bias towards the data and was an essential phase of the methodology. The data processing from the corpus to the exploration is the expansion of the reproducible and scalable method and composed as a general framework for future studies.

A model-based agenda tracking as such this study has introduced, excludes from conventional architectural reading due to its statistical background. Furthermore, a highlight of this study is the focus on several factors that are presumed to be effecting the architectural research agenda and the investigation of the factors is concluded with the mapping of this paradigm over the course of 15 years. This map reveals the research focus and the effects of the subject universities of the given timeline. Where the overall agenda shifts can be observed and compared to multiple layers of factors, the extracted knowledge has been briefly mapped with the general context of Turkey. Furthermore, although the model results present rather expected research themes, it became evident that this knowledge is no longer explicit and the results can be accepted as a basis for a possible hypothesis. With this hypothesis, the presented results can be tested through statistical means.

Regarding the mapping of the knowledge, the map contains the layers of knowledge produced per time and per source. This map provides a basis for further critical reading of the architectural research environment of Turkey. Although this study is limited to the corpus presented and concluded within these limits, with different critical attitude and interpretation, novel layers can be added to this map within the same corpus. Furthermore, findings of other related works can be super positioned to this map to expand the overall knowledge discovery.

Throughout this study, the adopted methods to trace of architectural research agenda has established itself as a feasible tool and furthermore a novel medium of architectural narration. At the age of big data, the adaptation of methods alike has become a necessity rather than a supplementary tool. Eventually, this necessity has urged similar studies and research. Although the mathematical, statistical and computational background of the method exceeds the field of architecture in terms of knowledge, the core concept of the method is yet comprehensible and adaptable to the architectural domain. While the computational process of model generation has been acknowledged as a partial black box, without in-depth knowledge of its internal workings, the results and further interpretations validate the model as a viable tool. Hence, the statistical outputs and the critical reading with an architectural basis has confirmed each other.

By the systematical approach of deduction throughout the data organizing, structuring and model construction, an output of manageable scale has been generated. However, different decisions and approaches along these deductions can eventually lead to different outputs. Nevertheless, the subjective interpretation of the results has revealed the implicit knowledge over the architectural research agenda.

6.1. Future Works

Although textual data analysis techniques and topic modeling in specific have a disciplinary establishment of nearly 20 years, their application scope is expanding drastically in many fields. For discursive fields such as social sciences, philosophy, architecture, and so on these techniques have immense potential of the application. While conventional data analysis is conducted with numerical data sets, the tokenization and hence enumeration of the language enables novel types of data analysis. With these novel data sets and methods, retrospective reading of history and agenda tracking research can be conducted under various disciplines. In this regard, by eliminating the limitations of this study in terms of data source and scope, future researchers can expand and contribute with further context to the architectural agenda tracking.

A key feature of the method adopted in this study, with other similar methods, is that texts generated in different languages can be processed to extract knowledge without prerequisite expertise in those languages. In this matter, for researchers who prefer and excels at their native languages, especially with a language with a non-Latin alphabet, can now be effectively accessible globally and furthermore they can access to knowledge written in a foreign language. Therefore; textual data analysis methods and techniques elevate as an emancipatory medium, by providing opportunities for researchers in various fields to conduct interdisciplinary studies with the ease of knowledge discovery. Furthermore, this medium sustains transparency between the epistemology and linguistic barriers of different fields.

By eliminating the barrier of language and information sharing, future works can alter the architectural discourse in multi-directions. With the practical usage of the text analytics, a broader reading of various cultures, languages and social dynamics becomes effectively possible. This possibility is not limited to architectural discourse alone but can further incorporate architectural praxis.

REFERENCES

Acar, Yiğit. "Atlas of Urban Design: Textual Analysis and Mapping of Production of Knowledge in Turkish Context." Middle East Technical Universtiy, 2017.

Aggarwal, Charu C., and ChengXiang Zhai. *Mining Text Data*. Edited by Charu C. Aggarwal and Chengxiang Zhai. *Mining Text Data*. Springer-Verlag New York, 2012. https://doi.org/10.1007/978-1-4614-3223-4.

Anscombe, F. J., and John W. Tukey. "The Examination and Analysis of Residuals." *Technometrics* 5, no. 2 (1963): 141–60. https://doi.org/10.1002/qaj.216.

Blei, David. "Probabilistic Topic Models." *Proceedings of the 17th ACM SIGKDD International Conference Tutorials on - KDD '11 Tutorials*, 2011, 1–1. https://doi.org/10.1145/2107736.2107741.

Blei, David M, Blei@cs Berkeley Edu, Andrew Y Ng, Ang@cs Stanford Edu, Michael I Jordan, and Jordan@cs Berkeley Edu. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (2003): 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993.

Chai, Kah Hin, and Xin Xiao. "Understanding Design Research: A Bibliometric Analysis of Design Studies (1996-2010)." *Design Studies* 33, no. 1 (2012): 24–43. https://doi.org/10.1016/j.destud.2011.06.004.

Chatfield, Chris. "Exploratory Data Analysis." *European Journal of Operational Research* 23 (1986): 5–13. https://doi.org/10.1017/cbo9780511626166.003.

D'Avanzo, E., A. Elia, T. Kuflik, A. Lieto, and R. Preziosi. "Where Does Text Mining Meet Knowledge Management? A Case Study." *Interdisciplinary Aspects of Information Systems Studies: The Italian Association for Information Systems*, 2008, 311–17. https://doi.org/10.1007/978-3-7908-2010-2_38.

Deerwester, Scott, George W Furnas, Thomas K Landauer, and Richard Harshman. "Indexing by Latent Semantic Analysis." *Journal of American Society for Information Science* 41, no. 6 (1990): 391–407. https://doi.org/10.1017/CBO9781107415324.004.

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From Data Mining to Knowledge Discovery in Databases." *AI Magazine* 17, no. 3 (1996): 37–54. https://doi.org/10.1007/978-3-319-18032-8_50.

Frawley, William J., Gregory Piatetsky-Shapiro, and Christopher J. Matheus. "Knowledge Discovery in Databases: An Overview." *AI Magazine* 13, no. 3 (1992): 57–70. https://doi.org/https://doi.org/10.1609/aimag.v13i3.1011.

Friendly, Micheal. "Milestones in the History of Thematic Cartography, Statistical

Graphics, and Data Visualization," 2009, 1–79. http://www.math.yorku.ca/SCS/Gallery/milestone/mileston%0Ae.pdf.

Hays, K. Michael. *Architecture Theory since 1968*. Cambridge, Massachusetts: The MIT Press, 1998.

Hofmann, Thomas. "Probabilistic Latent Semantic Indexing." *ACM SIGIR Forum* 51, no. 2 (2017): 211–18. https://doi.org/10.1145/3130348.3130370.

Jencks, Charles. *The Architecture of the Jumping Universe. A Polemic: How Complexity Science Is Changing Architecture and Culture*. Baffins Lane, Chichester: John Wiley & Sons, 1995. file:///Volumes/750G/# setting/# Papers2 Library/Papers2/Files/Architecture of the Jumping Universe, The - Wei Zhi.pdf%5Cnpapers2://publication/uuid/1A5CD109-9998-4BB0-87C0-3DD5AF51CF3D.

Lafferty, John D, and David M Blei. "Correlated Topic Models." *Advances in Neural Information Processing Systems 18,* 2006, 147–54. http://papers.nips.cc/paper/2906-correlated-topic-models.pdf.

Liu, Z.Y., W.P. Wang, Y. Wang, W.Y. Lu, and Z.Z. Ji. "Bayesian Parameter Estimation in LDA." In *Proceedings of the International Conference on Computer Information Systems and Industrial Applications*, 837–40, 2015. https://doi.org/https://doi.org/10.2991/cisia-15.2015.225.

Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. "Stm:R Package for Structural Topic Models." *Journal of Statistical Software,* no. 2014 (2017). https://doi.org/10.18637/jss.v000.i00.

Miller, Clayton, Matias Quintana, and Jason Glazer. "Twenty Years of Building Simulation Trends : Text Mining and Topic Modeling of the Bldg-Sim Email List Archive Topic Modeling of the Bldg-Sim Email List Archive," no. July (2019). https://doi.org/10.13140/RG.2.2.24955.46885.

Nesbitt, Kate, ed. *Theorizing a New Agenda for Architecture an Anthology of Architectural Theory 1965-1995*. New York, New York: Princeton Architectural Press, 1996.

Porter, M. F. "An Algorithm for Suffix Stripping." *Program* 14, no. 3 (1980): 130–37. https://doi.org/10.1108/eb046814.

Rajman, Martin, and Romaric Besançon. "Text Mining - Knowledge Extraction from Unstructured Textual Data." In *Advances in Data Science and Classification. Studies in Classification, Data Analysis, and Knowledge Organization*, edited by A. Rizzi, M. Vichi, and BockHH., 473–80. Berlin, Heidelberg: Springer, 1998. https://doi.org/https://doi.org/10.1007/978-3-642-72253-0_64.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, and Edoardo M. Airoldi. "The Structural Topic Model and Applied Social Science." *NIPS 2013 Workshop* *on Topic Models*, 2013, 2–5. http://mimno.infosci.cornell.edu/nips2013ws/slides/stm.pdf%5Cnhttp://structuraltopicmodel.com/.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. “Structural Topic Models for Open-Ended Survey Responses.” *American Journal of Political Science* 58, no. 4 (2014): 1064–82. https://doi.org/10.1111/ajps.12103.

Schofield, Alexandra, Mans Magnusson, Laure Thompson, and David Mimno. “Understanding Text Pre-Processing for Latent Dirichlet Allocation,” 2017.

Schumacher, Patrik. *The Autopoiesis of Architecture: A New Framework for Architecture, Volume I*. Vol. I, 2011. papers3://publication/uuid/6DD45F85-0900-4BC2-8508-A774612C0A4F.

Thomas P. Minka. “Estimating a Dirichlet Distribution,” 2000.

Tufte, Edward R. *The Visual Display of Quantitative Information*. Cheshire, Conn.: Graphics Press, 2001.

Tukey, John W. *Exploratory Data Analysis*. Addison-Wesley Pub, 1977.

———. “The Future of Data Analysis.” *The Annals of Mathematical Statistics* 33, no. 1 (1962): 1–67.

Tvinnereim, Endre, and Kjersti Fløttum. “Explaining Topic Prevalence in Answers to Open-Ended Survey Questions about Climate Change.” *Nature Climate Change* 5, no. 8 (2015). https://doi.org/10.1038/nclimate2663.

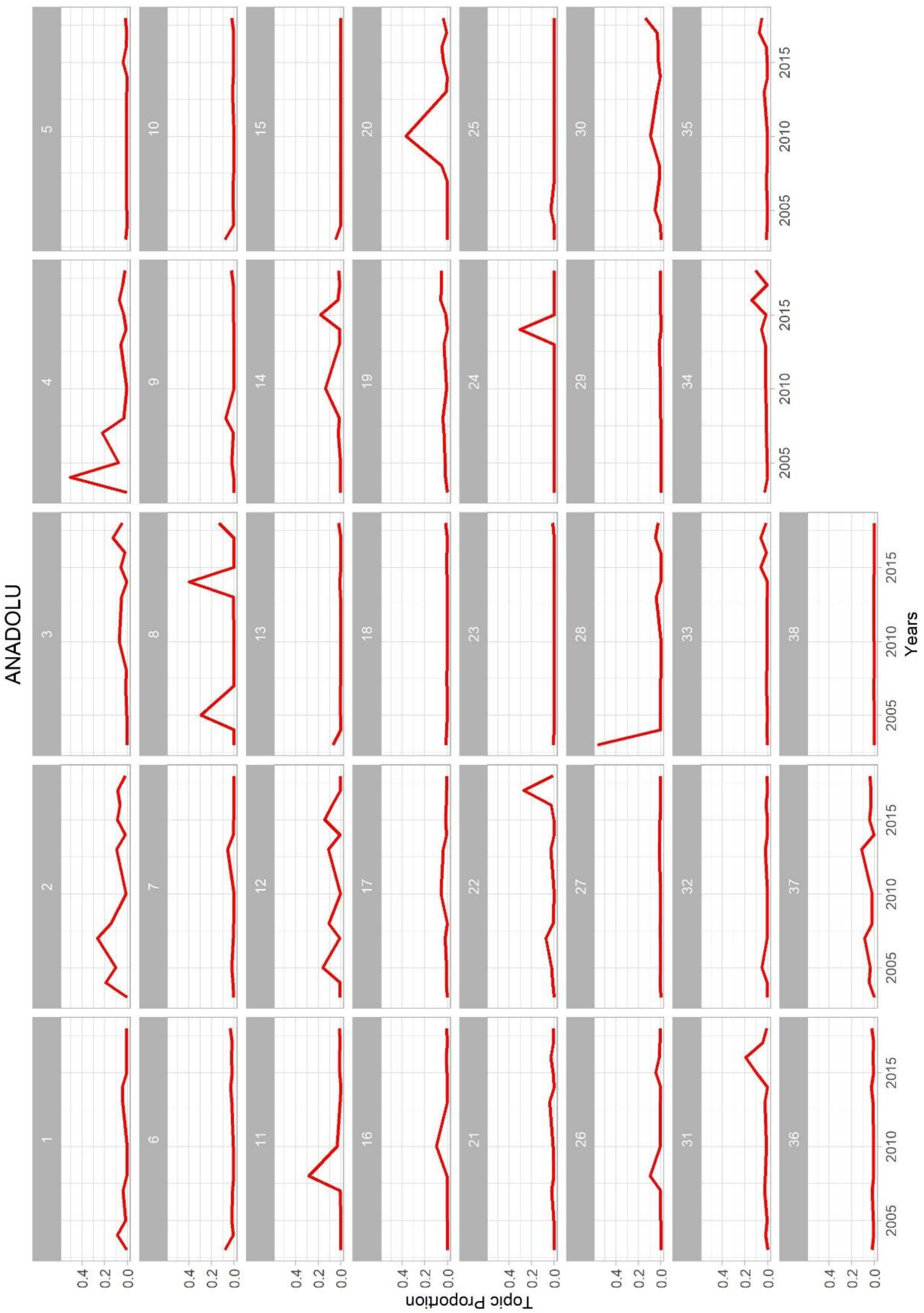
Wickham, Hadley. “Tidy Data.” *Journal of Statistical Software* 59, no. 10 (2015). https://doi.org/10.18637/jss.v059.i10.

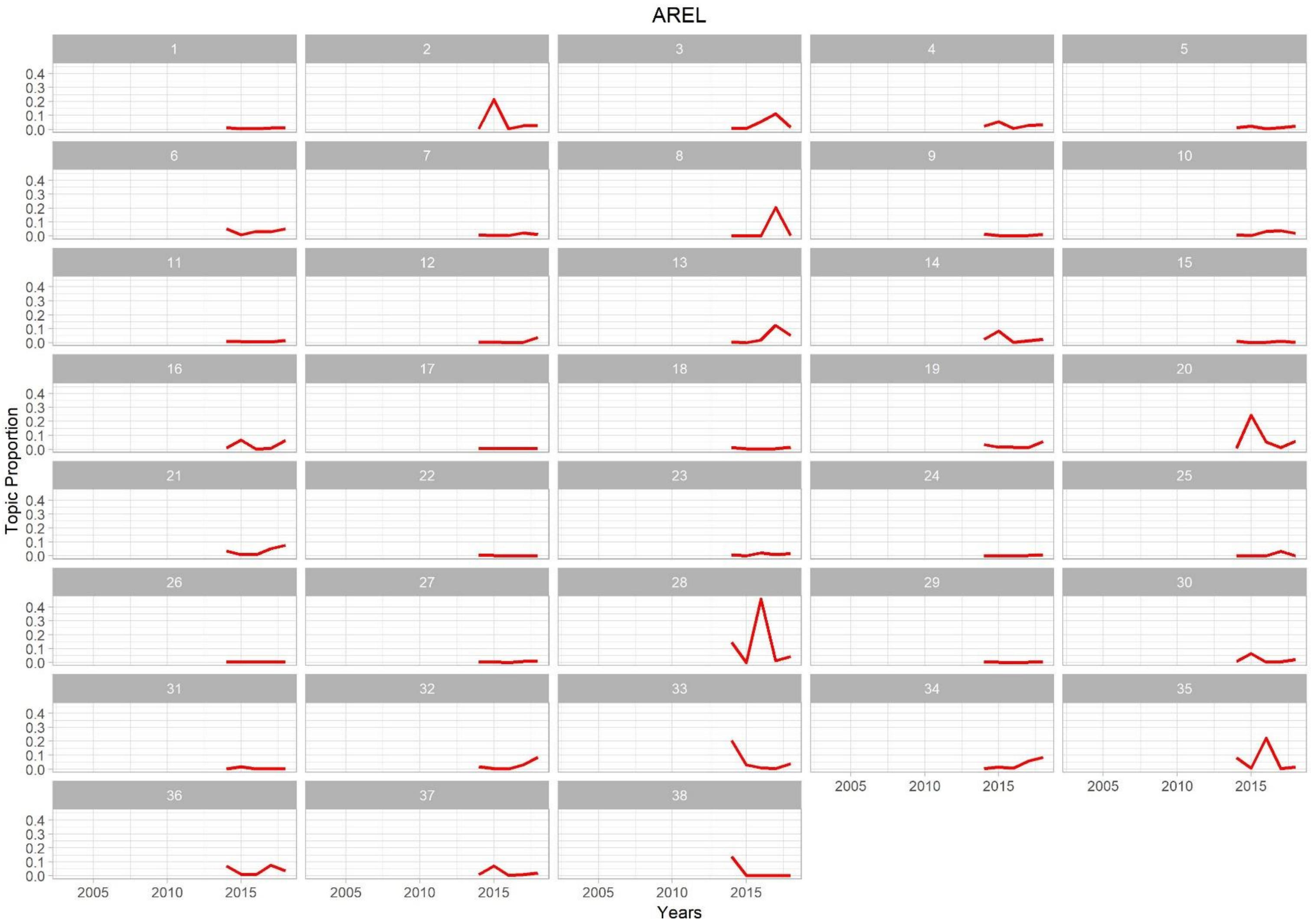
Wu, Hao, Jiajun Bu, Chun Chen, Jianke Zhu, Lijun Zhang, Haifeng Liu, Can Wang, and Deng Cai. “Locally Discriminative Topic Modeling.” *Pattern Recognition* 45, no. 1 (2012): 617–25. https://doi.org/10.1016/j.patcog.2011.04.029.

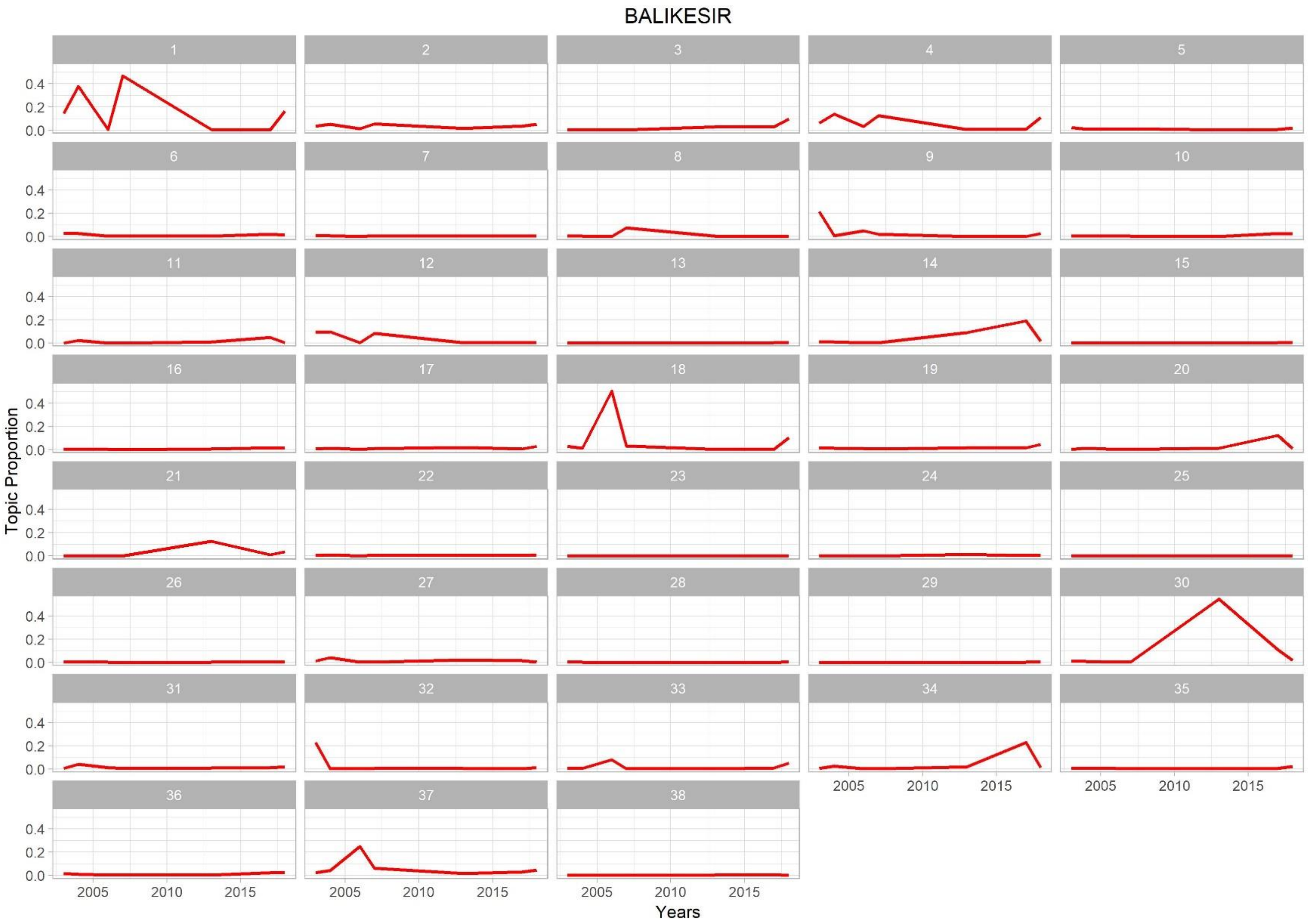
A. Model Vocabulary (sorted alphabetically)

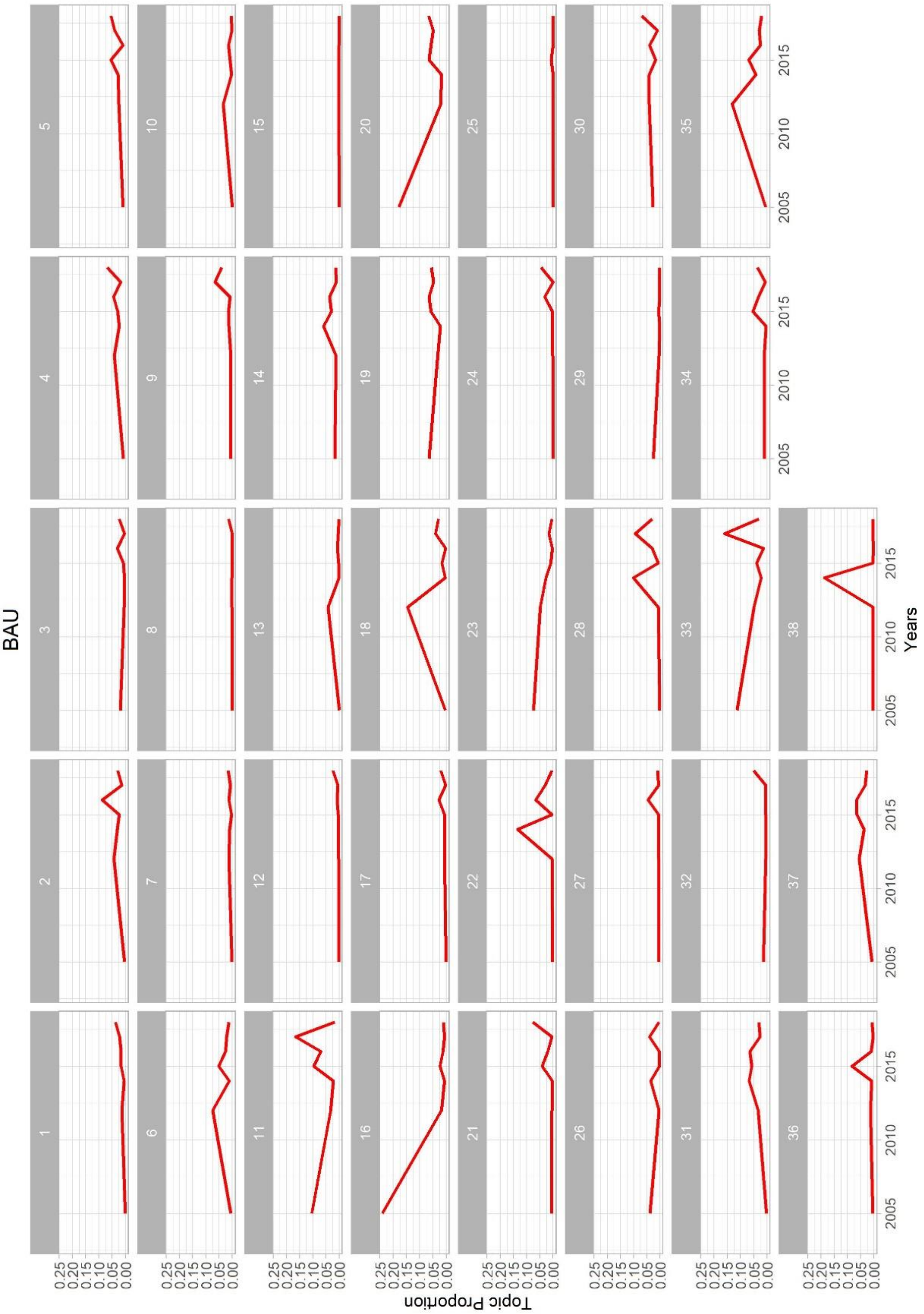
abandon, abil, absenc, absolut, absorb, absorpt, academ, acceler, accept, access, accid, accommod, accomod, accompani, accomplish, account, accumul, accur, accuraci, achiev, acknowledg, acoust, acquir, act, action, activ, actor, actual, adana, adapt, add, addit, address, adequ, adjac, adjust, administr, adopt, advanc, advantag, advers, advertis, aegean, aesthet, affect, affirm, afford, afore, afterth, aga, age, agenc, agenda, agent, aggreg, agre, agreement, agricultur, ahmet, aid, air, airport, algorithm, ali, alien, aliv, alloc, allow, alter, altern, ambigu, america, american, amount, anal, analog, analyt, analyz, anatolia, anatolian, ancient, andor, andth, angl, anim, ankara, annex, annual, answer, antalya, anticip, antiqu, apart, appar, appeal, appear, appendix, appli, applic, appreci, approach, appropri, approv, arch, archaeolog, archeolog, architect, architectur, archiv, arena, argu, argument, aris, armenian, arrang, arriv, art, articl, articul, artifact, artifici, artist, ascertain, asia, aspect, assembl, assert, assess, asset, assign, assist, associ, assum, assumpt, ataturk, ation, atmospher, attach, attain, attempt, attent, attitud, attract, attribut, audienc, auditori, augment, authent, author, autom, autonom, avail, avant, avenu, averag, avoid, awar, award, axe, axi, background, bad, balanc, balconi, bank, barrier, base, basement, basi, basic, bath, bazaar, beam, bear, beauti, begin, begun, behalf, behavior, behaviour, belief, believ, belong, benefici, benefit, berlin, bey, beyoglu, bigger, biggest, bim, biolog, birth, black, block, board, bodi, bodili, bond, book, border, born, bosphorus, bound, boundari, branch, brand, break, breeam, brick, bridg, bring, broad, broader, brought, budget, build, builder, built, bursa, busi, byon, byth, byzantin, cad, calcul, call, camera, campus, capabl, capac, capit, capitalist, captur, car, carbon, care, career, carri, castl, catalog, catalogu, categor, categori, caus, ceil, cell, cement, center, centr, central, centuri, certif, certifi, chain, challeng, chamber, chanc, chang, changeabl, channel, charact, character, characterist, charg, chart, check, chemic, child, children, choic, choos, choosen, chosen, christian, chronolog, church, cinema, circl, circul, circumst, cite, citi, citizen, civil, clad, clarifi, class, classic, classif, classifi, classroom, clean, cled, client, climat, close, closer, cloth, cls, clue, coast, coastal, coat, code, coeffici, coexist, cognit, coher, cold, collabor, collaps, collect, coloni, color, colour, column, combin, come, comfort, comment, commerc, commerci, commiss, commit, common, communic, communiti, compani, compar, comparison, compat, compet, competit, compil, complet, complex, compli, complianc, complic, compon, compos, composit, comprehend, comprehens, compress, compris, comput, conceiv, concentr, concept, conceptu, concern, conclud, conclus, concret, condit, conduct, confer, configur, confirm, conflict, conform, confront, connect, conscious, consequ, conserv, consid, consider, consist, consolid, constant, constitu, constitut, constraint, construct, consult, consum, consumpt, contact, contain, contemporari, content, context, contextu, contin, continu, contract, contractor, contradict, contrari, contrast, contribut, control, controversi, conveni, convent, convers, convert, convey, cool, cooper, coordin, cope, corbusi, core, corner, corpor, correct, correl, correspond, corrupt, cost, council, count, counter, counti, countri, coupl, cours, courtyard, cover, craft, creat, creation, creativ, crisi, criteria, criterion, critic, cross, crowd, crucial, cultur, culturel, current, curtain, curv, custom, cut, cycl, daili, damag, danger, data, databas, date, day, daylight, deal, dealt, death, debat, decad, decay, decid, deciph, decis, declar, declin, deconstruct, decor, decreas, dedic, deep, deeper, deepli, defect, defend, defens, defici, defin, definit, deform, degrad, degre, delay, deleuz, deliv, deliveri, demand, democrat, demograph, demolish, demolit, demonstr, dens, densiti, depart, depend, depict, deplet, depth, deriv, describ, descript, dese, design, desir, destroy, destruct, detach, detail, detect, deterior, determin, develop, devic, devot, diagram, dialect, dialogu, differ, differenti, difficult, difficulti, diffus, digit, dimens, dimension, dioxid, direct, director, disabl, disadvantag, disappear, disast, disciplin, disciplinari, disclos, discours, discov, discoveri, displac, display, disput, disrupt, dissert, distanc, distinct, distinguish, distribut, district, disturb, divers, diversifi, divid, divis, diyarbakir, document, domain, dome, domest, domin, door, doubl, dramat, draw, drawn, dri, drive, driven, dualiti, durabl, durat, duti, dwell, dweller, dynam, earlier, earliest, earn, earth, earthquak, eas, easi, easier, easili, east, eastern, eat, eco, ecolog, econom, economi, ecosystem, edg, edirn, edit, educ, effect, effici, effort, eighth, elabor, elast, electr, electron, element, elev, elimin, embodi, embrac, emerg, emiss, emot, emphas, emphasi, emphasis, empir, employ, employe, enabl, enclos, encount, encourag, energi, engag, engin, england, enhanc, enjoy, enlarg, enlighten, enorm, enrich, ensur, enter, enterpris, entertain, entir, entiti, entitl, entranc, entri, envelop, enviro, enviroment, environ, environment, equal, equip, equival, era, erect, error, eskisehir, essenc, essenti, establish, estat, esthet, estim, ethnic, europ, european, evalu, event, eventu, everyday, evid, evolut, evolutionari, evolv, exact, excav, exceed, except, excess, exchang, exclud, execut, exemplifi, exhaust, exhibit, exist, expand, expans, expect, expens, experi, experienc, experiment, expert, expertis, explan, explic, exploit, explor, expos, exposur, express, extend, extens, extent, exterior, extern, extinct, extra, extract, extrem, eye, fabric, facad, face, facil, facilit, factor, factori, faculti, fail, failur, fall, famili, familiar, famous, fashion, fast, faster, fatih, fault, favor, feasibl, featur, feed, feedback, feel, fiber, fiction, field, fieldwork, figur, file, fill, film, filter, final, financ, financi, fine, finish, fire, firm, first, fit, fix, flat, flexibl, floor, flow, focal, follow, food, forc, fore, forefront, foreign, forest, forgotten, form, formal, format, formul, fort, fossil, found, foundat, fountain, fourth, fragment, frame, framework, franc, freedom, french, frequenc, frequent, friend, fuel, fulfil, function, fund, fundament, furnitur, futur, gain, galata, galleri, game, gap, gard, garden, gas, gase, gate, gather, gaziantep, gender, general, generat, geograph, geographi, geolog, geometr, geometri, german, germani, glass, glaze, global, goal, golden, govern, government, grade, gradual, graduat, grammar, grand, graph, graphic, grasp, great, greec, greek, green, greenhous, grid, ground, grow, grown, growth, guid, guidanc, guidelin, habit, habitat, hall, han, hand, handl, happen, harbor, hard, harm, harmon, harmoni, hazard, head, headlin, health, healthi, heat, heavi, heavili, height, held, help, heritag, hidden, hierarch, hierarchi, high, highlight, highway, hill, hisher, histor, histori, historian, hold, holist, home, homogen, hope, horizont, horn, hospit, host, hot, hotel, hour, hous, household, huge, human, humankind, humid, hybrid, hypo, icon, idea, ideal, ident, identif, identifi, ideolog, ignor, iii, illumin, illustr, imag, imagin, imaginari, imit, immigr, immov, impact, imperi, implement, impli, implic, import, impos, imposs, impress, improp, improv, inadequ, inadequaci, incent, incid, inclin, includ, inclus, incom, incorpor, increas, independ, indic, indirect, indispens, individu, indoor, industri, inevit, infer, influenc, influenti, inform, infrastructur, ing, inhabit, inher, inherit, initi, inn, innov, input, inquir, inquiri, insight, inspect, inspir, instal, instanc, instant, institut, instruct, instrument, insuffici, insul, intang, integr, intellectu, intellig, intend, intens, intensifi, intent, inter, interact, interdisciplinari, interfac, interfer, interior, intern, internet, interpret, interrel, interrog, intersect, intertwin, interv, intervent, interview, inth, introduc, introductori, intuit, invent, inventori, invest, investig, investor, invis, involv, ion, iran, iron, irregular, islam, island, iso, isol, issu, istanbul, itali, italian, item, izmir, japan, job, join, joint, journal, journey, kadikoy, kayseri, keep, key, keyword, khan, kind, kinet, kingdom, kitchen, knowledg, konya, labor, laboratori, labour, lack, land, landmark, landscap, languag, larger, largest, last, late, later, law, lay, layer, layout, lead, leadership, learn, leav, led, leed, left, legal, legisl, lesson, level, liber, librari, lie, life, lifestyl, lifetim, light, lightweight, limit, line, linear, link, list, listen, literari, literatur, livabl, live, load, local, locat, lodg, logic, look, loos, lose, loss, lost, lot, lowest, luxuri, lynch, machin, macro, magazin, main, maintain, mainten, major, maker, mall, manag, manifest, manipul, mankind, manner, mansion, manual, manufactur, map, mardin, mark, market, marmara, masonri, mass, massiv, master, match, materi, mathemat, matrix, matter, maximum, mean, meaning, meant, measur, mechan, media, mediat, medic, mediev, mediterranean, medium, meet, mehmet, memori, mental, mere, merg, mersin, messag, met, metal, metaphor, meter, metropoli, metropolis, metropolitan, metu, micro, mid, middl, migrat, militari, mimar, mind, minim, minimum, ministri, minor, miss, mission, mistak, mix, mobil, mode, model, modern, modernist, modif, modifi, modul, modular, moistur, moment, money, monitor, month, monument, morpholog, mortar, mosqu, motion, motiv, mountain, move, movement, movi, multi, multidisciplinari, multipl, municip, museum, music, muslim, mustafa, mutual, name, narrat, narrow, nation, natur, nearbi, necess, necessit, negat, neglect, negoti, neighbor, neighborhood, neighbourhood, neo, network, newli, newspap, night, nineteenth, node, nois, norm, normal, north, northern, notabl, note, notic, notion, nowaday, number, numer, object, oblig, observ, obstacl, obtain, obvious, occup, occupi, occur, occurr, offer, offic, offici, ofth, oil, olog, ongo, ontolog, open, oper, opinion, opportun, oppos, opposit, optic, optim, optimum, option, oral, ordinari, organ, organis, organiz, orient, origin, ornament, osman, ottoman, outcom, outdoor, outer, outlin, output, overcom, overlap, overlook, overview, own, owner, ownership, packag, paid, paint, palac, panel, paper, paradigm, parallel, paramet, parametr, parcel, park, part, parti, partial, particip, participatori, partner, partnership, pasa, pasha, pass, passeng, passiv, path, patient, pattern, pave, pavilion, pay, peculiar, pedestrian, penetr, peninsula, peopl, pera, perceiv, percent, percentag, percept, perceptu, perfect, perform, period, perman, permeabl, permiss, permit, person, perspect, peter, phase, phenomena, phenomenolog, phenomenon, philosoph, philosophi, photo, photograph, photographi, physic, physiolog, pictur, piec, pilot, pioneer, placement, plain, plan, plane, planner, plant, plaster, plastic, platform, play, plot, polici, polit, pollut, pool, poor, popul, popular, port, portion, posit, possess, possibl, post, postmodern, potenti, power, practic, practis, pre, precaut, preced, precious, precis, predict, prefabr, prefer, preliminari, premis, prepar, presenc, present, preserv, pressur, prestig, prevail, preval, prevent, previous, price, primari, primit, princip, principl, print, prior, prioriti, privaci, privat, probabl, problemat, procedur, proceed, process, procur, produc, product, profess, profession, profil, profit, profound, program, programm, progress, project, promin, promis, promot, proof, proper, properti, proport, propos, proposit, prospect, prosper, protect, prototyp, prove, provid, provinc, provis, psycholog, public, publish, purchas, pure, purpos, pursu, push, put, qualif, qualifi, qualit, qualiti, quantit, quantiti, quarter, quest, question, questionnair, quick, radiat, radic, railway, rain, rais, random, rang, rank, rapid, rare, rate, ratio, ration, raw, reach, react, reaction, read, reader, readi, real, realis, realist, realiti, realiz, realm, reason, rebuilt, receiv, recogn, recognit, recommend, reconsid, reconstitut, reconstruct, record, recoveri, recreat, rectangular, recycl, redefin, reduc, reduct, refer, reflect, reform, refunct, regard, regener, regim, region, regist, registr, regul, regular, rehabilit, reign, reinforc, reinterpret, relat, relationship, releas, relev, reli, reliabl, relief, religi, religion, remain, remark, rememb, remind, remov, renaiss, render, renew, renov, rent, repair, repeat, repetit, replac, report, repres, represent, reproduc, reproduct, republ, republican, request, requir, resembl, reserv, reshap, resid, residenti, resist, resolut, resolv, resort, resourc, respect, respond, respons, rest, restaur, restitut, restor, restrict, restructur, result, return, reus, reveal, reverber, revers, review, revis, revit, reviv, revolut, rich, right, rigid, rise, risk, river, road, rock, role, roman, rome, roof, root, rout, routin, ruin, rule, run, rural, safe, safeti, sake, sale, sampl, satisfact, satisfi, save, scale, scan, scenario, scene, schedul, schema, schemat, scheme, scholar, school, scienc, scientif, scientist, score, screen, scrutin, sculptur, sea, search, season, secondari, section, sector, secur, seek, segment, seismic, select, seljuk, sell, semant, semi, sens, sensat, sensit, sensori, separ, seper, sequenc, seri, serv, servic, set, settl, settlement, setup, seventh, sever, shade, shadow, shape, share, sheet, shelter, shift, shop, shore, short, shortag, sight, sign, signific, silhouett, similar, simpl, simul, simultan, sinan, singl, singular, sit, site, situ, situat, sixth, size, sketch, skill, skin, slope, slow, slum, smart, social, societi, socio, sociocultur, sociolog, softwar, soil, solar, sold, sole, solid, solut, solv, son, sort, sought, sound, sourc, south, space, span, spatial, speak, special, specialist, specif, specul, speech, speed, spend, spent, sphere, spirit, spiritu, spite, sport, spot, spread, spss, squar, stabil, stabl, staff, stage, stair, stakehold, stand, standard, standart, star, start, state, statement, static, station, statist, status, stay, steel, stem, step, stimul, stock, stone, stop, storag, store, storey, stori, strateg, strategi, stratif, stream, street, strength, strengthen, stress, strict, strong, structur, struggl, student, style, subject, submit, subsequ, substanc, substanti, substructur, subtitl, succeed, success, suffer, suffici, suggest, suitabl, sultan, sum, summar, summari, summer, sun, superior, supervis, suppli, supplier, support, suppos, surfac, surround, survey, surviv, sustain, symbol, symmetr, syn, syntax, system, systemat, tabl, tactic, take, talk, tall, tangibl, tanzimat, target, task, teach, teacher, team, technic, techniqu, technolog, tecton, temper, temperatur, templ, tempor, temporari, tend, tendenc, tender, tension, term, termin, terminolog, territori, test, text, textur, theater, theatr, theme, theoret, theori, thermal, these, thick, think, threat, threaten, threshold, tie, tile, timber, time, tissu, titl, today, toki, told, tomb, tool, topic, topograph, topographi, total, touch, tourism, tourist, tower, town, trabzon, trace, track, trade, tradit, traffic, train, transfer, transform, transit, translat, transmiss, transmit, transpar, transport, travel, treat, treatment, tree, trend, trial, trigger, TRUE, turk, turkey, turkish, twentieth, type, typic, typolog, ultim, uncertainti, unconsci, uncontrol, uncov, under, undergo, undergon, underground, underlin, understand, understood, undertak, undertaken, unexpect, uniform, union, uniqu, unit, uniti, univers, unknown, unplan, unpredict, unqualifi, updat, upper, urban, urgent, usa, usabl, usag, user, uskudar, util, utopia, valid, valu, valuabl, vanish, vari, variabl, variat, varieti, vast, vault, veget, vehicl, ventil, venu, verbal, verifi, vernacular, version, vertic, vicin, video, view, viewpoint, vii, villag, virtual, visibl, vision, visit, visitor, visual, vital, voic, void, volum, wait, walk, wall, war, warm, wast, watch, water, weak, wealth, weather, week, weight, west, western, wide, wider, widespread, win, wind, window, winter, wit, women, wood, wooden, word, worker, workmanship, workshop, world, worldwid, worth, write, writer, written, wrong, yield, yildiz, zone

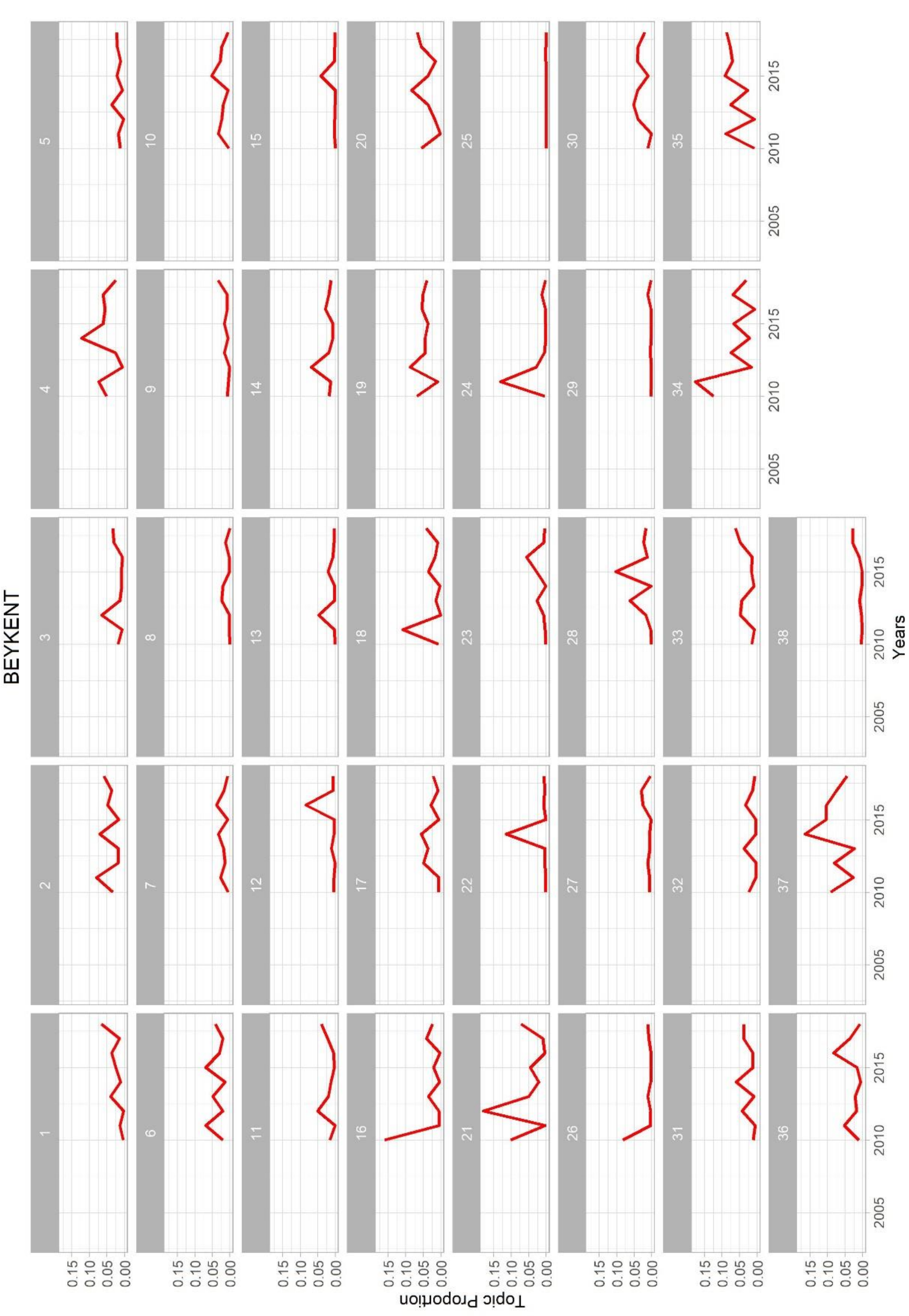
B. Yearly Expected Proportions of Universities

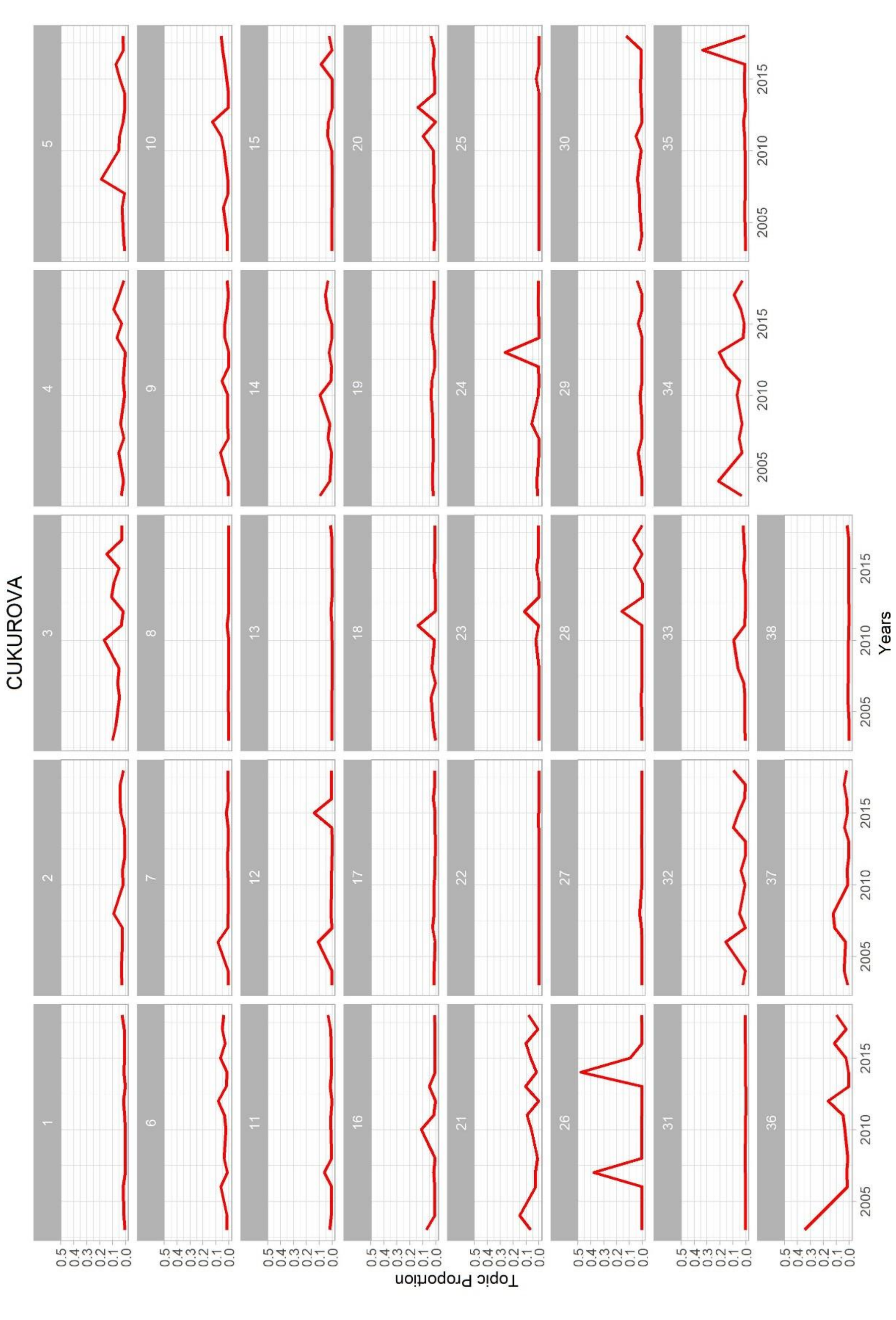



AREL
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
0.4
0.3
0.2
0.1
0.0
Topic Proportion
2005
2010
2015
Years

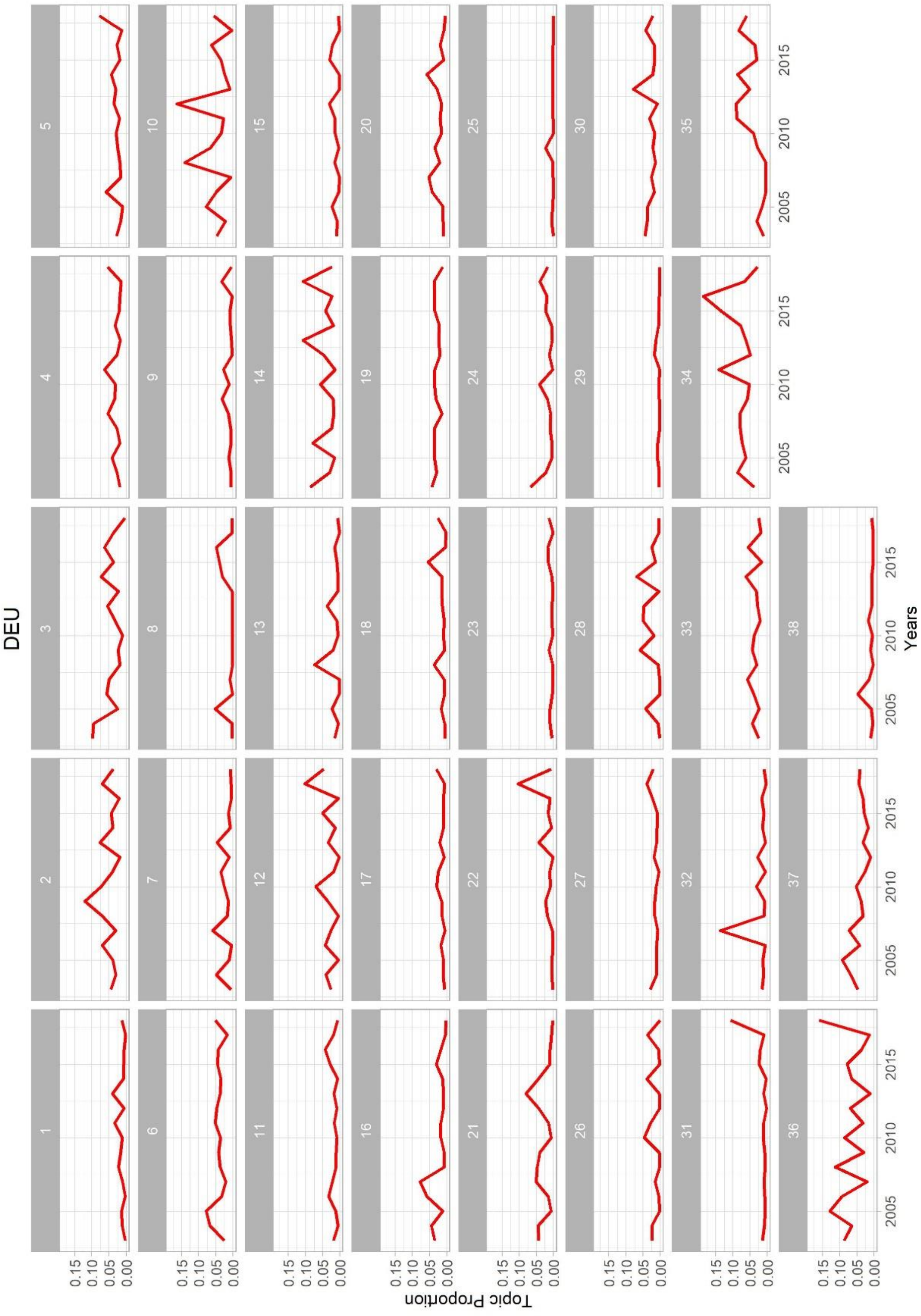

BALIKESIR
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
Topic Proportion
0.4
0.2
0.0
2005
2010
2015
Years

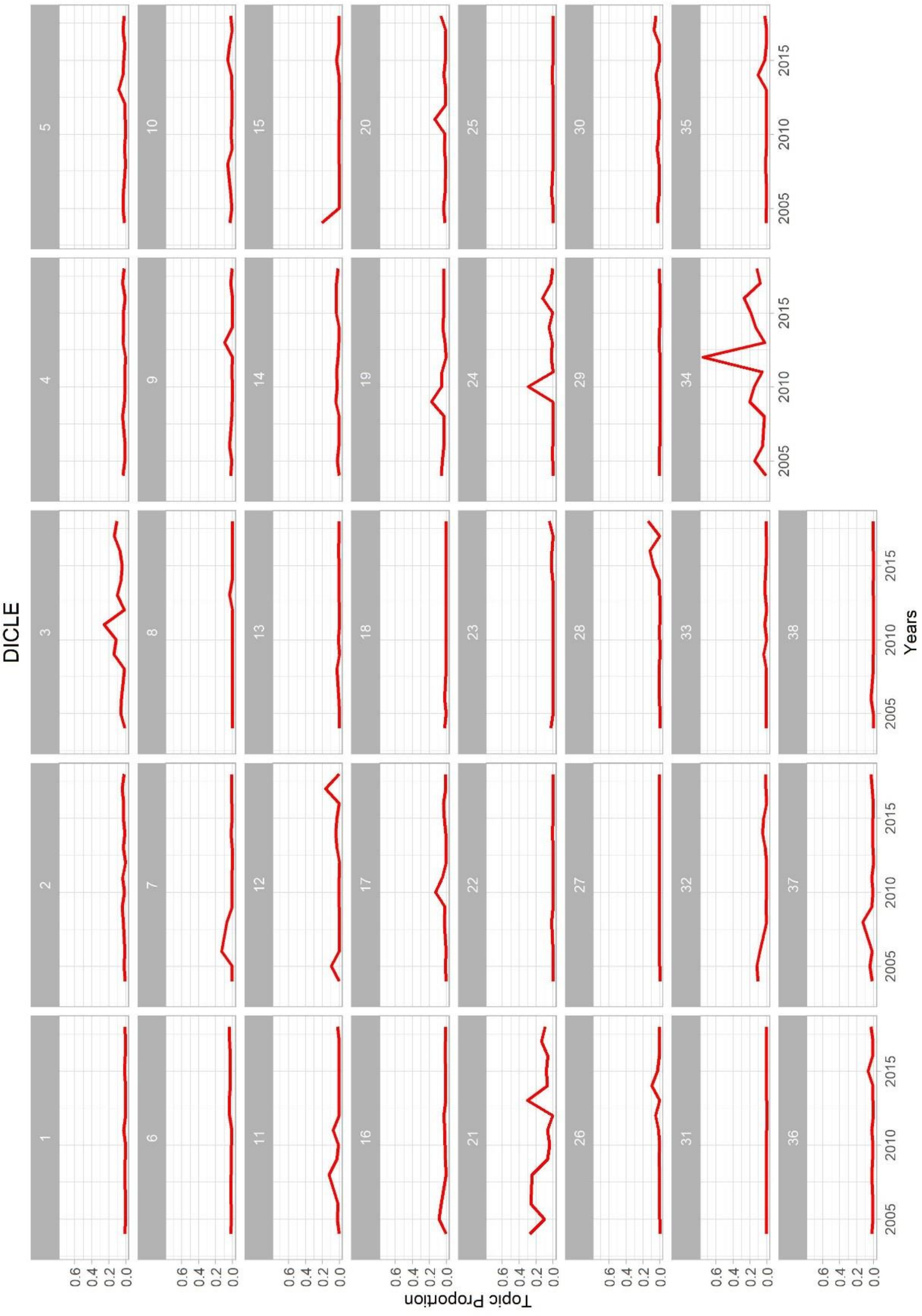

BAU
Topic Proportion
Years
0.25
0.20
0.15
0.10
0.05
0.00
2005
2010
2015
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38

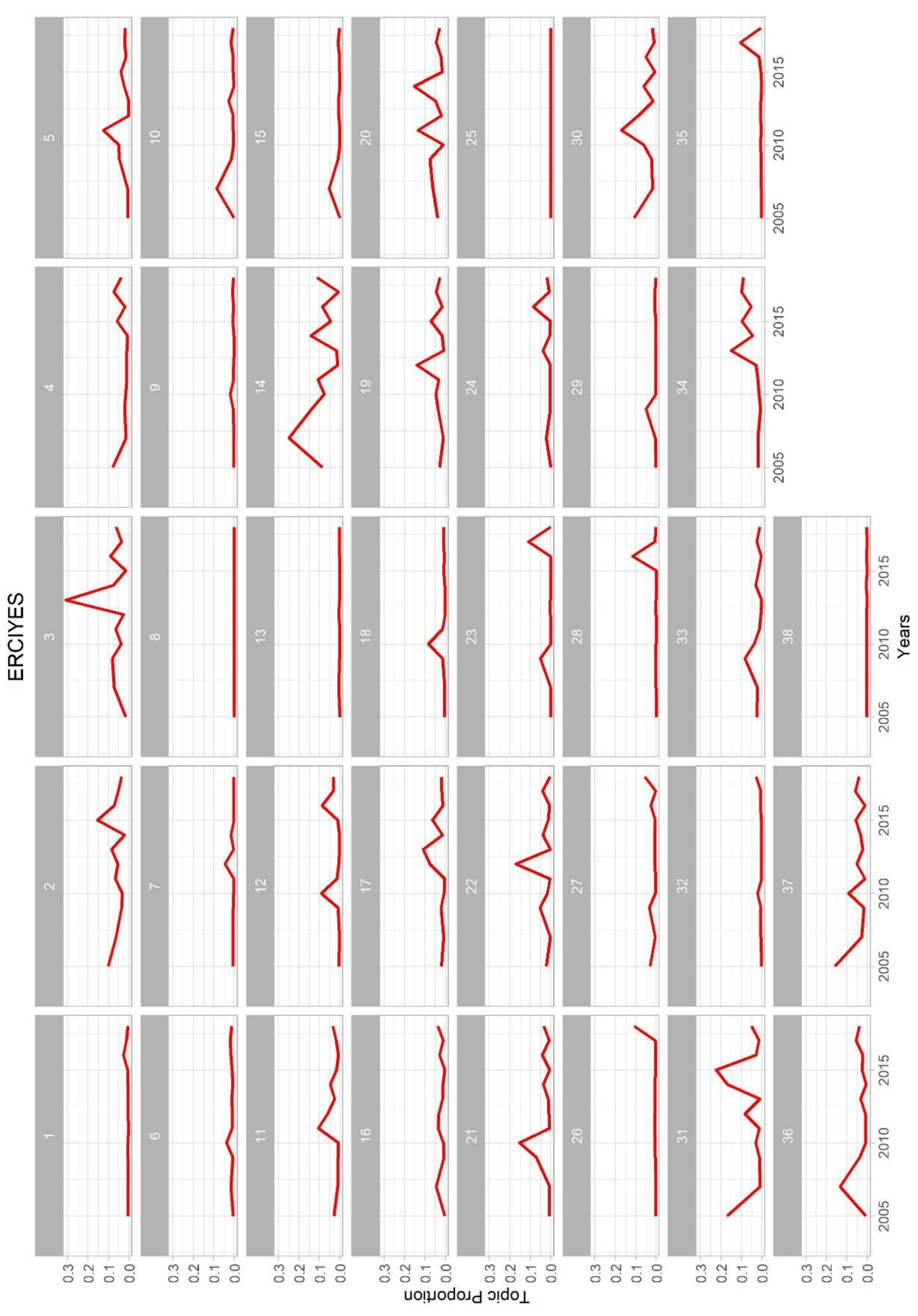

BEYKENT
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
0.15
0.10
0.05
0.00
2005
2010
2015
Years
Topic Proportion

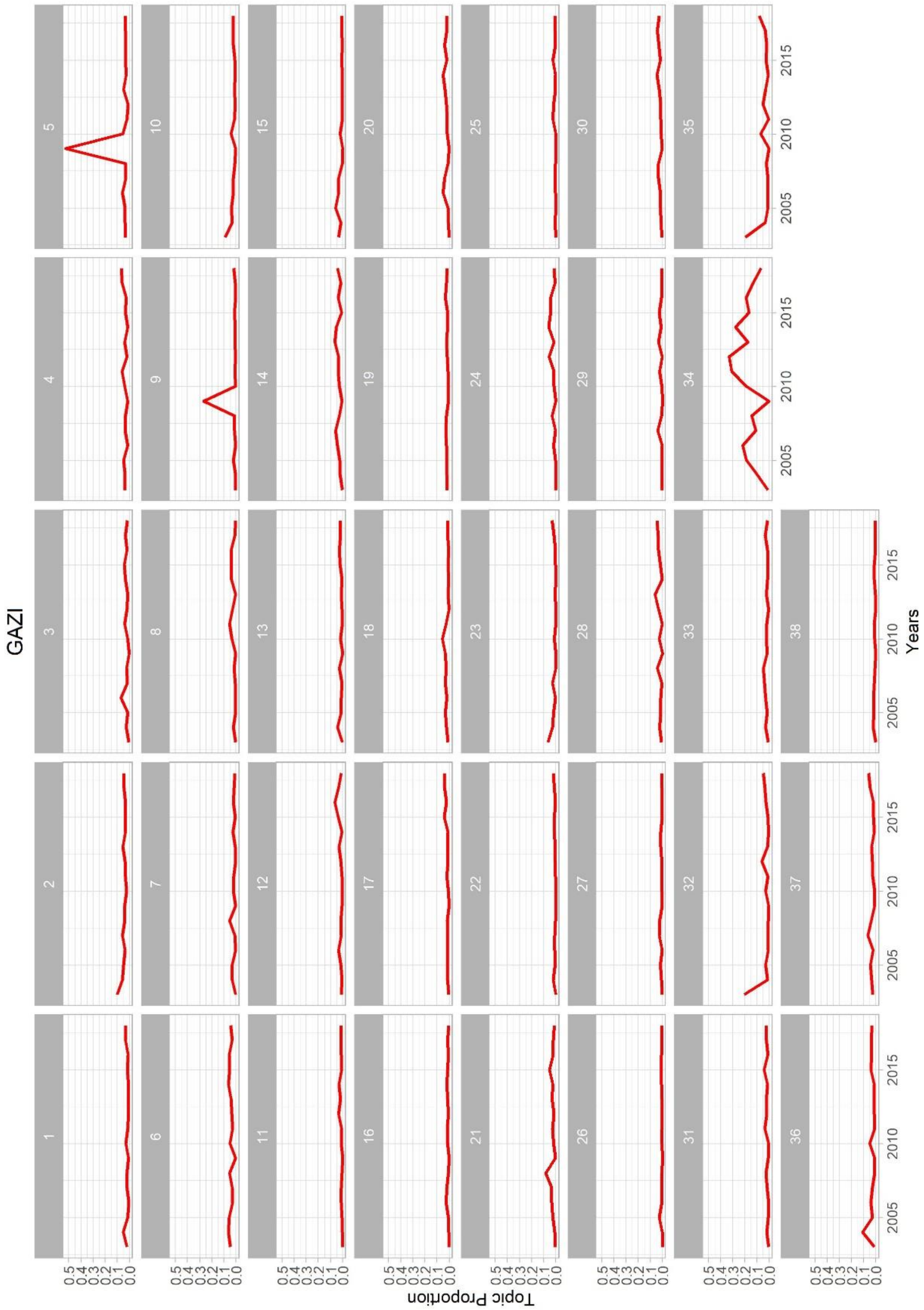

CUKUROVA

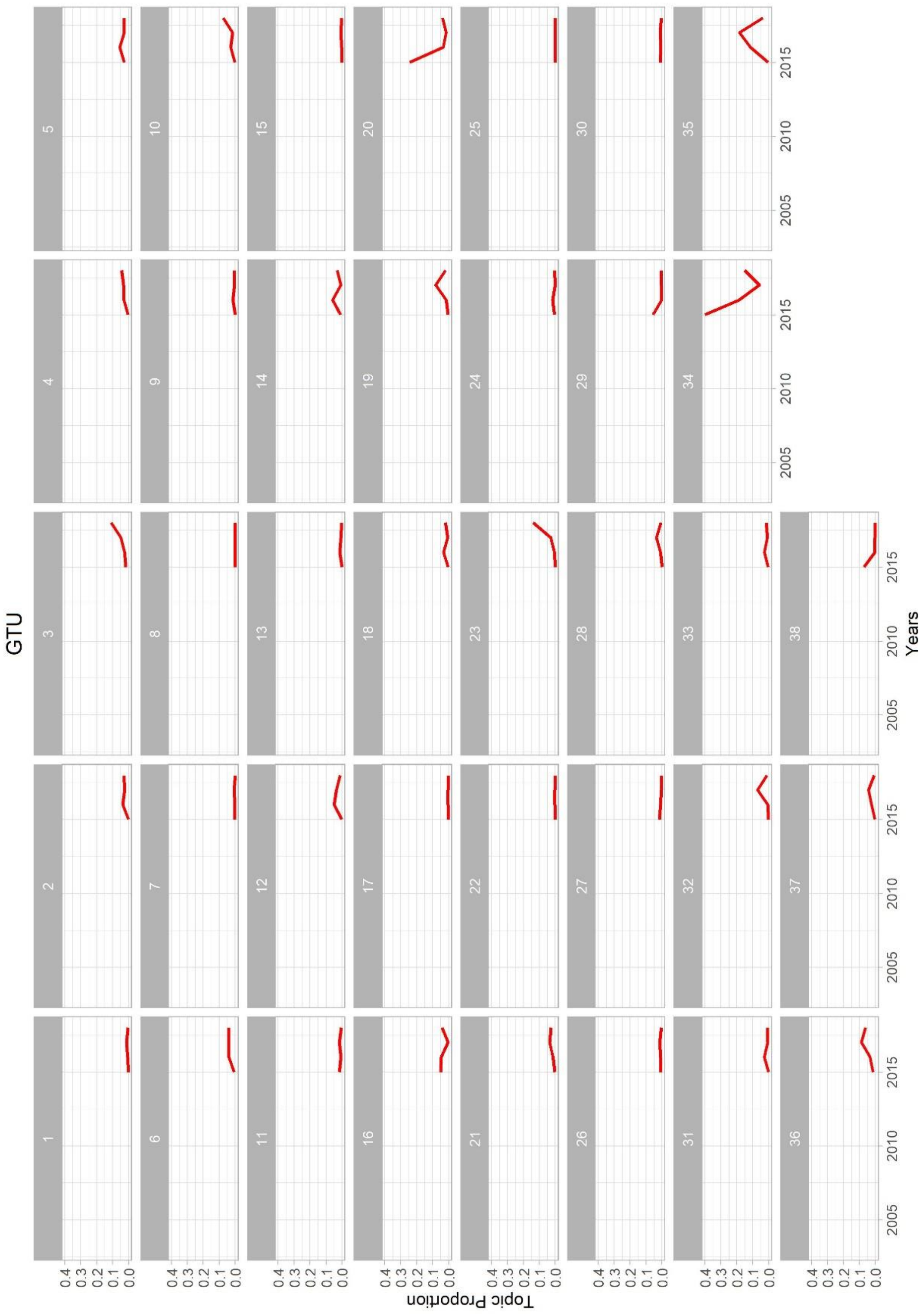
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
0.0
0.1
0.2
0.3
0.4
0.5
Topic Proportion
2005
2010
2015
Years

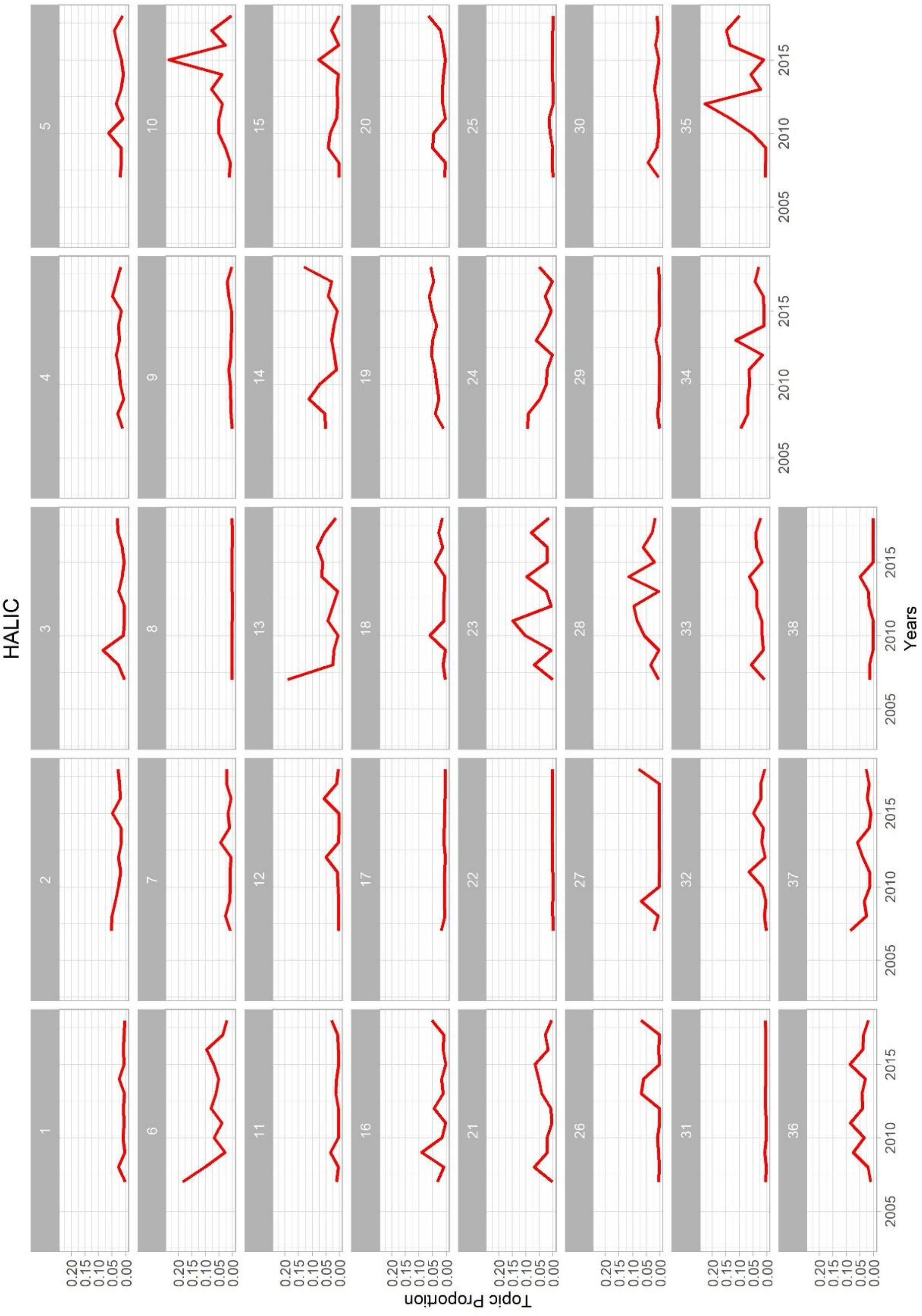

DEU
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
0.15
0.10
0.05
0.00
2005
2010
2015
Years
Topic Proportion

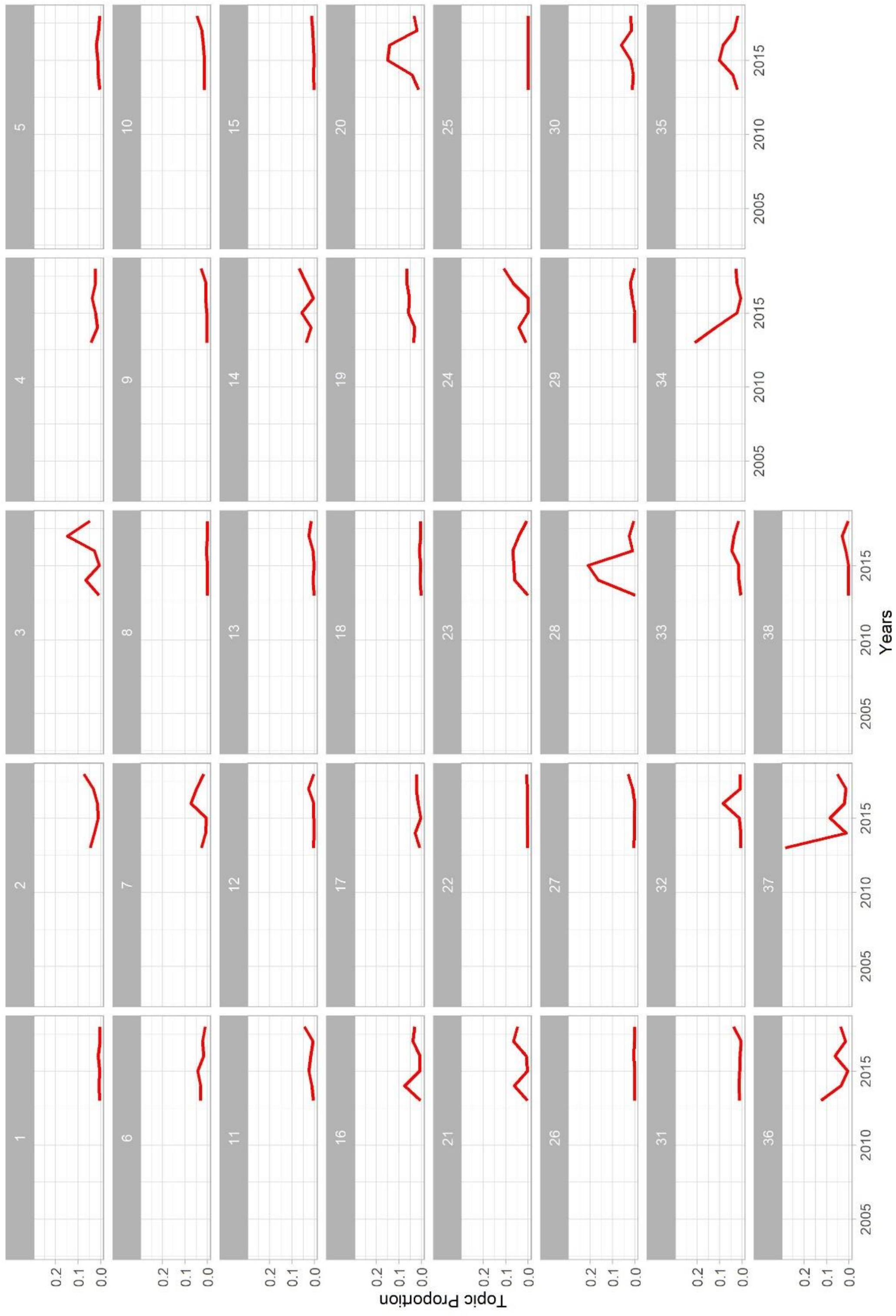

DICLE
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
0.0
0.2
0.4
0.6
2005
2010
2015
Years
Topic Proportion

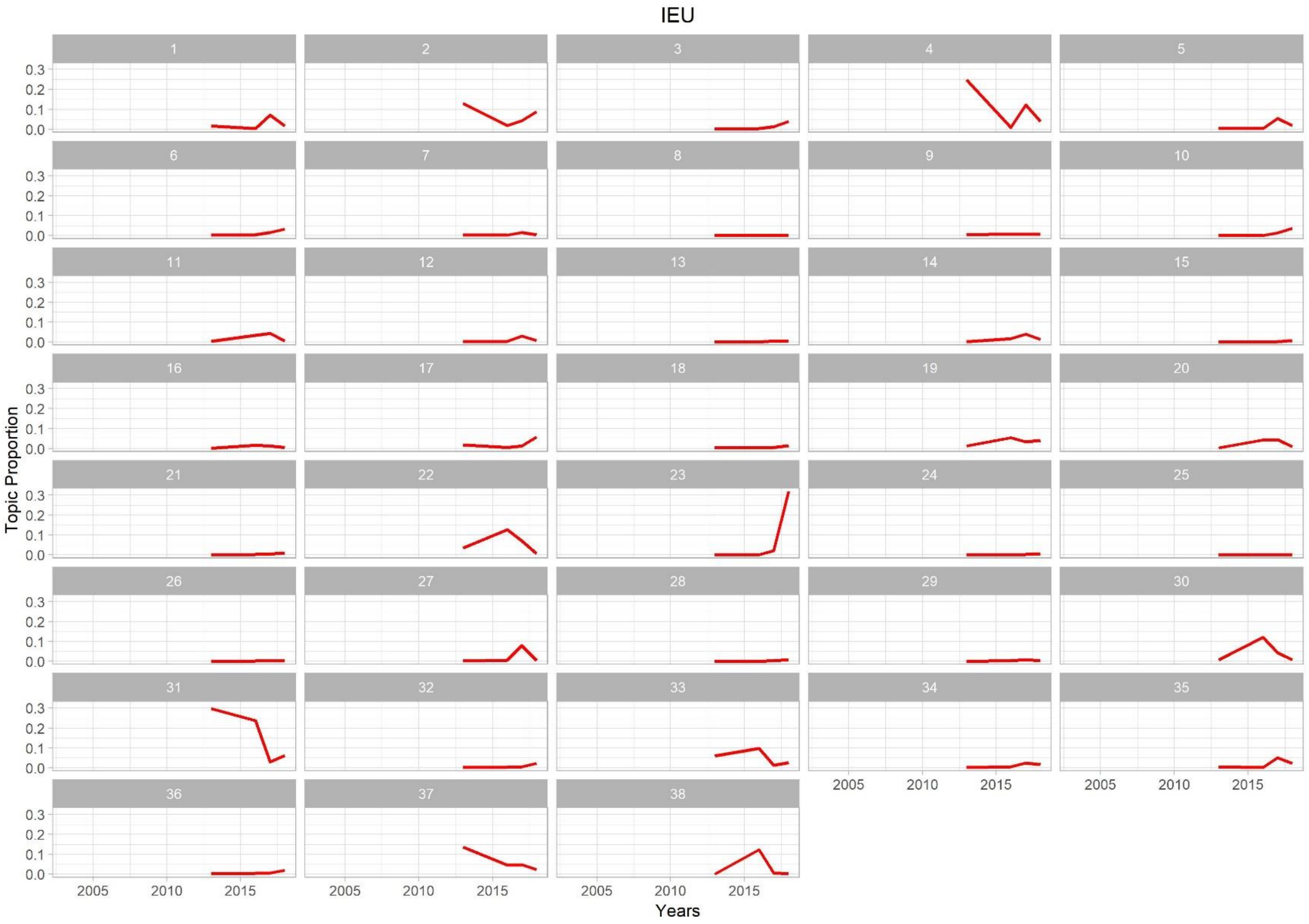

ERCIYES
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
Topic Proportion
0.0
0.1
0.2
0.3
Years
2005
2010
2015

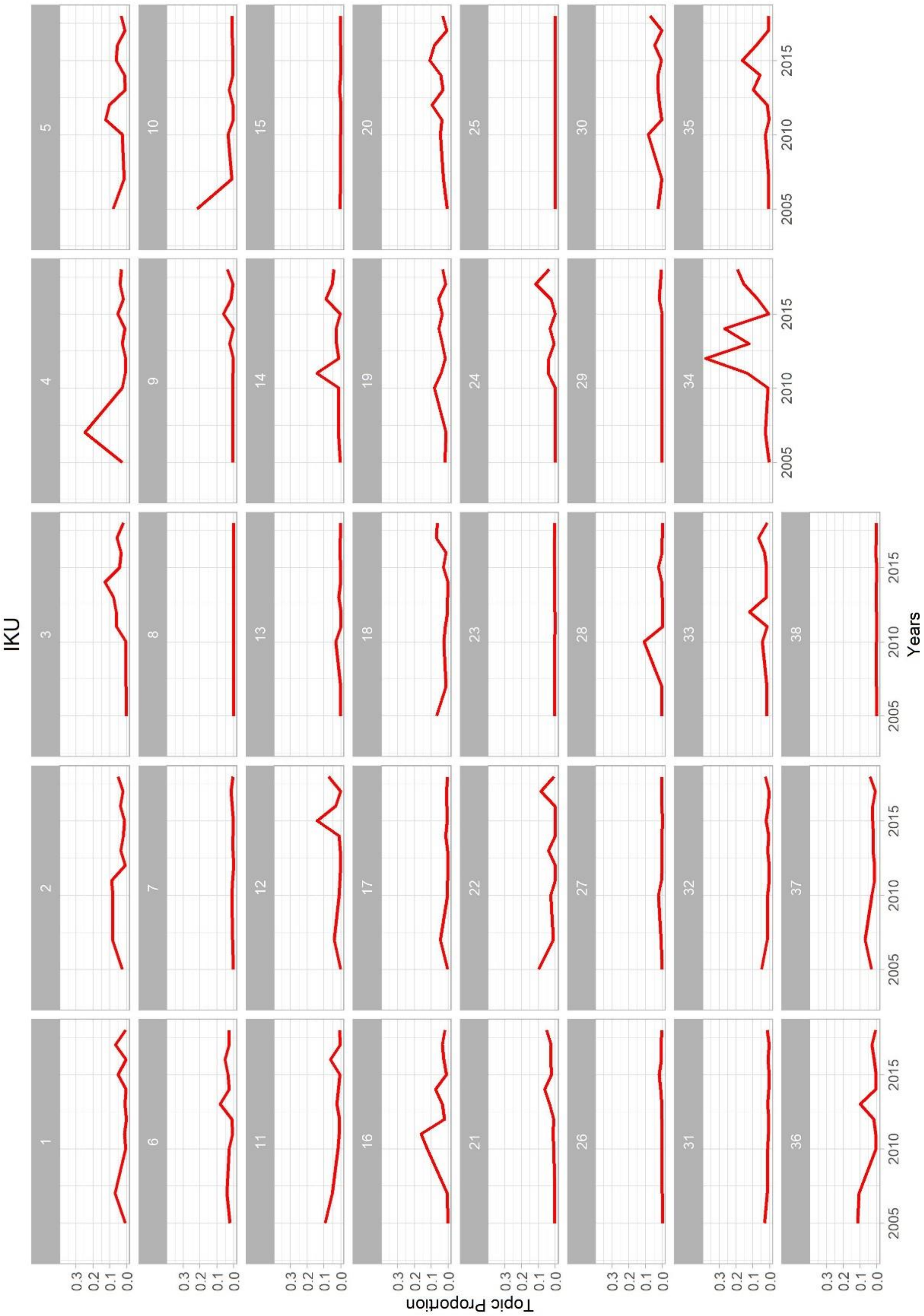

GAZI
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
0.0
0.1
0.2
0.3
0.4
0.5
Topic Proportion
2005
2010
2015
Years

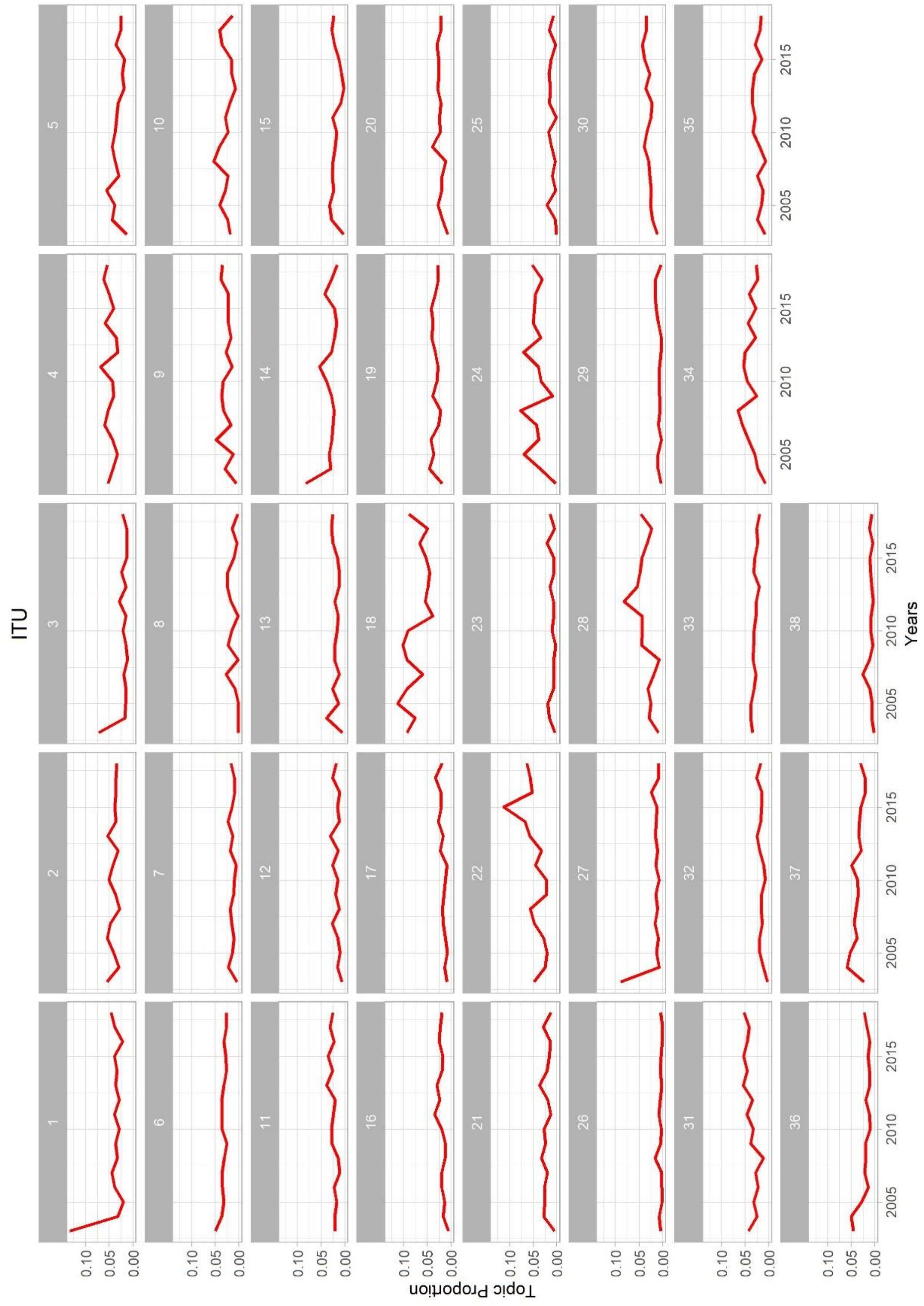

GTU
Topic Proportion
Years
0.0
0.1
0.2
0.3
0.4
2005
2010
2015
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38

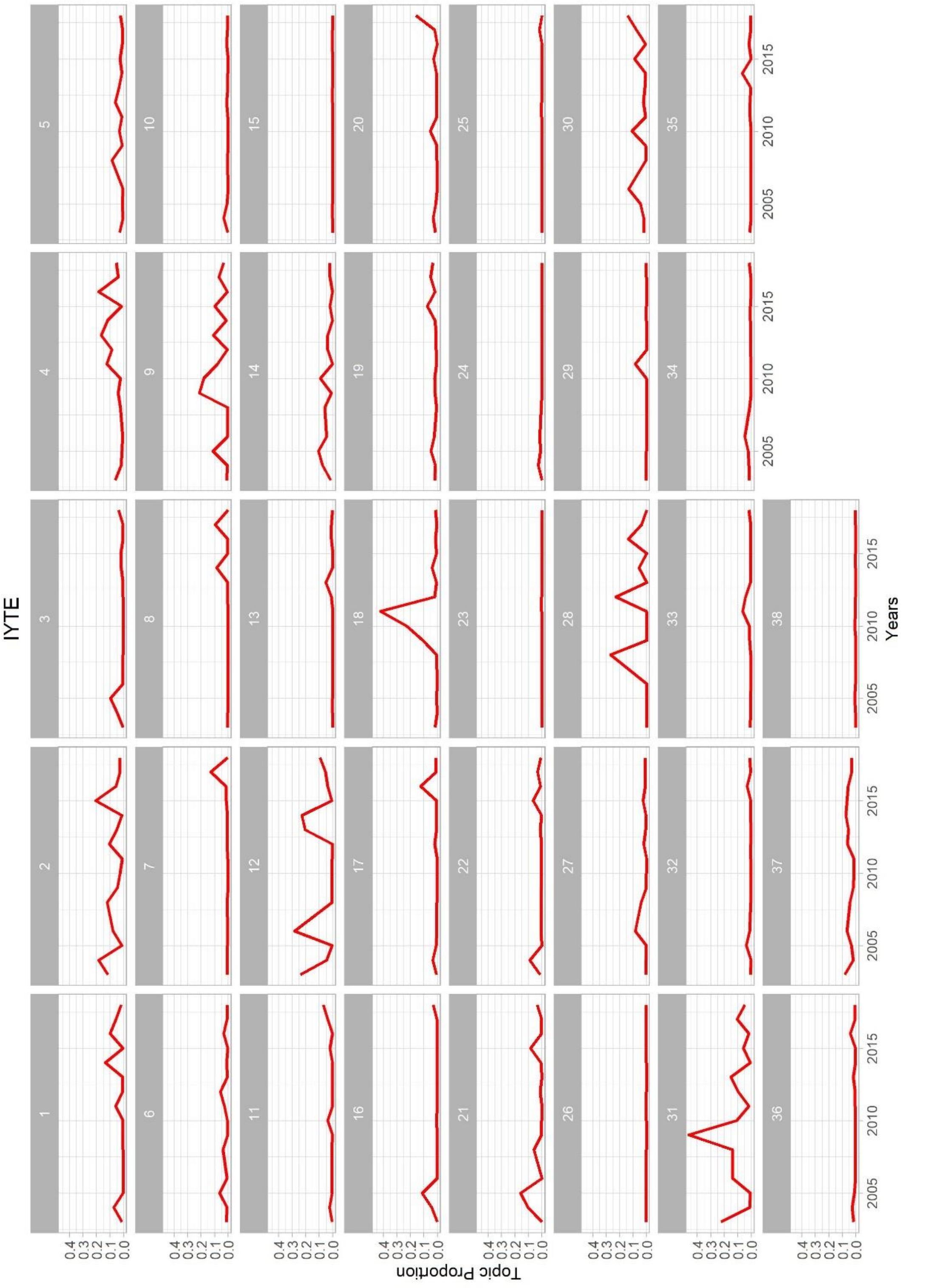

HALIC
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
0.20
0.15
0.10
0.05
0.00
2005
2010
2015
Years
Topic Proportion

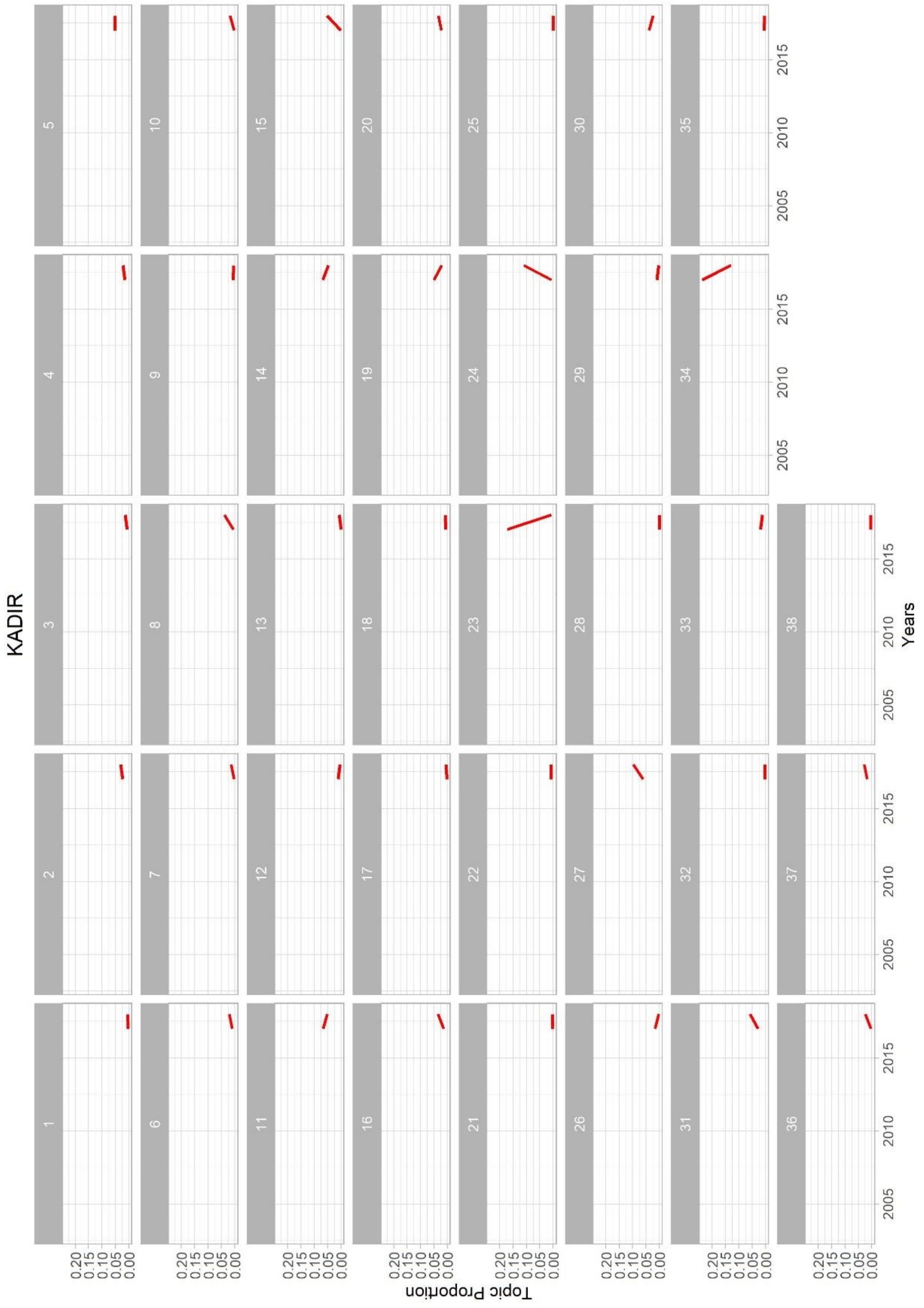

IAU
Topic Proportion
Years
0.0
0.1
0.2
2005
2010
2015
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38


IEU
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
Topic Proportion
0.0
0.1
0.2
0.3
2005
2010
2015
Years

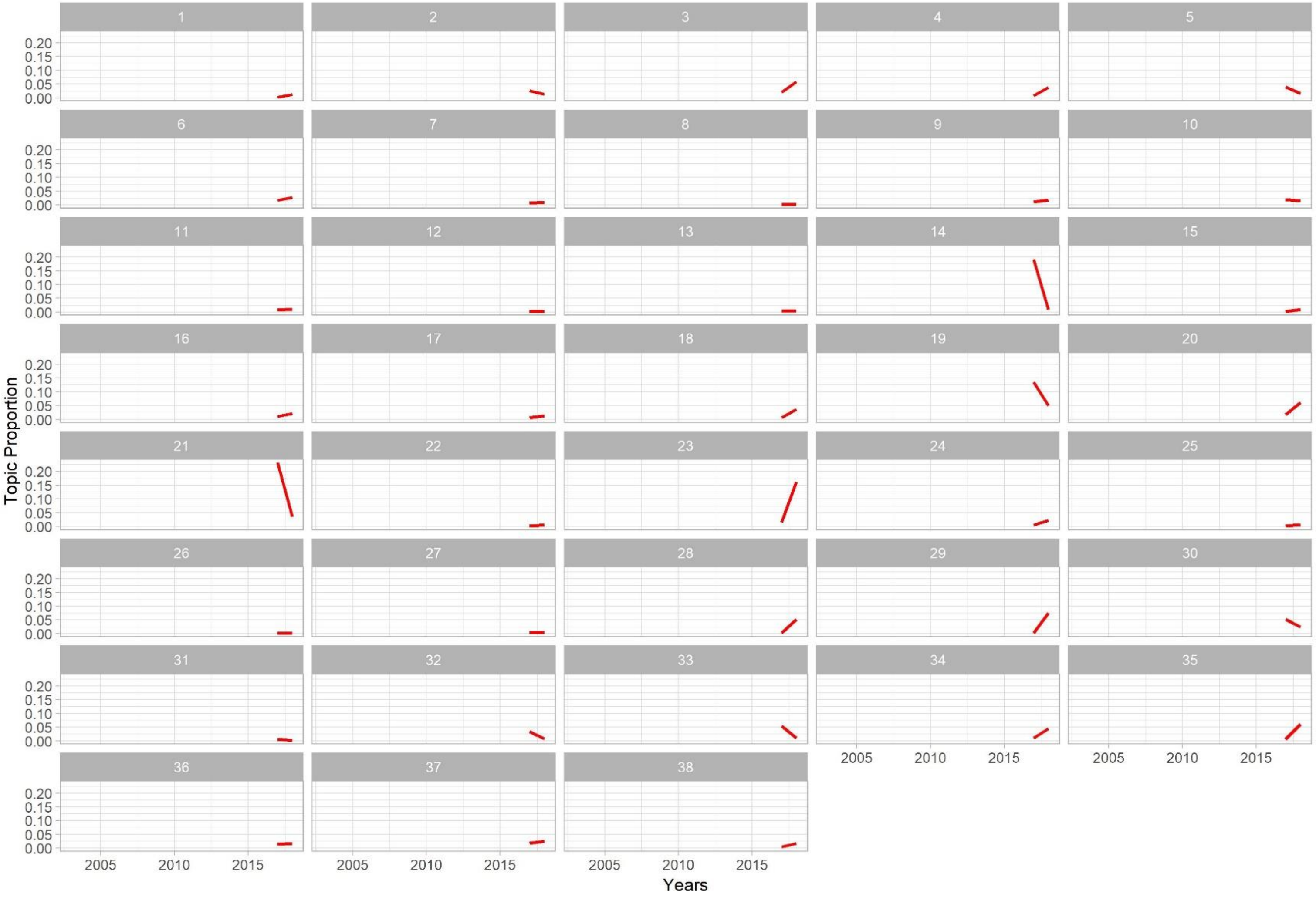

IKU
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
Topic Proportion
0.0
0.1
0.2
0.3
Years
2005
2010
2015

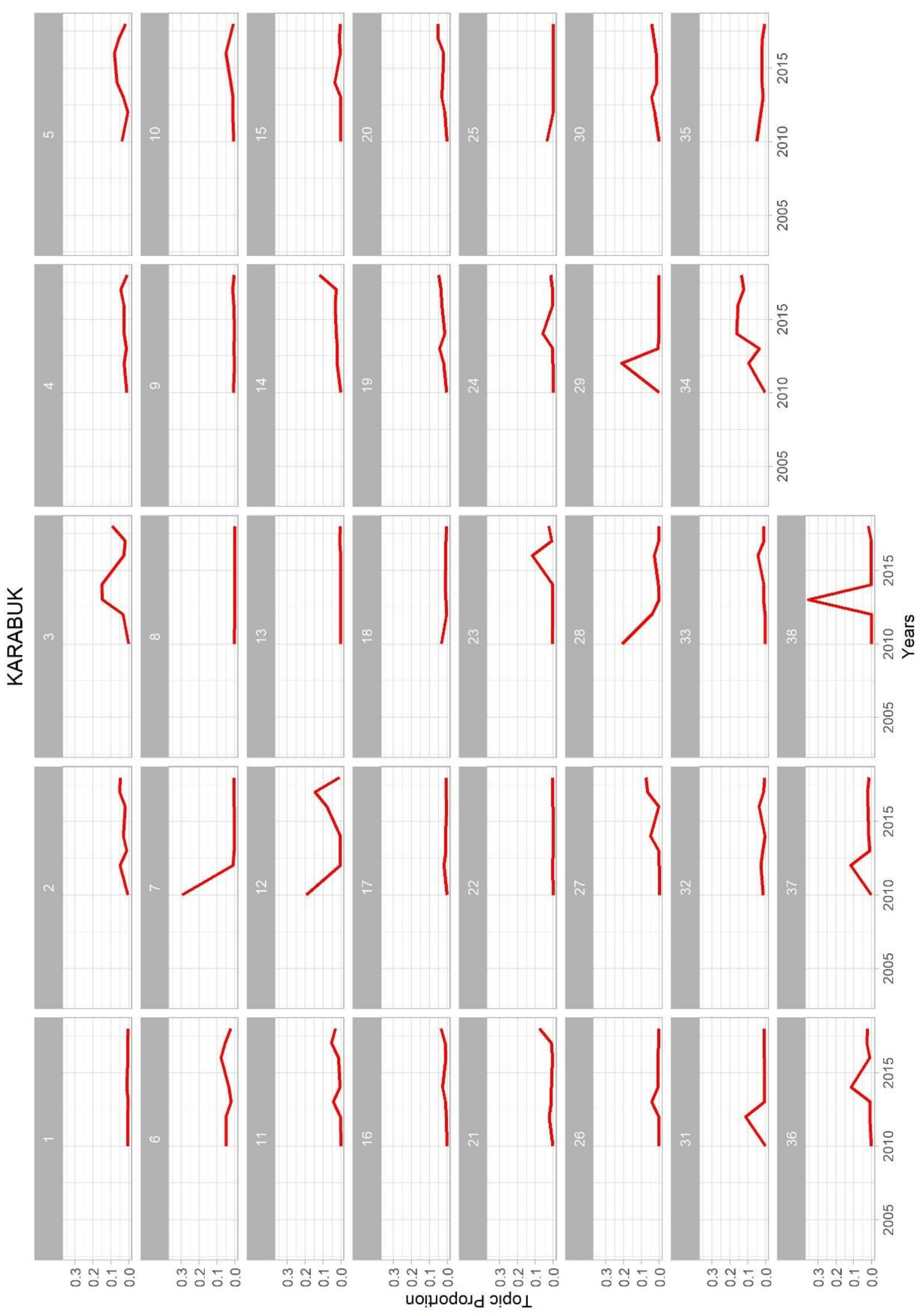

ITU
Topic Proportion
Years
0.00
0.05
0.10
2005
2010
2015
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38

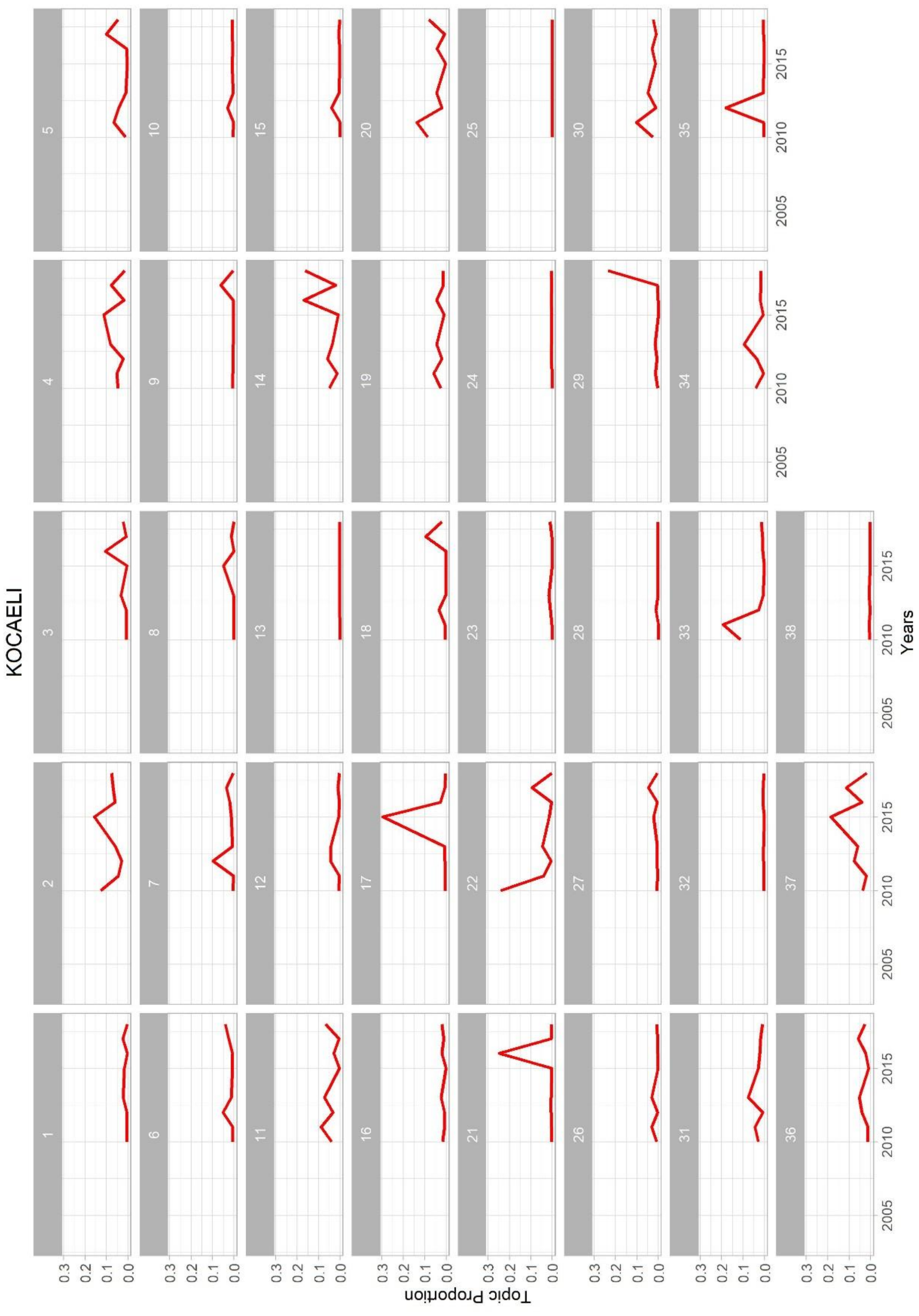

IYTE
Topic Proportion
Years
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
0.0
0.1
0.2
0.3
0.4
2005
2010
2015

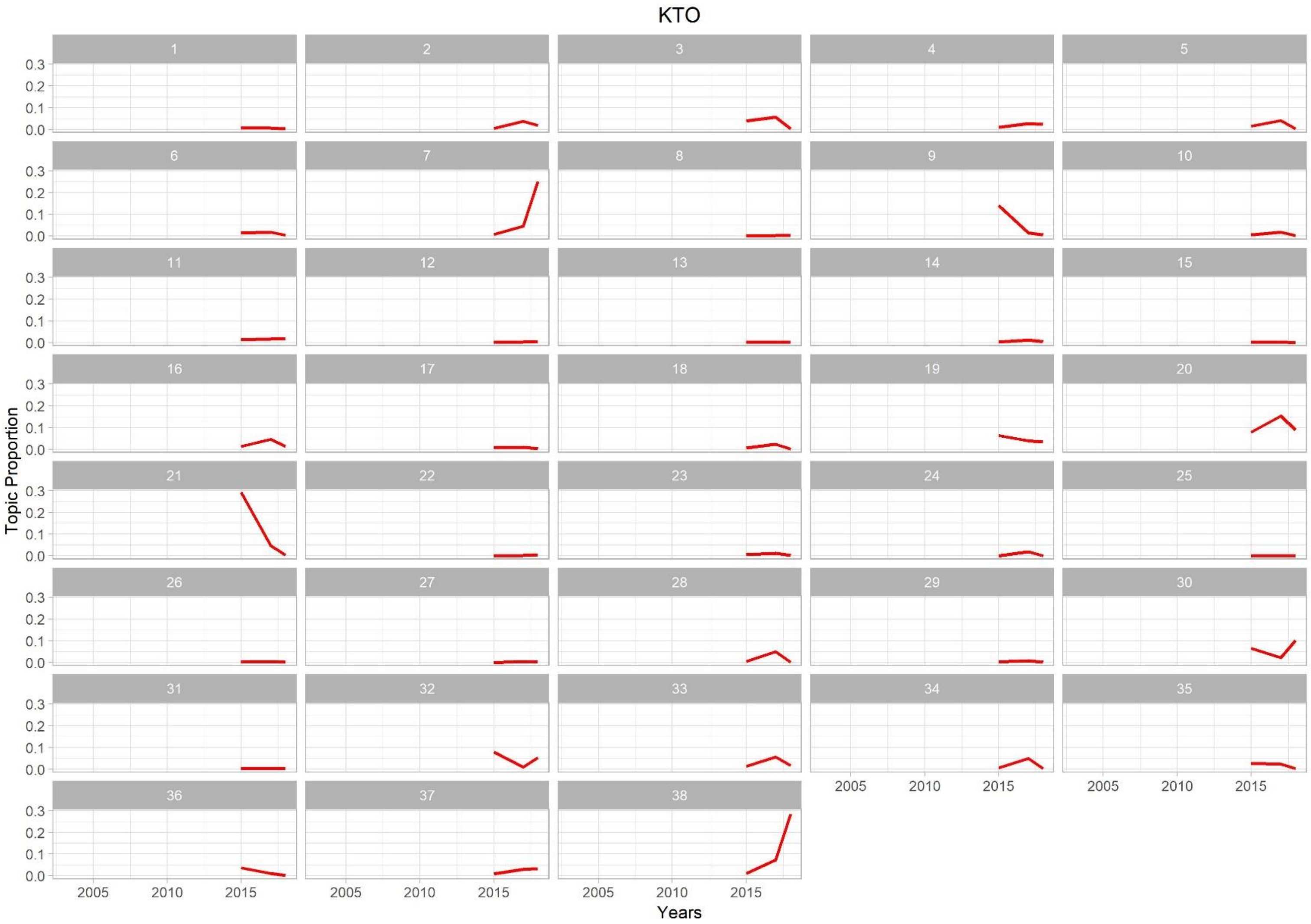

KADIR
Topic Proportion
Years
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
0.00
0.05
0.10
0.15
0.20
2005
2010
2015

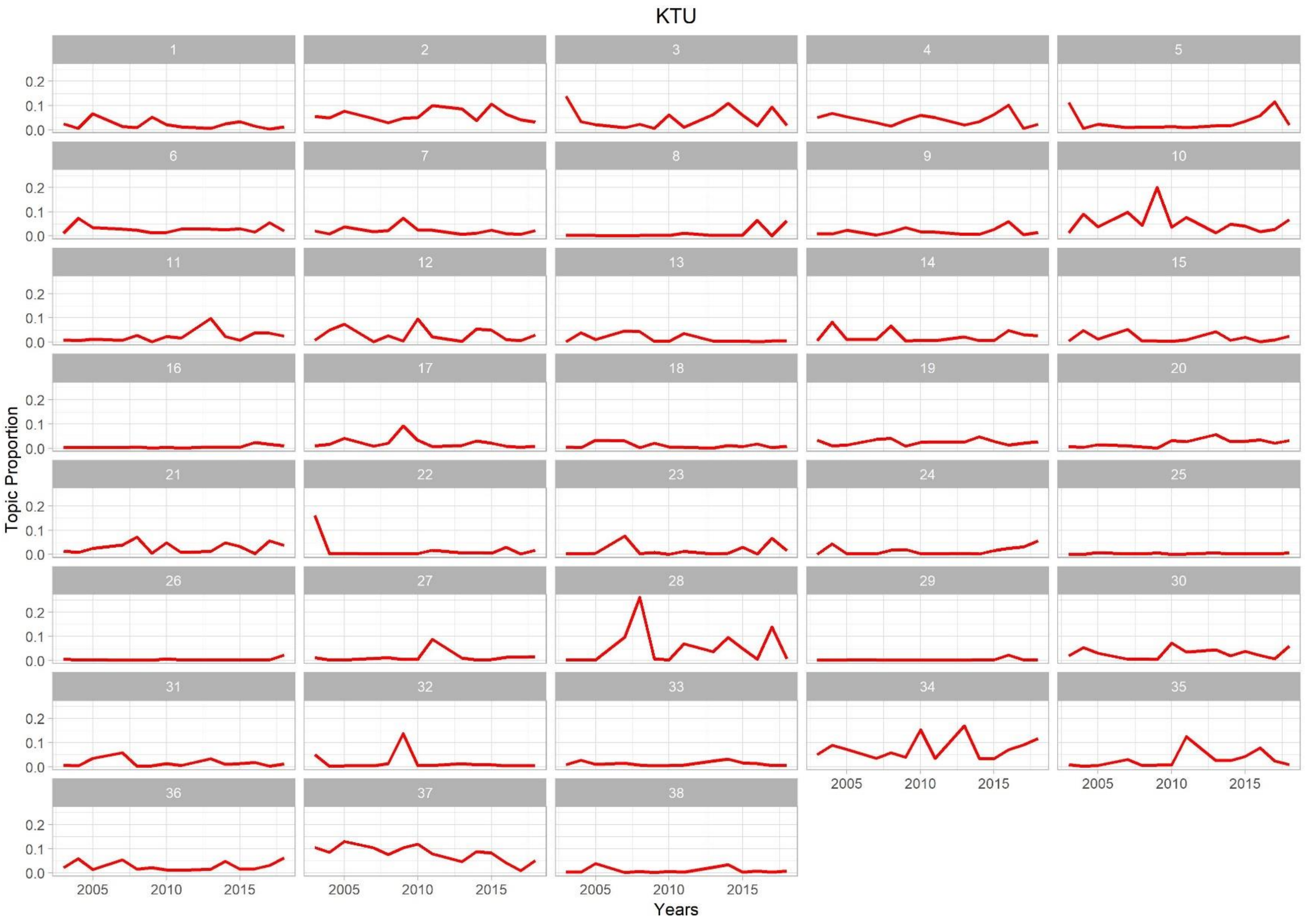
KALYONCU

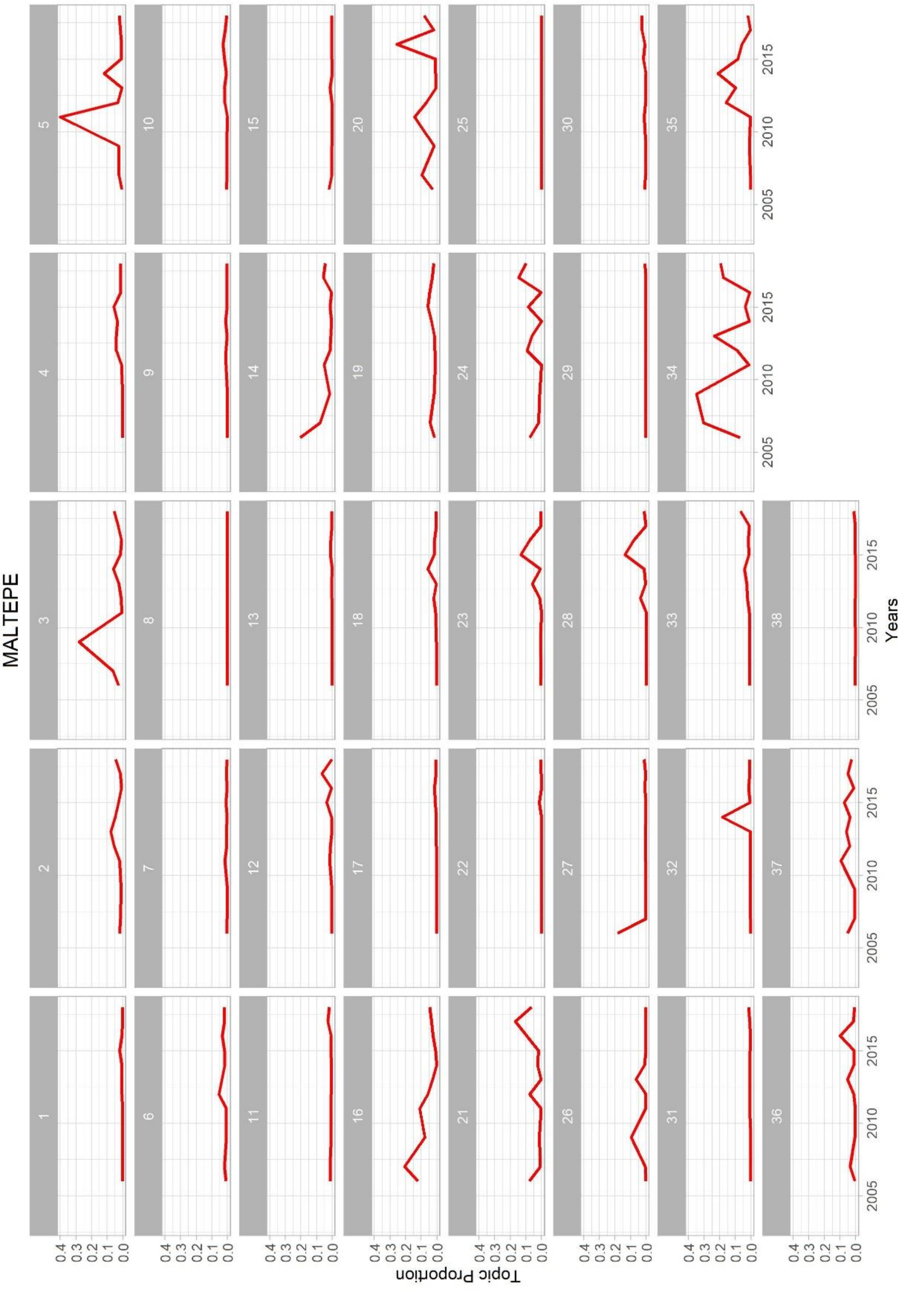

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
Topic Proportion
0.20
0.15
0.10
0.05
0.00
2005
2010
2015
Years

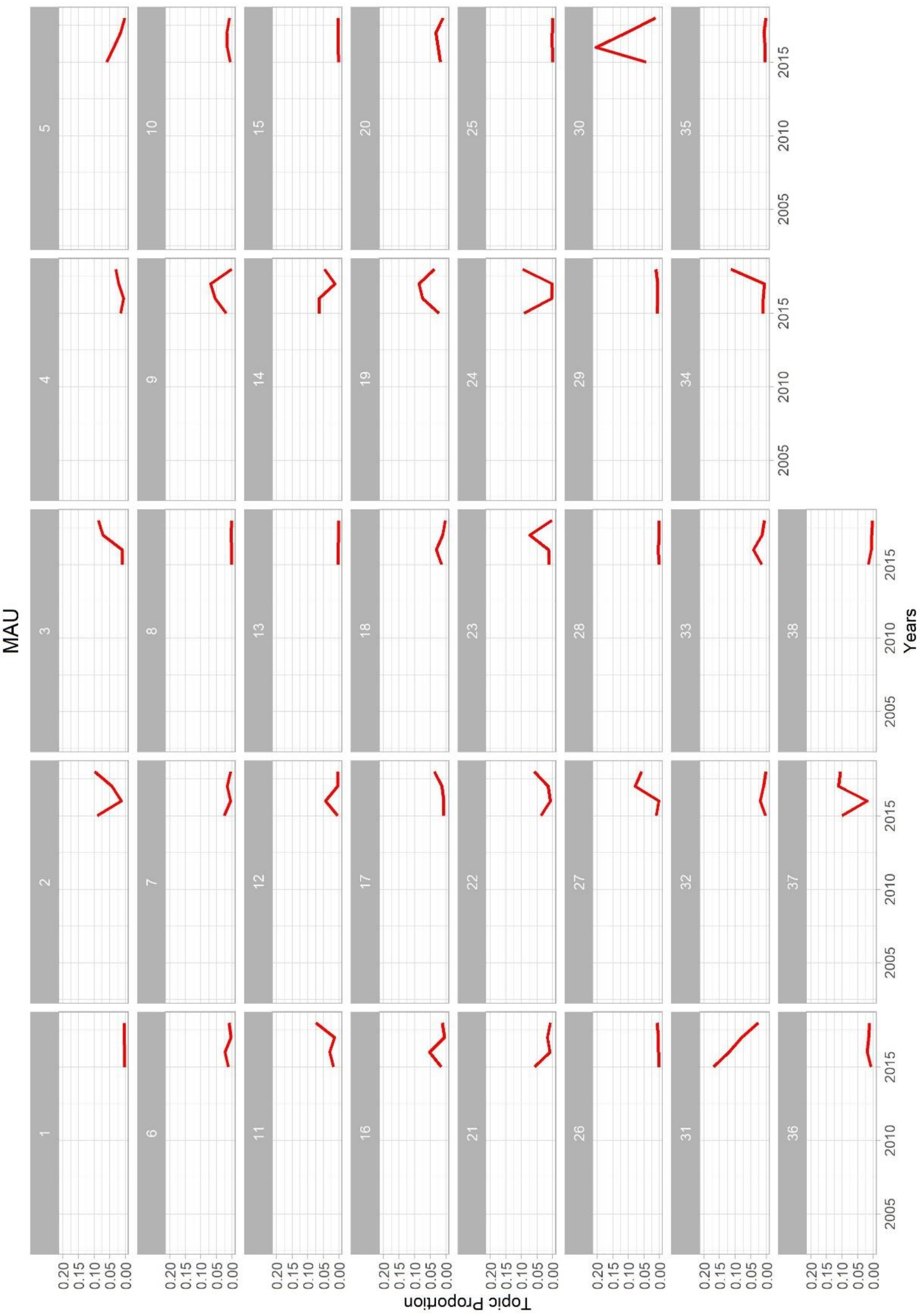

KARABUK
Topic Proportion
Years
0.0
0.1
0.2
0.3
2005
2010
2015
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38

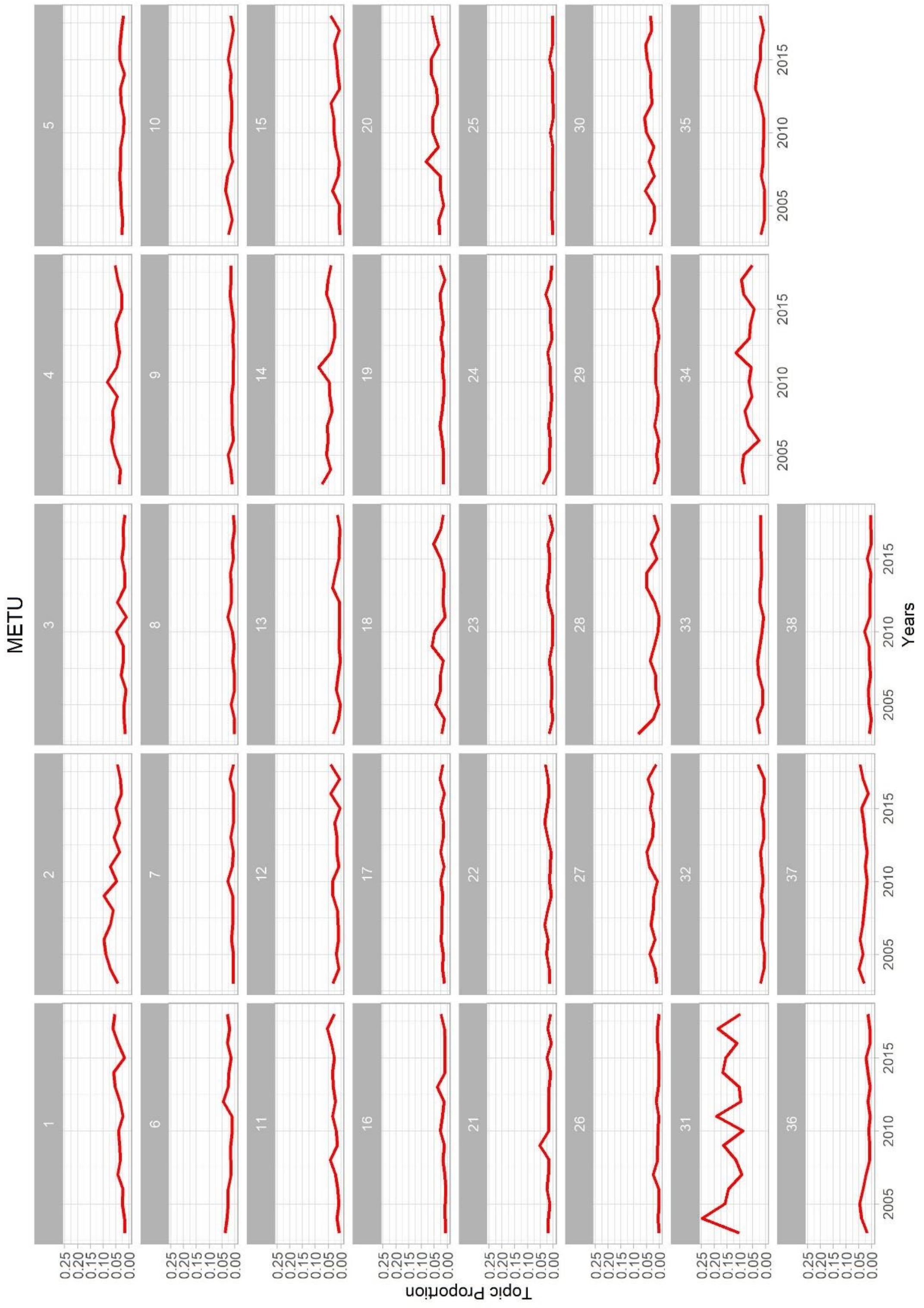

KOCAELI
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
Topic Proportion
0.0
0.1
0.2
0.3
Years
2005
2010
2015

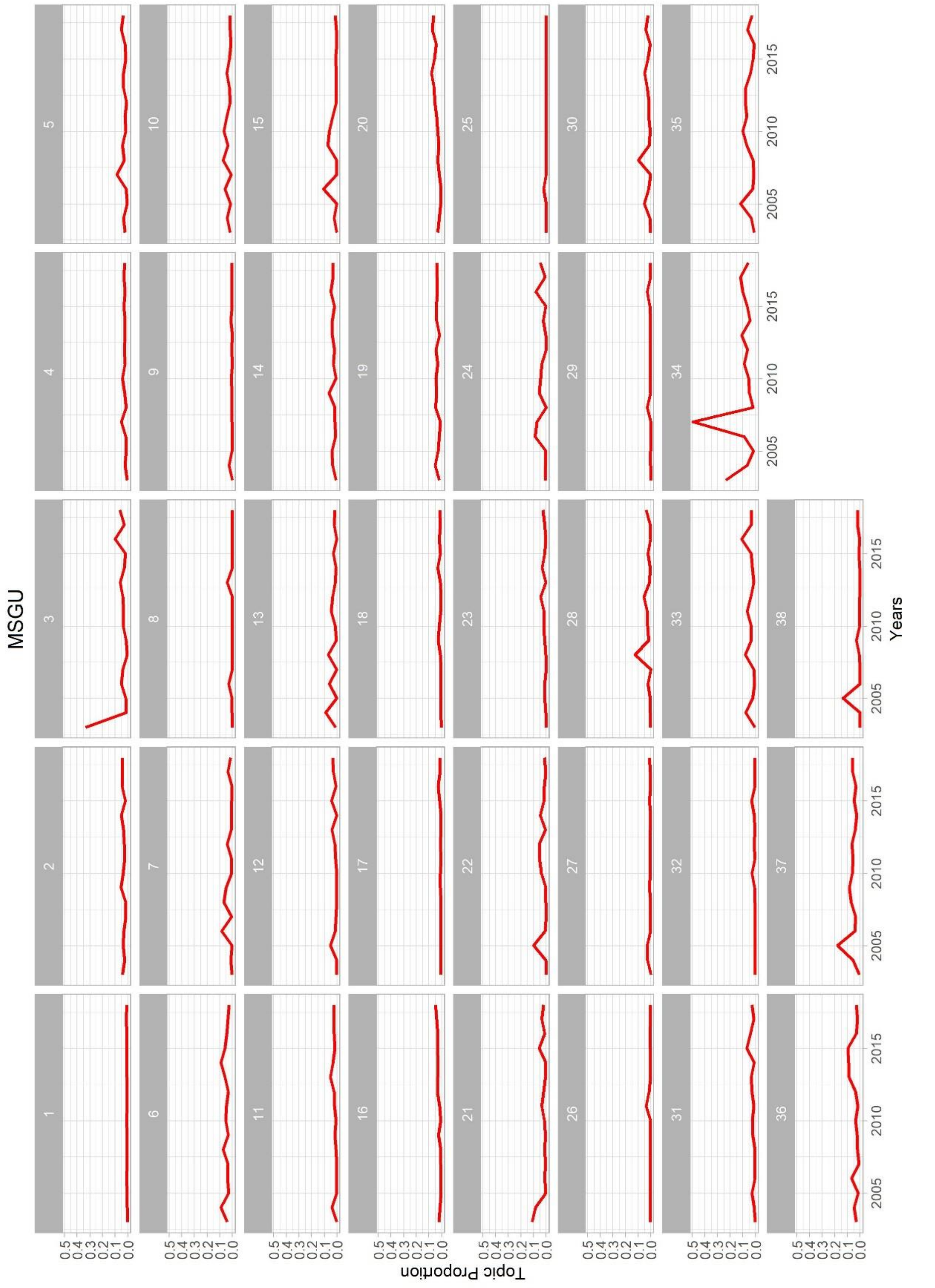

KTO
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
0.0
0.1
0.2
0.3
2005
2010
2015
Topic Proportion
Years

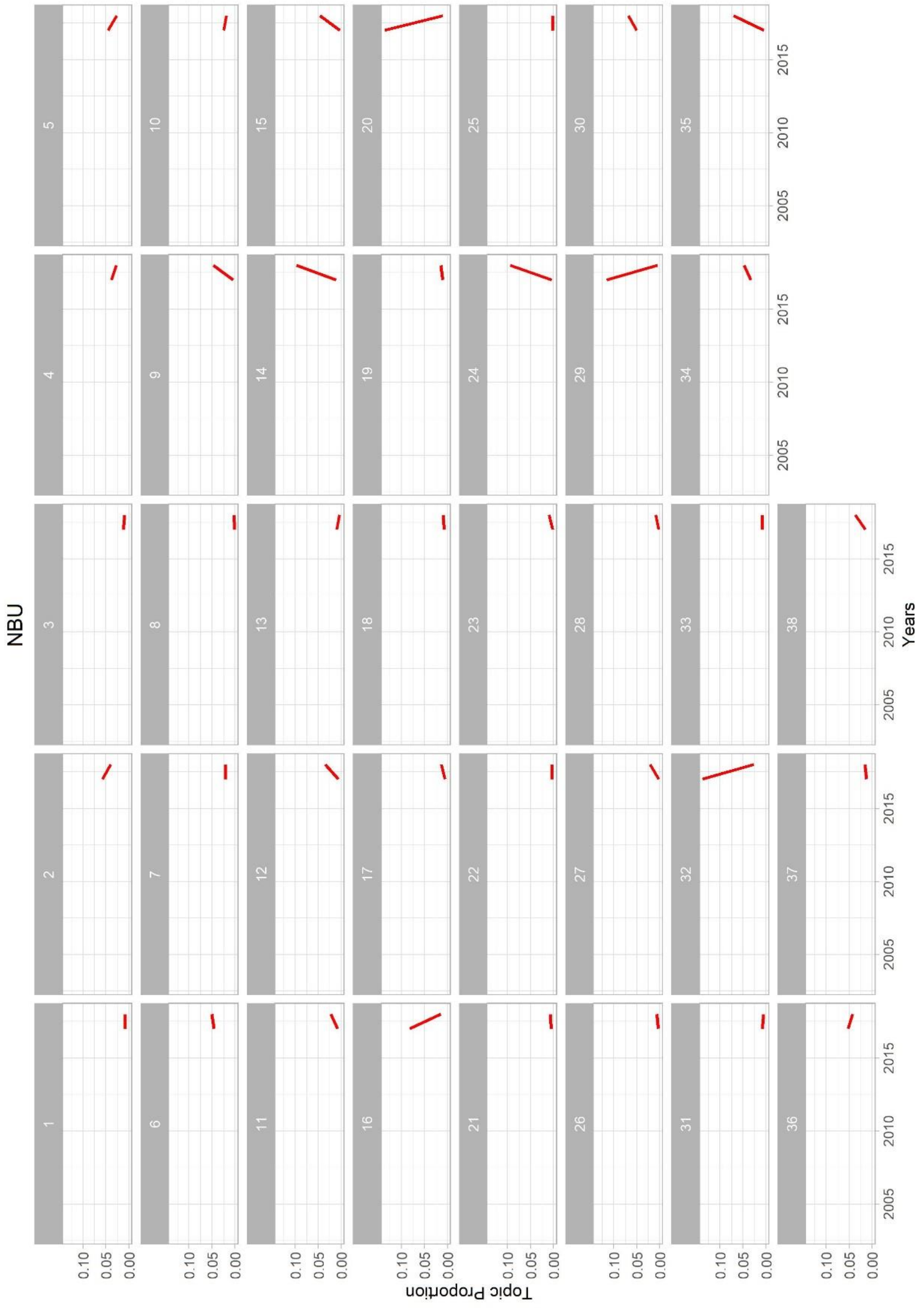

KTU
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
Topic Proportion
0.0
0.1
0.2
2005
2010
2015
Years

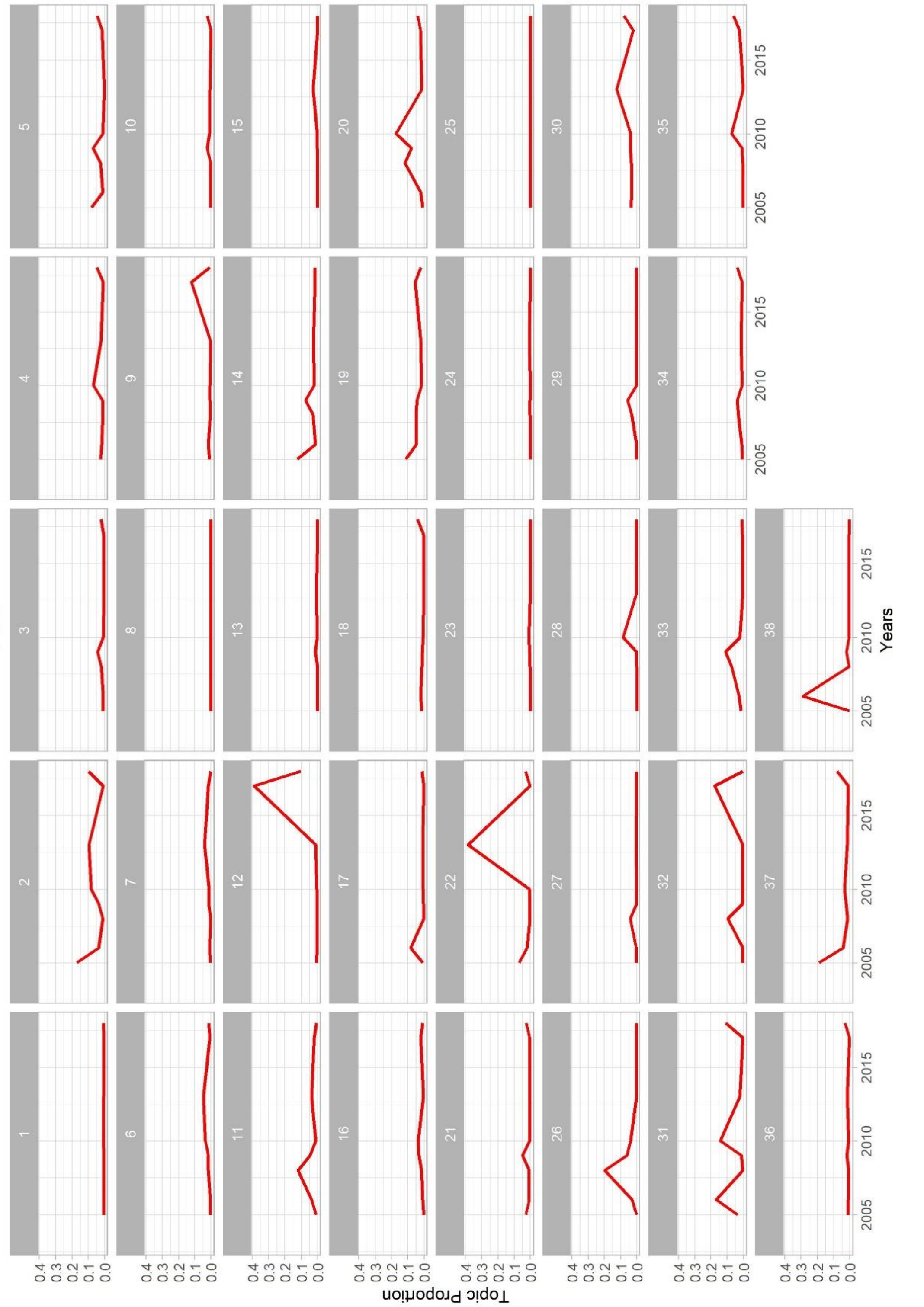

MALTEPE
Topic Proportion
Years
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
0.0
0.1
0.2
0.3
0.4
2005
2010
2015

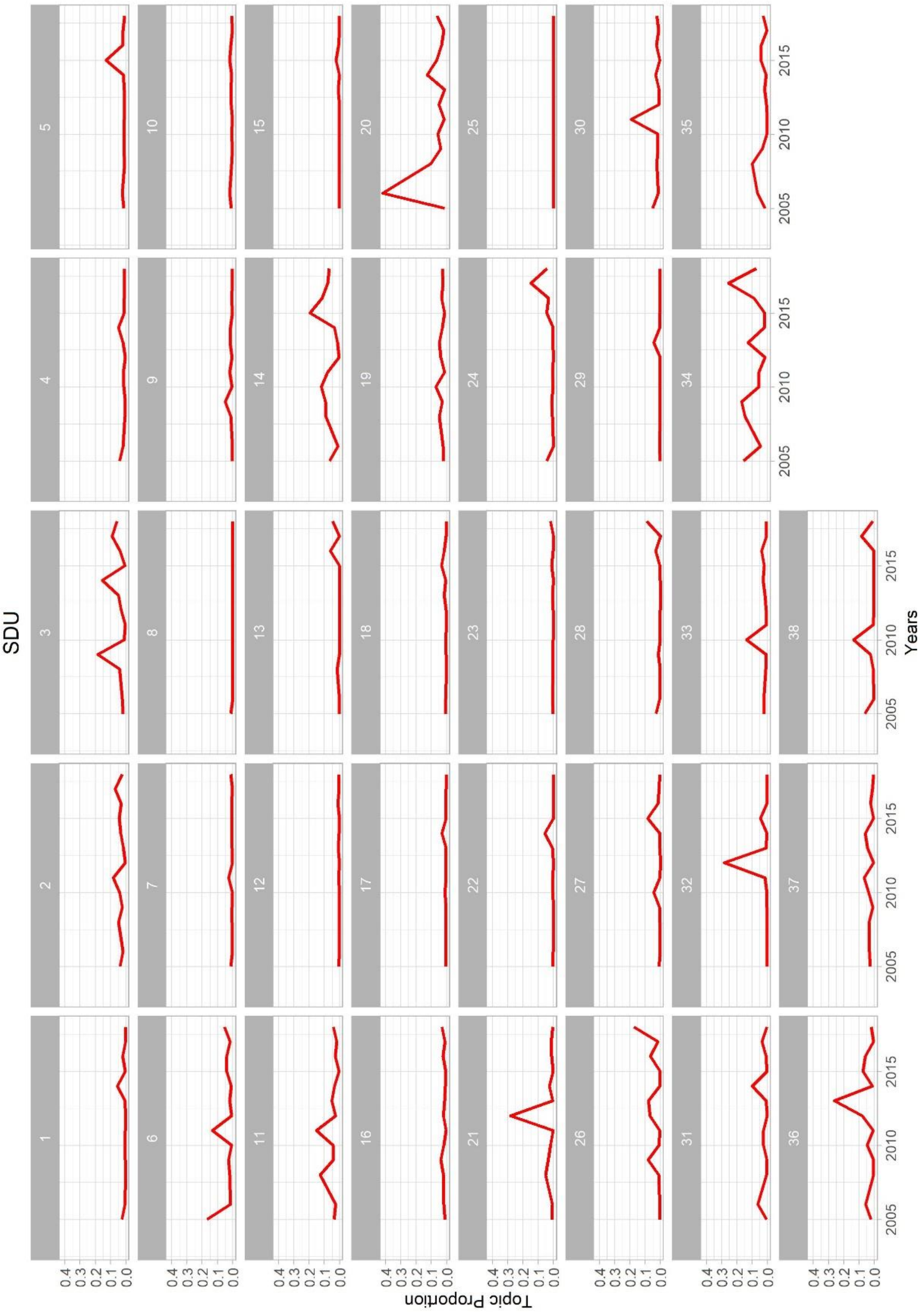

MAU
Topic Proportion
Years
0.00
0.05
0.10
0.15
0.20
2005
2010
2015
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38

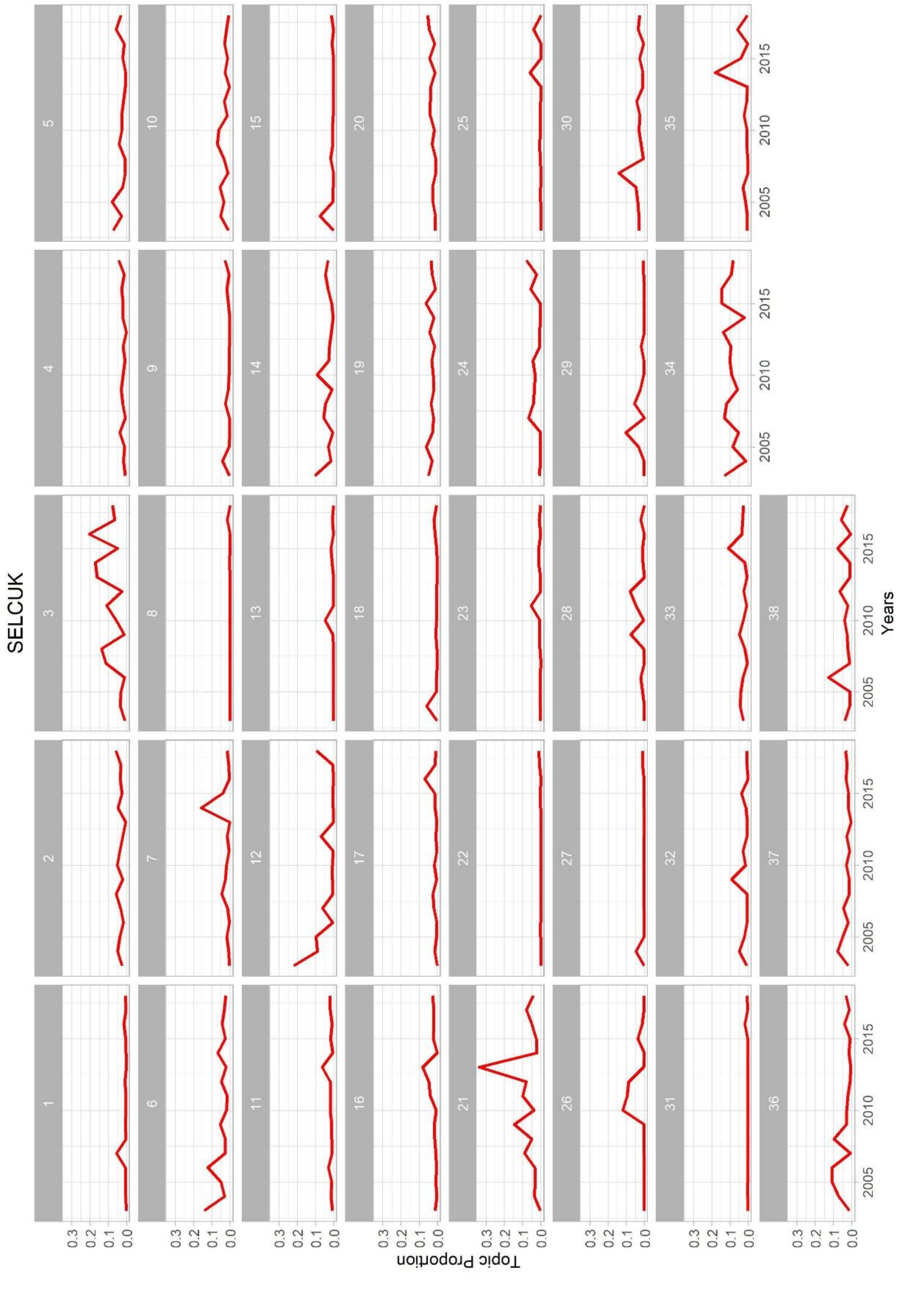

METU
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
0.25
0.20
0.15
0.10
0.05
0.00
Topic Proportion
2005
2010
2015
Years

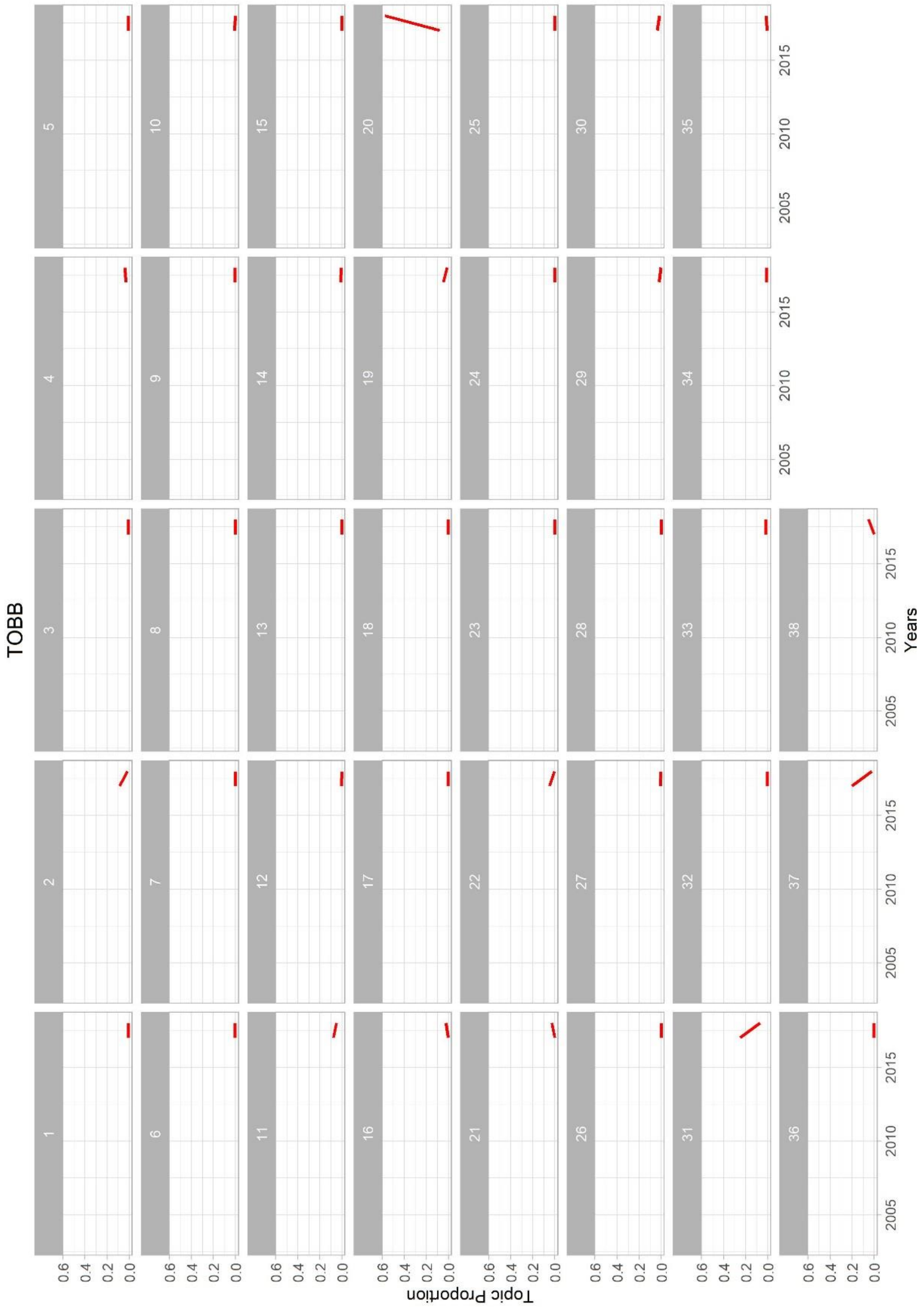

MSGU
Topic Proportion
Years
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
0.0
0.1
0.2
0.3
0.4
0.5
2005
2010
2015

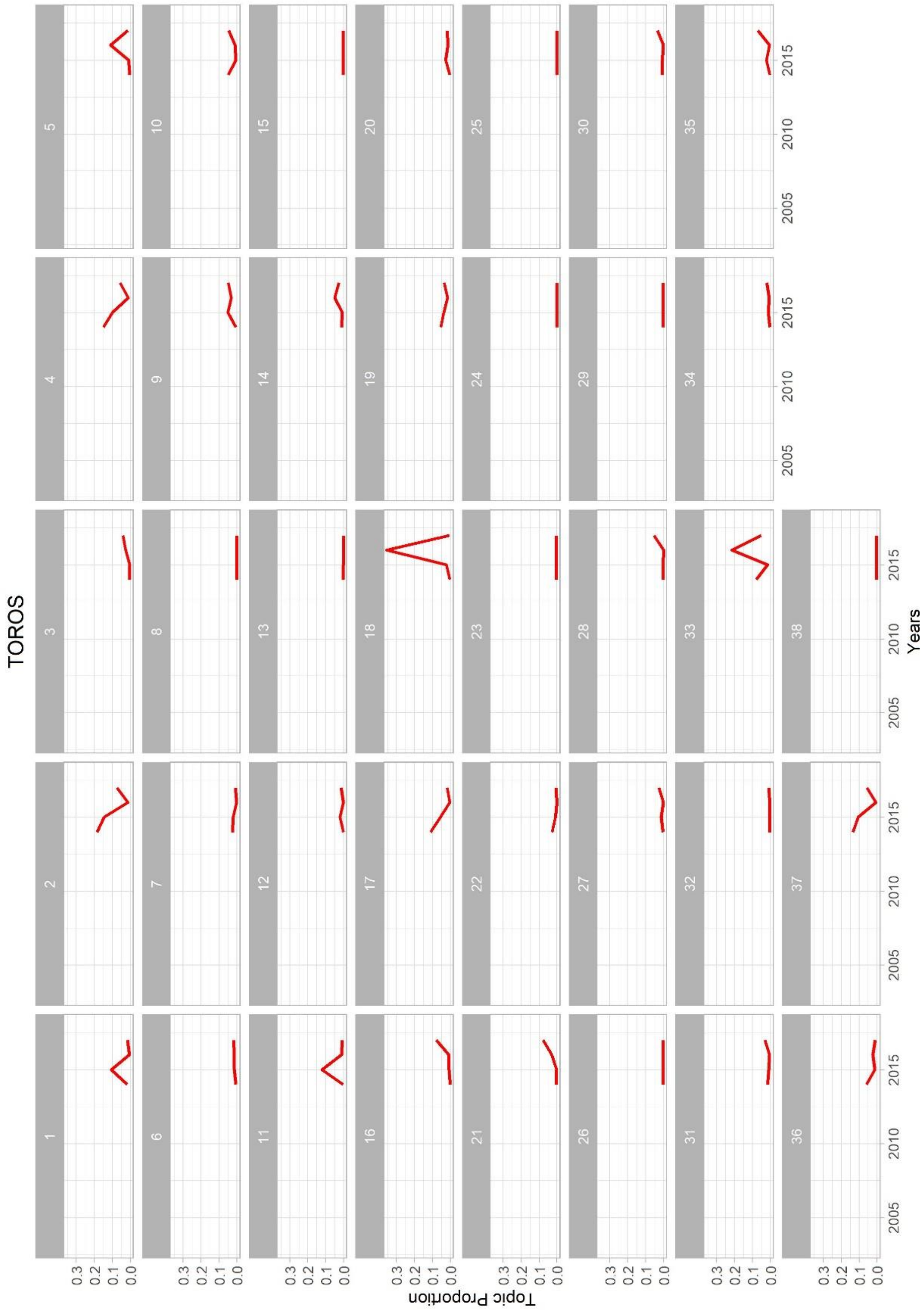

NBU
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
0.10
0.05
0.00
2005
2010
2015
Topic Proportion
Years

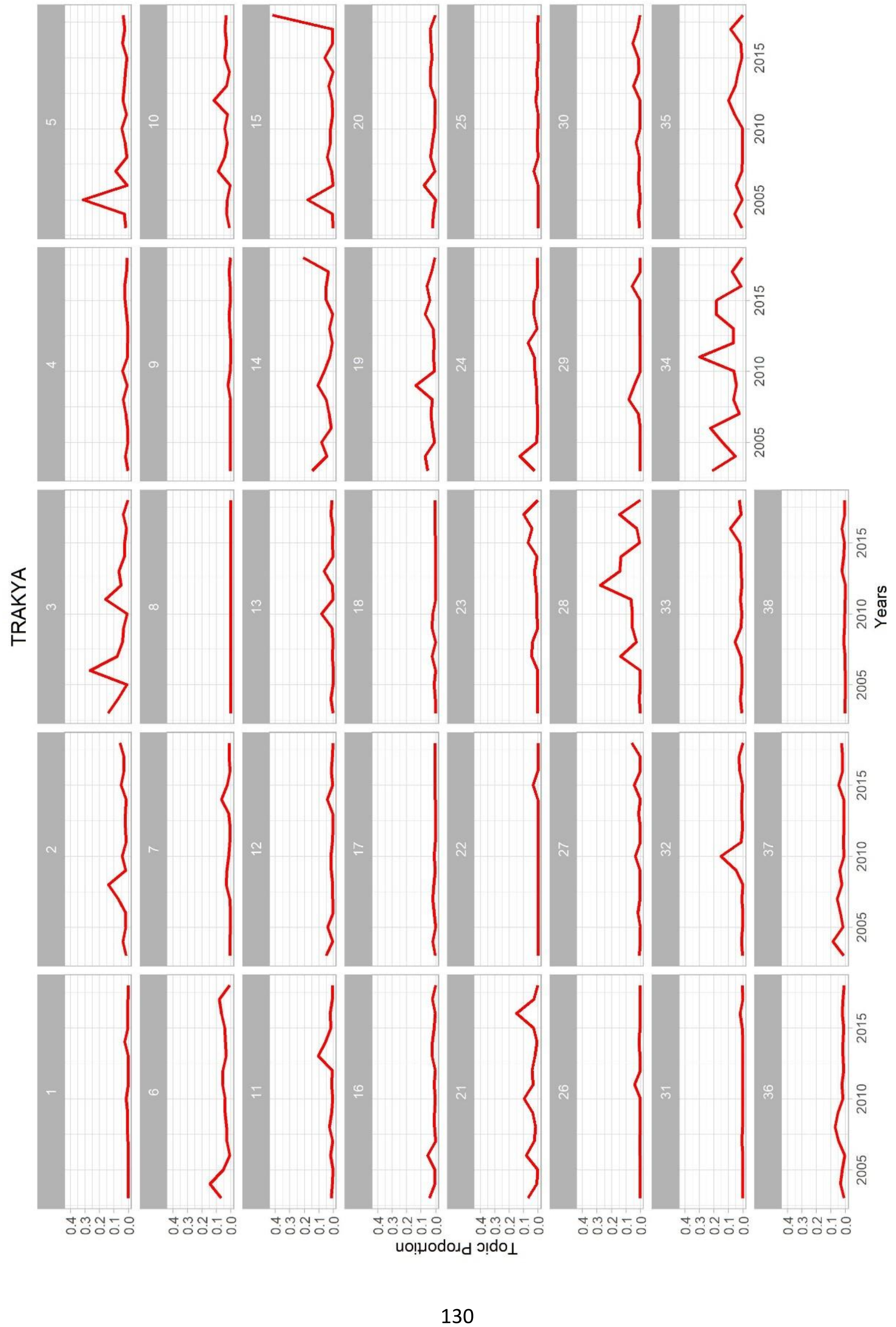

OGU
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
Topic Proportion
0.0
0.1
0.2
0.3
0.4
Years
2005
2010
2015

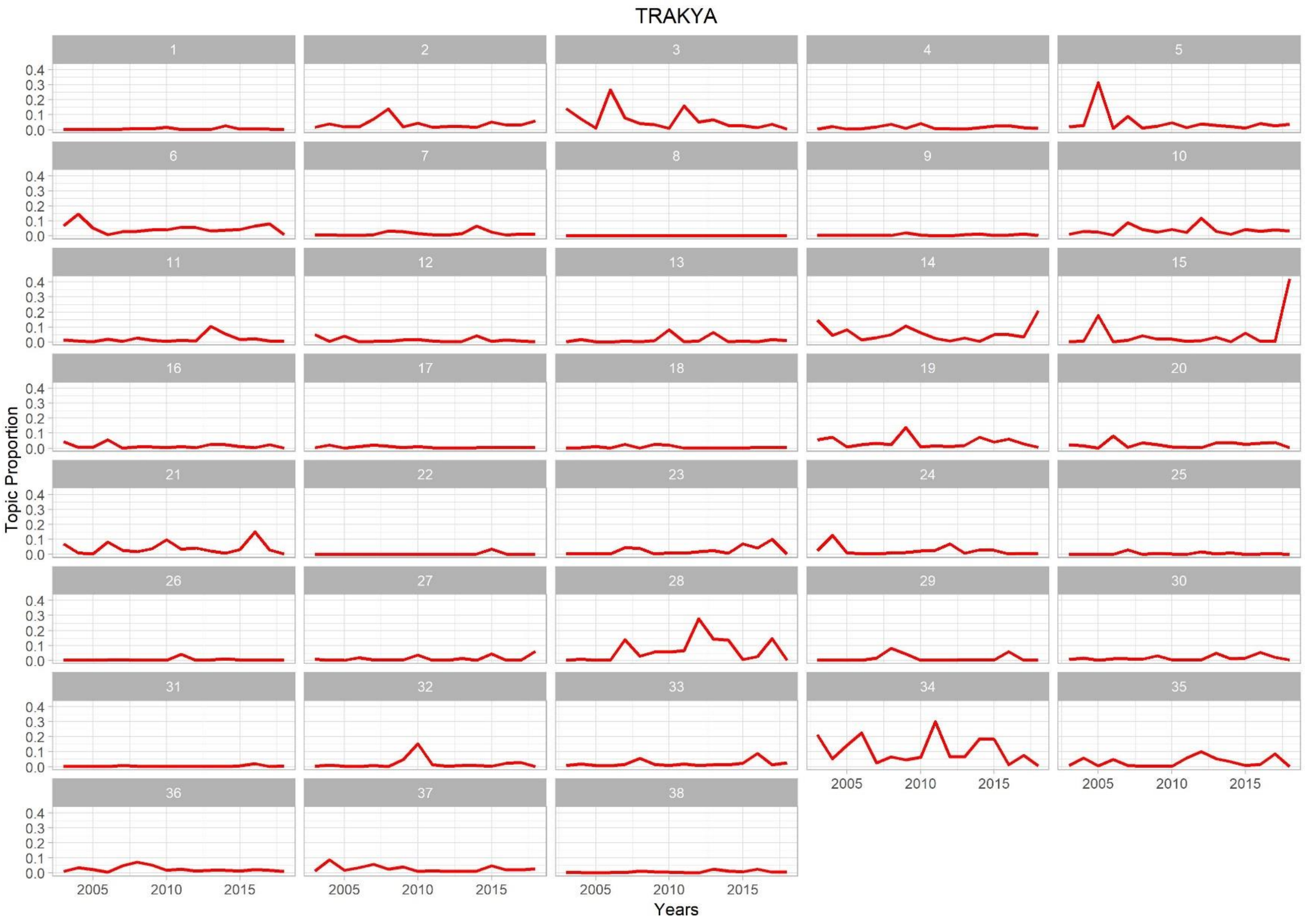

SDU
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
0.0
0.1
0.2
0.3
0.4
Topic Proportion
2005
2010
2015
Years

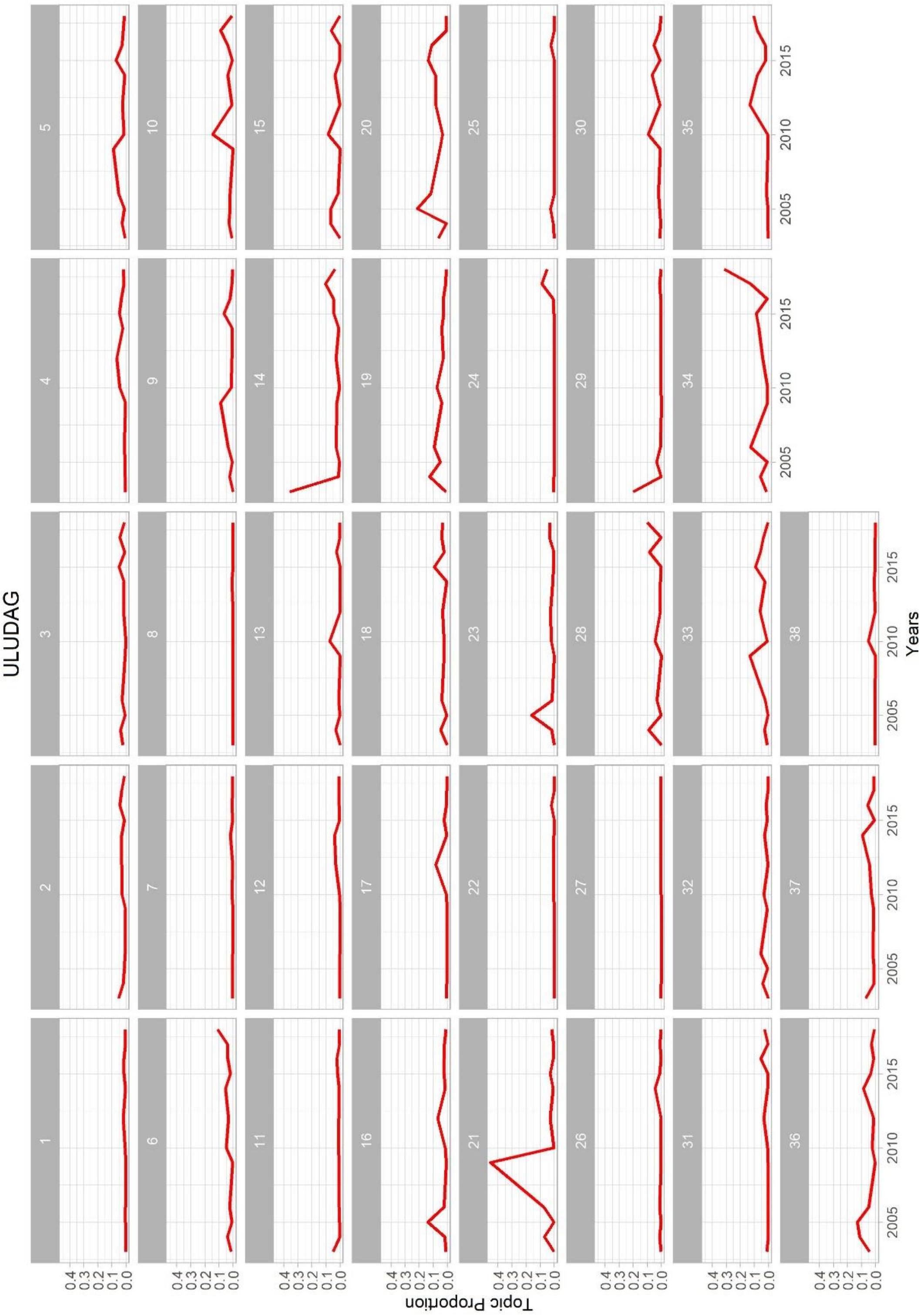

SELCUK
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
0.0
0.1
0.2
0.3
Topic Proportion
2005
2010
2015
Years

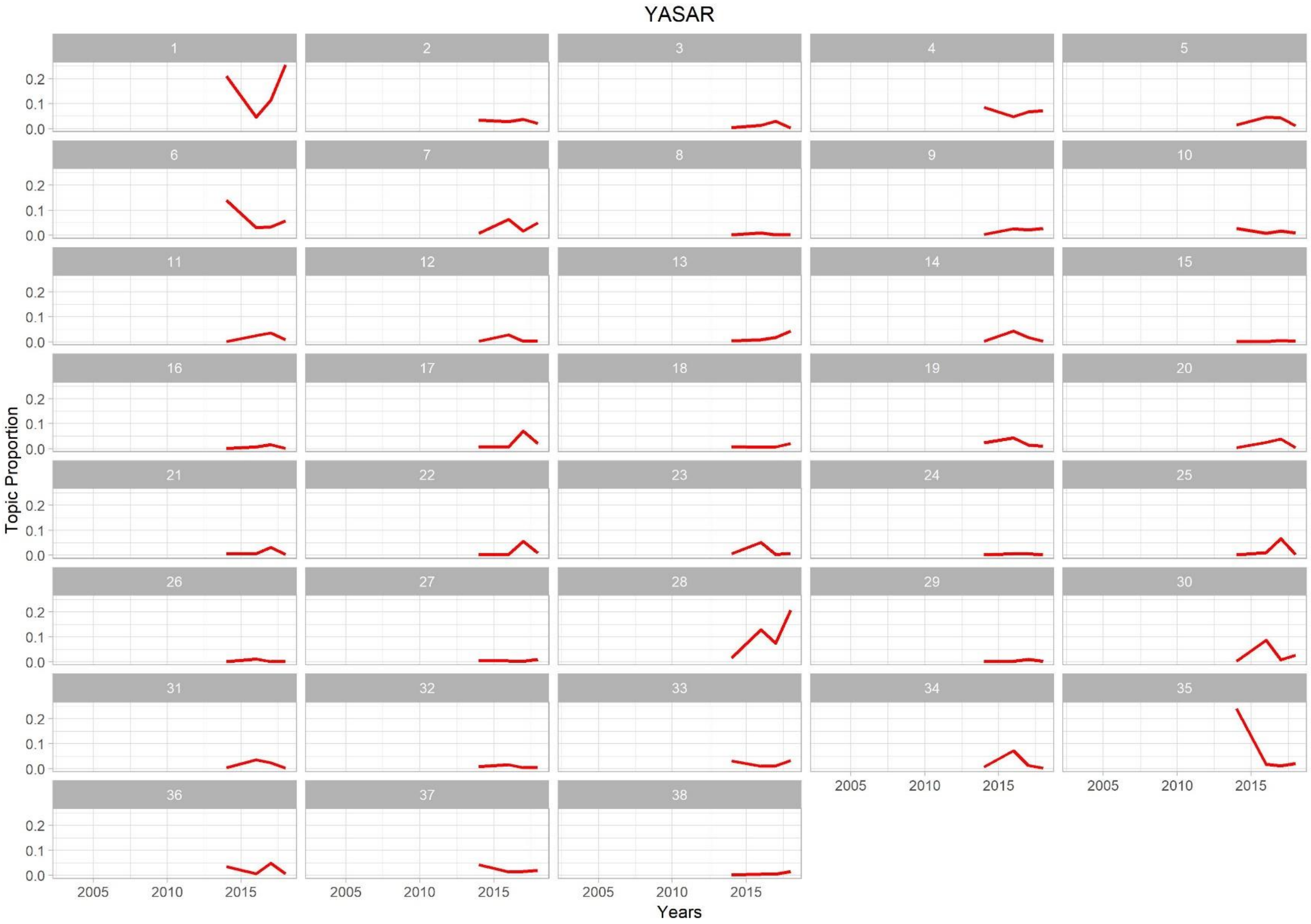

TOBB
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
Topic Proportion
0.0
0.2
0.4
0.6
Years
2005
2010
2015

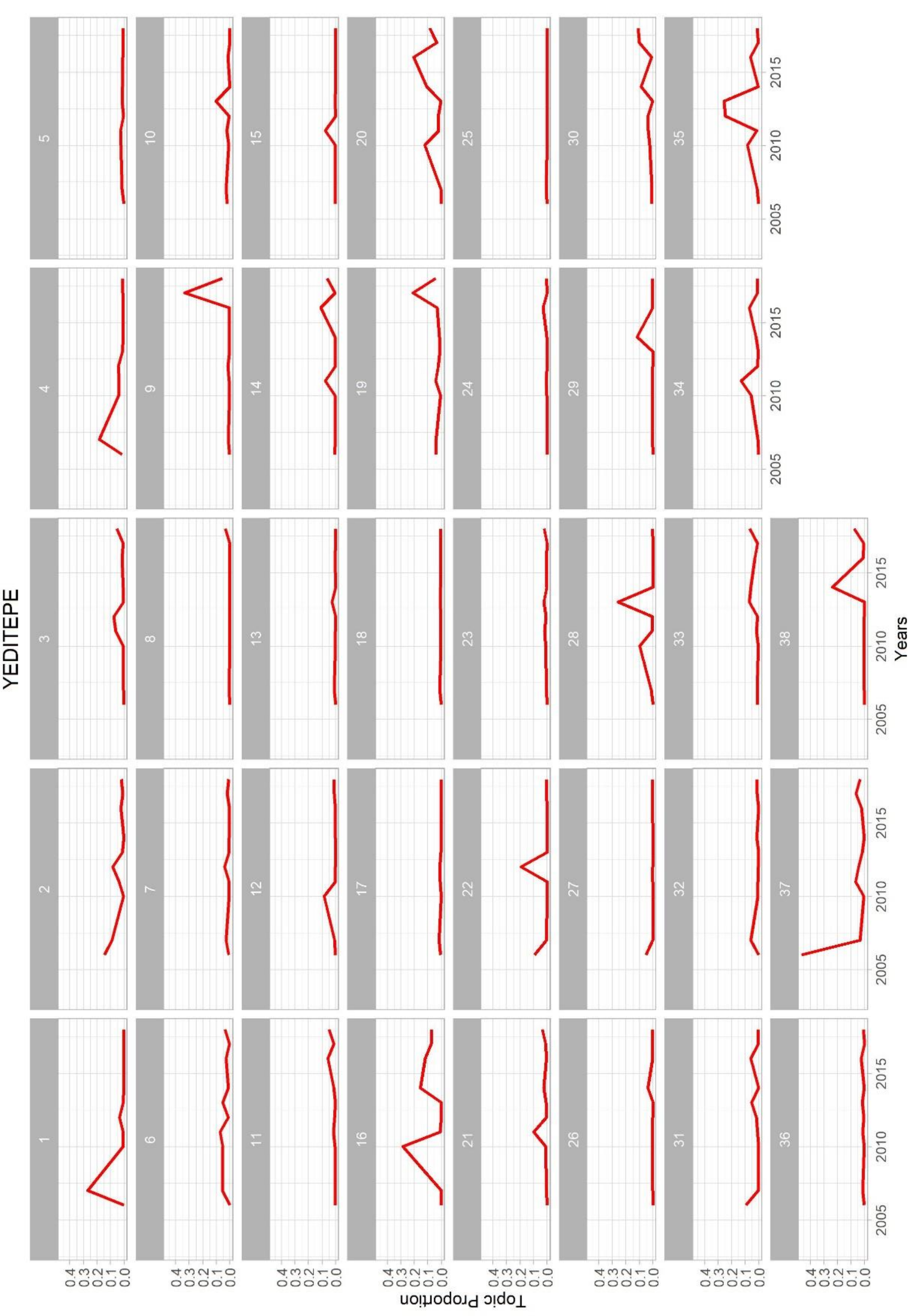

TOROS
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
Topic Proportion
0.0
0.1
0.2
0.3
Years
2005
2010
2015

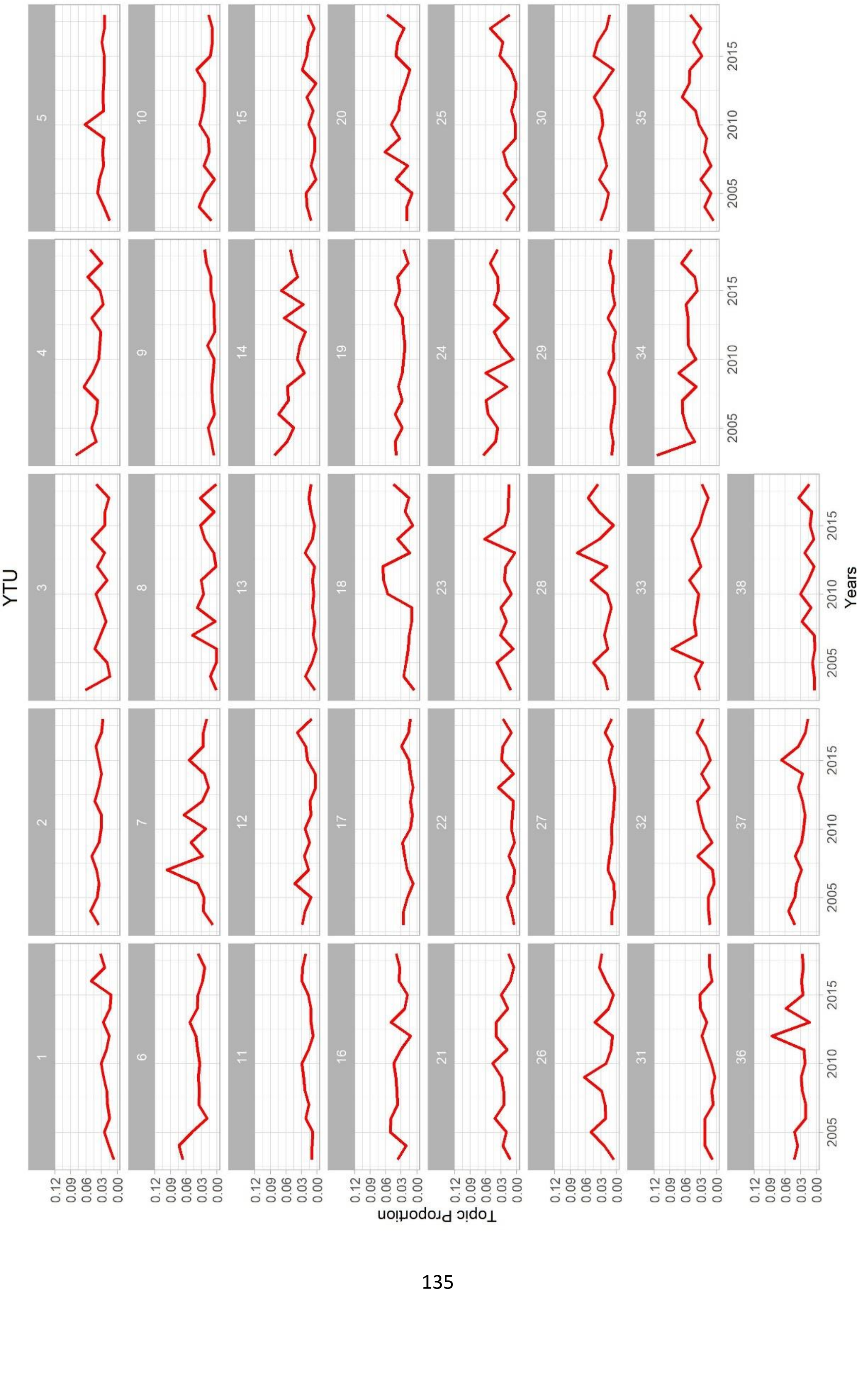

TRAKYA
Topic Proportion
Years
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
0.4
0.3
0.2
0.1
0.0
2005
2010
2015

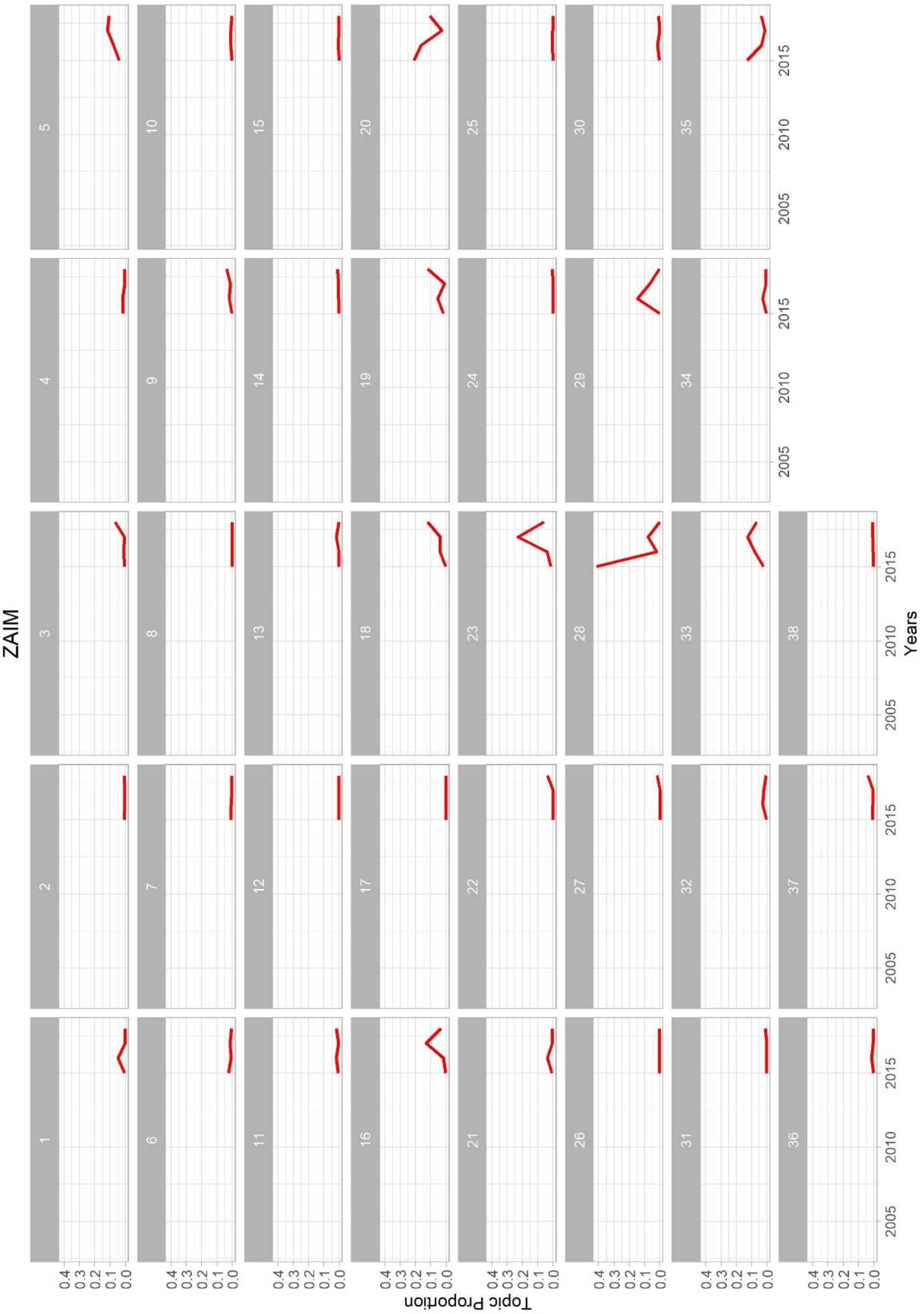

TRAKYA
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
0.4
0.3
0.2
0.1
0.0
Topic Proportion
2005
2010
2015
Years


ULUDAG
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
Topic Proportion
0.4
0.3
0.2
0.1
0.0
Years
2005
2010
2015


YASAR
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
0.2
0.1
0.0
Topic Proportion
2005
2010
2015
Years


YEDITEPE
Topic Proportion
Years
0.0
0.1
0.2
0.3
0.4
2005
2010
2015
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38


YTU
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
Topic Proportion
0.00
0.03
0.06
0.09
0.12
Years
2005
2010
2015


ZAIM
Topic Proportion
Years
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
0.0
0.1
0.2
0.3
0.4
2005
2010
2015