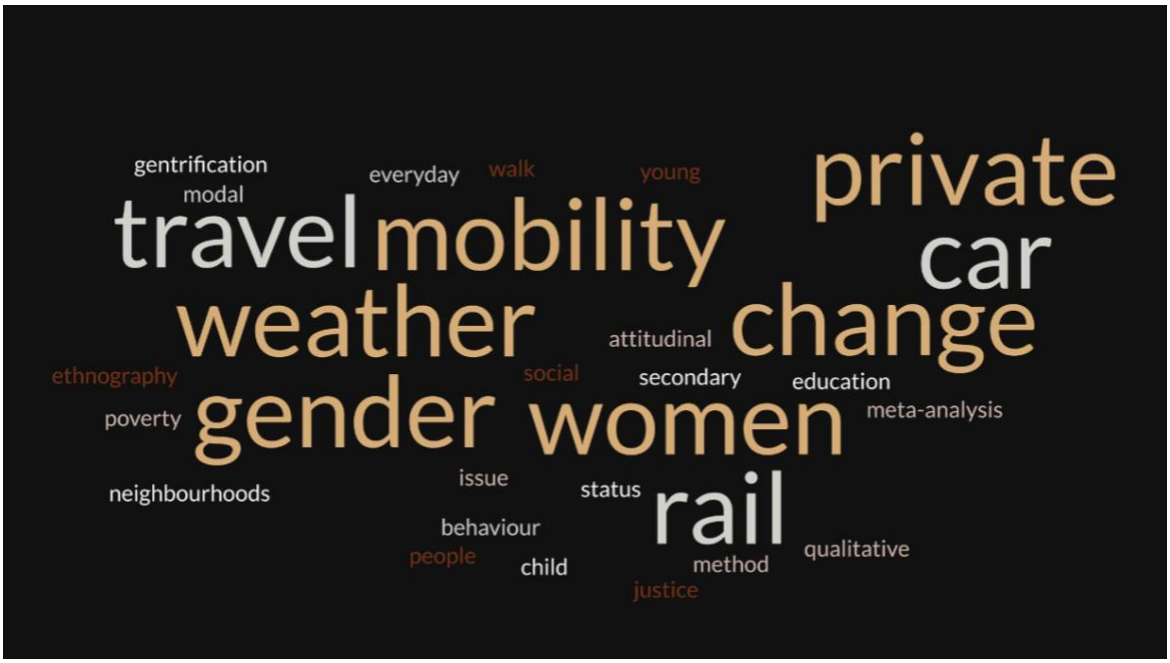
Exploratory Analysis and Discussion on Female-led Research Topics and Contribution to Transportation Geography Utilizing BERTopic and Logistic Regression

Ann Zian Zhang & Sofia Fasullo

May 2024



Key topics associated with female-led research articles

Abstract

This capstone project aims to reveal unique contribution of female-identifying researcher-led research in the field of transport geography. Three approaches were taken for topic analysis on female vs. non-female led research articles - 1) keyword frequency counts, 2) BERTopic, and 3) logistic regression. The results show that female-led research help explore agencies' subjective experience of transit, and gender and social issues, in comparison to the objectivity and physicality of transportation that non-female research tend to focus on.

Introduction: Women in Transportation Geography

Professor Elizabeth Delmelle at the University of Pennsylvania was appointed the co-chief editor of the ***Journal of Transport Geography***, an influential academic journal in the field. She became the first female-identifying scholar in this leadership role tracing back to the establishment of this journal back in 1993. Personally, as female-identifying student researchers, we find this moment particularly celebratory and meaningful, and we hope to explore the long way we, female researchers, have come through. Through examining topics researched by female scholars and published on the journal, we hope to highlight unique contribution, perspectives, and insights female researchers are bringing into this field, and at the same time advocate for more equity and inclusivity for female scholars in this field.

Some previous studies have taken similar paths of utilizing machine learning / language models to examine differences in topics by researchers of different gender identities in other fields. One recent example is a paper by J. Conde-Ruiz et al. in 2021, in which they looked at gender distribution across topics in the top five economics journal utilizing STM (structure topic models).[1] Research as such are enlightening in two aspects of their methodological thinking – 1) the specific language models used for identifying topics, and 2) the ways in which the researchers used to identify authors' gender identity. In this case, Conde-Ruiz et al. chose unsurprised STM, which was considered new and advance at the time when the paper was published. And they relied on three databases for identifying gender from authors' first names – U.S. Social Security Administration (SSN database), and two databases developed by other researchers utilizing social media. Undoubtedly, STM is now no longer considered an advanced model with development of new tools, and the three databases the researchers used in their study has a strong Angelo-centric nature with better prediction on English names' gender but fail to account for international names.

Though with similar fundamental theme of looking for gender-related topics in a chosen academic field, this project aims to offer a fresher view by applying BERT model instead of older language models for topic analysis, by looking at the field of transport geography, a field that has been overseen in previous research, and by using more reliable methods for identifying authors from diverse background.

[1] J. Ignacio Conde- Ruiz, Juan-Jose Ganuza, Manu Garcia, and Luis A. Puch, "Gender distribution across topics in the top five economics journals: a machine learning approach," *Journal of the Spanish Economics Association*, no.13 (Nov 2022): 269 – 308.

Journal of Transport Geography and Women in Transport Geography

Since the journal's establishment in 1995, we are witnessing an increasing trend of both number of overall articles and number of female-led articles. (Fig. 1) The percentage of female-led articles also have been increasing over the years. Interesting, however, there is a huge drop in 2021 and 2022, potentially due to COVID and societal expectations on female taking more family responsibilities at those critical healthcare moments. Geographically speaking, US, UK, and China are among top affiliation countries where female researchers are from, while New Zealand, Sweden, and Denmark has a higher percentage. Note that affiliation does not equal to nationality, but potentially linked more to sponsorship and funding sources – suggesting academic institutions in those countries or regions are sponsoring more female scholars to lead research projects.

Among all articles, female-led research articles have an average page count of 9.22 pages (comparing to 9.13 in non-female), an average time cited at 40.99 (comparing to 39.59 in non-female), and an average number of authors at 2.89 (comparing to 2.65 in non-female). Overall, the stats demonstrate an increasing trend in female researchers' involvement and a high quality of their work.

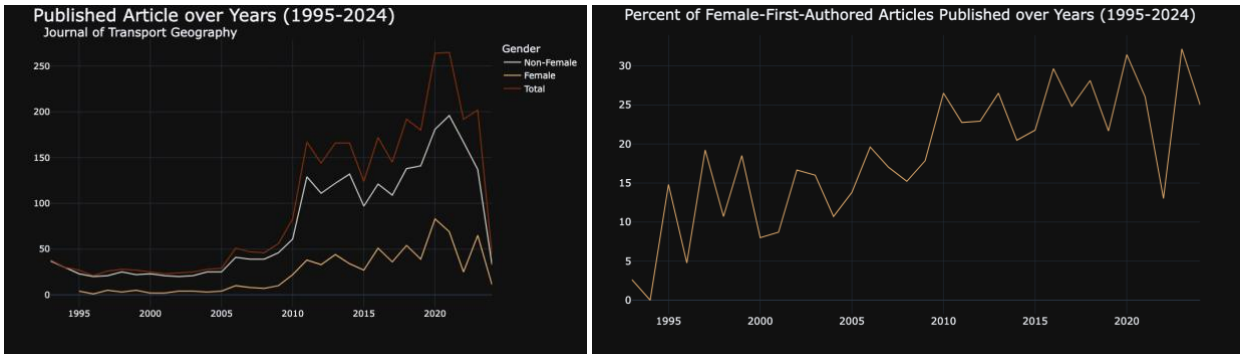


Fig. 1 – Published article over years (1993 – 2024) (Left) and
Percent Female-first authored articles published over years (1993 – 2024) (Right)

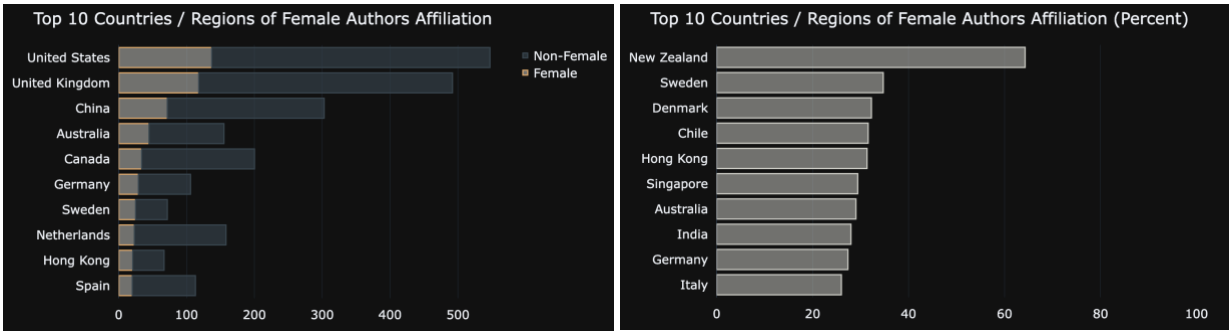


Fig. 2 – Top 10 Countries / Regions - most female-first author articles (left) and
Top 10 Countries / Regions – highest percentage of female-led research (right)

Research Question and Objectives

The key research questions that this project aims to explores are – 1) how female-identifying researchers have been contributing to the field of Transport Geography, and 2) what knowledge gaps there would be without contribution of female researchers. Through this exploratory analysis on unique topics covered more frequently by female-led, we hope to reveal their irreplaceable unique contribution to the field, and advocate for further equity and more research funding and research opportunities in the future.

Methodology: Data Source

Two types of data that are critical to our research are **– 1) data that potentially reveals topics of articles**, and **2) gender of authors of each article.**

We utilized ***Scopus*** for the topic-related data, which is an abstract and citation database under Elsevier, an academic publisher through which *Journal of Transport Geography* can be accessed electronically. The database includes information about individual articles from 1993 till early 2024, including all authors' names and affiliations, page counts, citations[2], index words, author keywords[3], and most importantly, abstracts. Index, author keywords, and abstract were used in the actual topic extrapolation and analysis, while the other information were used for exploratory analysis, shown earlier in the Introduction section. Until February 2024 when we accessed the database, the journal has published a total of 3060 articles (704 female-led and 2356 non-female-led). Among those, 306 did not have abstract information available through the Scopus dataset, hence, they were excluded. Therefore, 653 female-led and 2101 non-female-led research articles were examined in this project.

Regarding access to the second type of data, we decide to limit the scope to first authors and manually searched up online each author's preferred pronouns. Given the diverse background of researchers, we believe this is a better way, despite more time consuming, in comparison to similar precedent research that uses lists of common female English names to do quick filter through. For future studies with larger scale, nonetheless, better data entries (i.e. incorporating self-reporting gender when receiving and publishing articles) or more automated identification mechanism is necessary, given the limited capacity of manual search.

[2] In this circumstance, citations show how many times each article has been cited by other articles.
[3] Similar to index words, but less officially used as tag but more for the purpose of author self-reporting keywords they recognize as relevant.

Methodology: Analytical Lens, BERTopics, and Regression

The project takes three approaches / analytical lens to explore gender-related topics published on *Journal of Transport Geography,* each of which will be discussed in detail.

- **Part 1 – Keyword Counts**

The whole dataset is split into female and non-female led groups, containing information about articles, index words, and author keywords. For the first part, abstract paragraphs are split into a long list of individual words, stopper words (e.g., the, a, an) are removed from this list, then frequencies of occurrence of individual words are summarized. Intra-group differences in the top-hit words are compared for generating insights. This is the most straightforward approach to look at the most-frequently occurring keywords in abstract.

- **Part 2 – BERT Topic Analysis**

The second approach is to utilize BERT language model to identify topics across 1) female-led research articles and 2) all articles. BERT model mainly consists of 6 layers, including: 1. Document Embeddings (SBERT), 2. Reducing dimensionality (UMAP), 3. Clustering (HBDSCAN) for reducing embeddings, 4. Tokenizing documents (Count-Vectorizer), 5. Word-weight Scheme (c-TF-IDF), and 6. Tune Topic Representation (Key-BERT). Each topic (output) consists of a group of 8-10 keywords, and they are ranked in order of topics with which the most individual articles are associated with.

- **Part 3 – Logistic Regression**

Logistic Regression is particularly helpful in revealing intra-group differences between the two gender groups. Logistic regressions are run for predicting dependent variable of gender (female / non-female 1/0 dummy) and three sets of independent variables – topics identified using BERT in the previous section (47 variables or 337 if exploding individual words wrapped in a topic), index words (2573 variables), and author keywords (6878 variables). Magnitude of coefficient of individual variables are compared to unveil the top keywords that are most associated with the gender groups.

Results and Discussion

1. Keyword Frequency

The result of abstract keyword frequency count is shown below in table 1. There is a huge overlap between the two groups regarding the most frequently occurring keywords, such as urban, accessibility, and spatial. The more unique ones such as mobility, different, and research, are not too useful for distinguishing content that are unique to female-led articles. Since counting frequency is the most straightforward step, it supports the rationale of why we would need more complex analysis in revealing topics than simply counting frequency of words occurring in an abstract.

Female	**Non-Female**
Mobility	Accessibility
Urban	Urban
Accessibility	Spatial
Public	***Model***
Analysis	Analysis
Areas	Public
Spatial	Network
Different	***Time***
Research	Areas
Network	***Transportation***

Table 1. Top 10 Keywords in Abstract

2. BERT Topic Identification

2.1 Topics of Female-led Research Articles

This section is for testing out BERT model on our dataset and if we can generate any useful insights from simply looking at female-led articles' topics. As shown below (Fig. 3), the top topics identified matches common expectations on transportation geography articles, including mobility, car, shipping, bikes, etc. This, however, is not particularly helpful for comparing and revealing topics specifically associated with female.

Fig. 3 – Topics identified using BERT on Female-led Research Articles

2.2 Comparison of Topic Distribution with Topics Identified with All Articles

The second round of BERT was run on all articles. The one version of the BERT output is a list of all articles and the specific topics that have the strongest tie to their content. The intra-group difference is compared by counting the number of topics in each group. While we are witnessing some overlaps in top ranked topics, we start to see trends such as Topic 4 (school, students, children, travel) ranking higher in the female group and then non-female group.

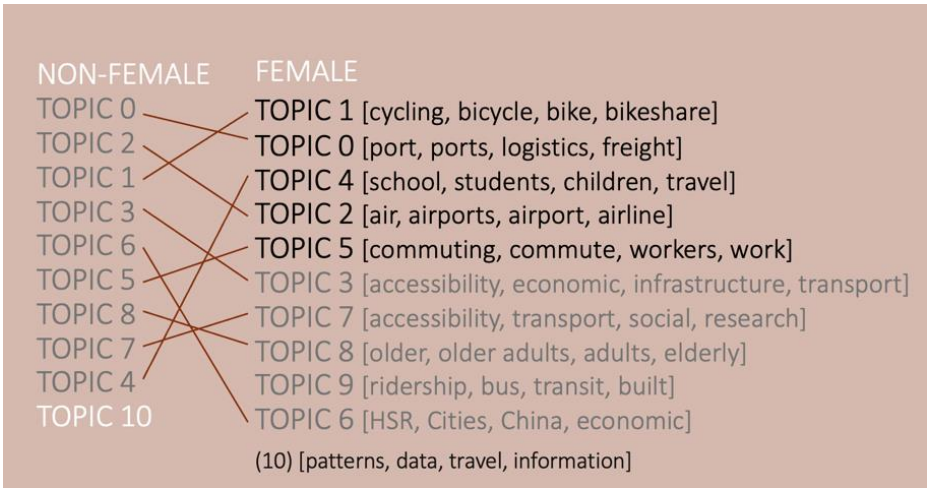


Fig. 4 – Top Ranked Topics for Female vs. Non-Female articles

3. Logistic Regression

To reveal the intra-group differences most clearly, we are conducting logistic regression on BERT topics identified in section 2.2 (a total of 47 topics), index words, and author keywords. The results are shown below in Table 2. It appears to be the trend that many female-led articles look at subjective experience of transit, specific agencies such as children, students, or young people, and utilizing less-commonly used methodology including attitudinal survey, qualitative analysis, and ethnographic studies. Comparing to non-female groups' focus on the physicality of transit and networks, female-led articles are particularly meaning in bringing in human-centric perspectives and qualitative aspects.

	Female	**Non-Female**
BERT Topics	[Practices, Mobility, Transport, Private] ***[Rail, Neighborhood, Neighborhoods, light rail]*** ***[School, Students, children, travel, childrens]*** ***[Weather, ridership, weather conditions]*** ***[Gender, car, women, men, car uses]***	[Railway, Europe, Baltic, States] [Rail, Freight, Closure, Intermodal] [Ride Hailing, Ride Sourcing, Transit, Ride Pooling] [Geography, Research, Transport] [Air, Airports, Airline, Airlines]
Index Words	***Secondary Education*** ***Child*** ***Women Status*** ***Gender Issues*** Meta-Analysis ***Qualitative Analysis*** ***Poverty*** Survey Methods ***Attitudinal Survey*** ***Social Justice***	Waterway Transport Regional Economy Planning Process Model Test Infrastructure Planning Sensitivity Analysis Distribution System Uncertainty Analysis Classification Service Sector
Author Keywords	***Walk*** ***Ethnography*** ***Young People*** Content Analysis ***Quality of Service*** Information and Communication ***Travel Diary*** Motility Mobile Methods ***Inequalities***	Governance Light Rail Location Airlines Air Transport High Speed Rail Networks Globalization Developing Countries Regulation

Table 2. Top 10 Keywords that are most associated with gender groups resulting from logistic regression on BERT topics, index words, and author keywords

Conclusion

In conclusion, female-led research articles are bringing in human-centric perspectives, qualitative research methods, and attention to subjective experiences of agencies traveling into the field of transportation geography, which would otherwise be missing, as non-female researchers tend to focus more on the physical infrastructure and larger network systems as opposed to individual's experience. If non-female articles aim to study for more efficient transportation, then female articles are hoping to build a more user/rider-friendly, more accessible, more equitable transportation system, both of which are irreplaceable and critical parts in the field.

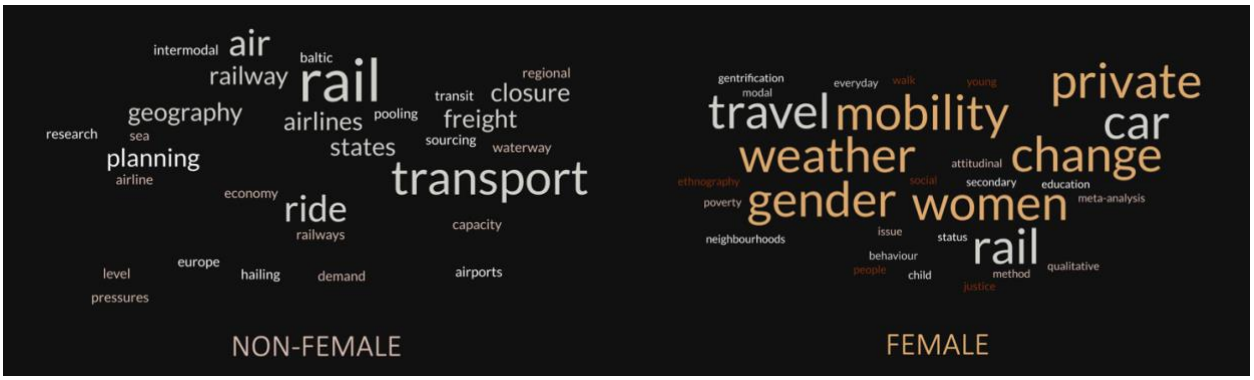


Fig. 5 – Word clouds of topics uniquely to non-female vs. female groups

GIRL POWER

in Transportation Geography

(Topic Analysis on female authors' contribution)

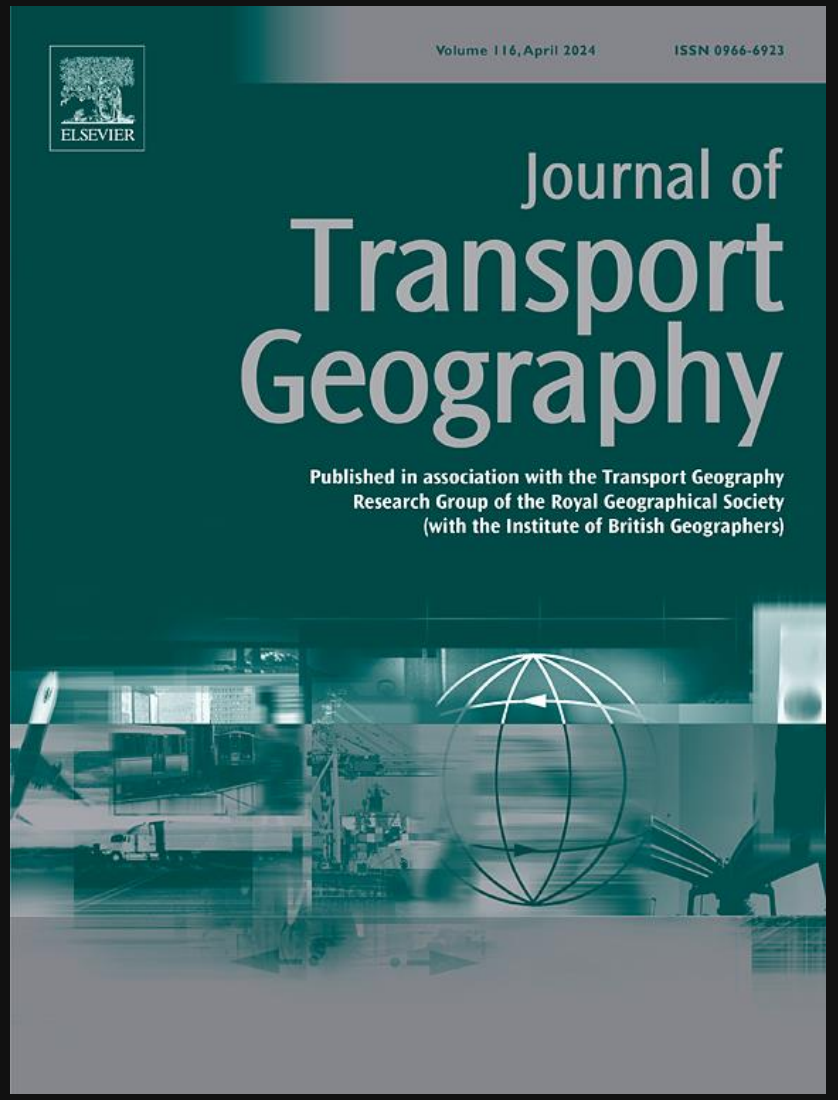
CAPSTONE | SPING 2024

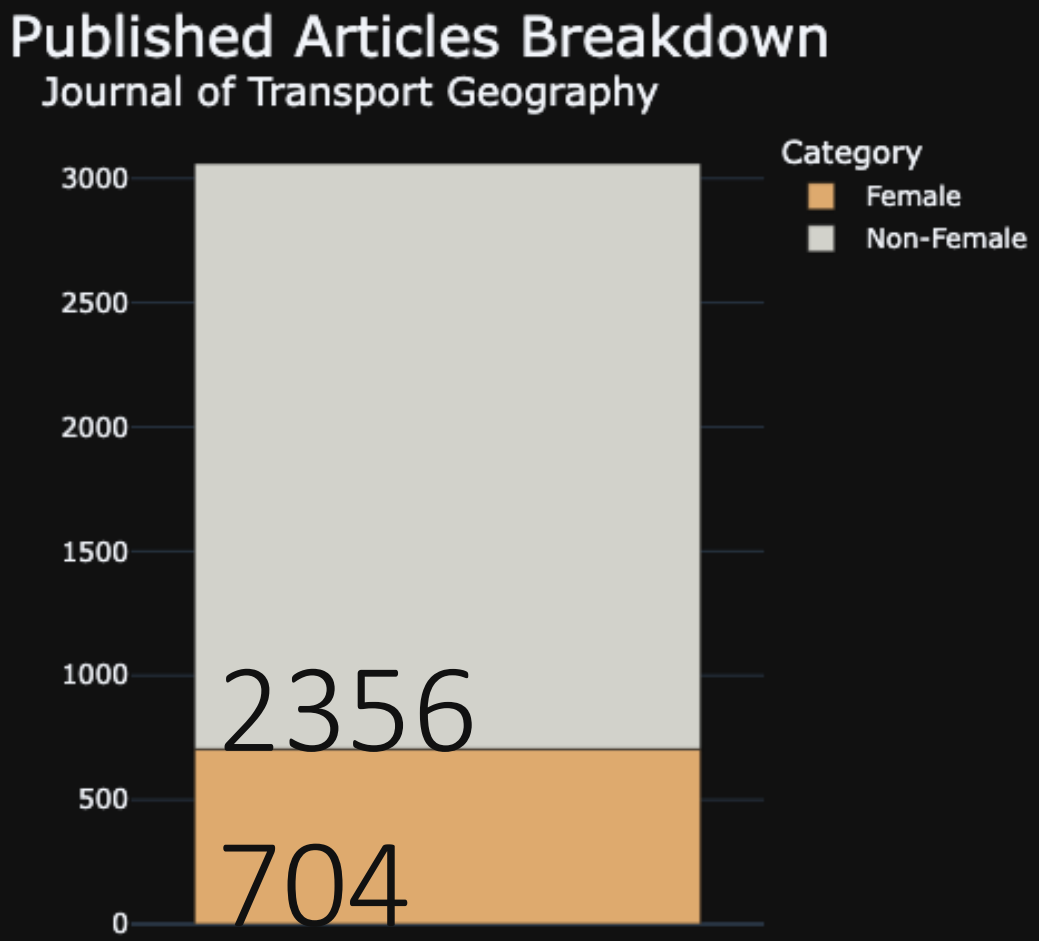
ANN ZIAN ZHANG & SOFIA FASULLO

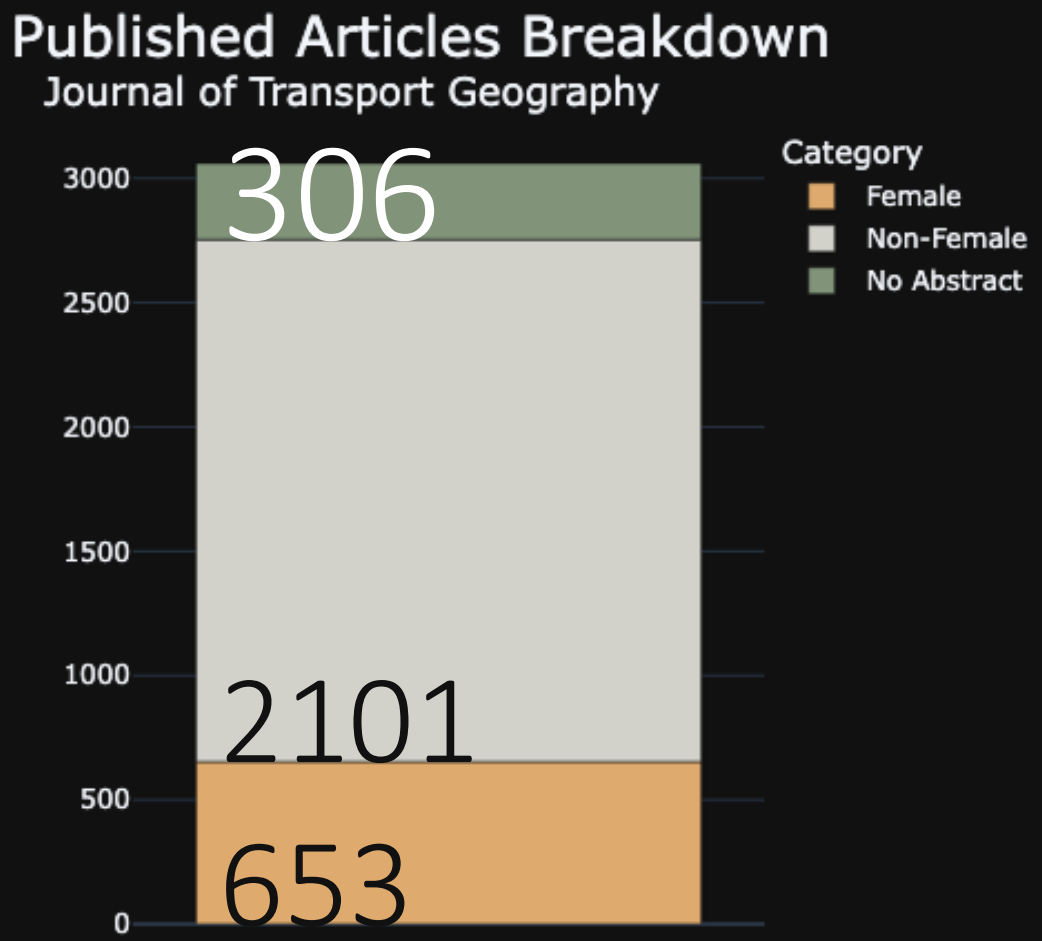
- Background
- Data Sources + Framework
- Part I - Key Words Frequency Count
- Part II - BERTopic Text analysis
- Part III - Regression
- Comparative Analysis & Conclusion

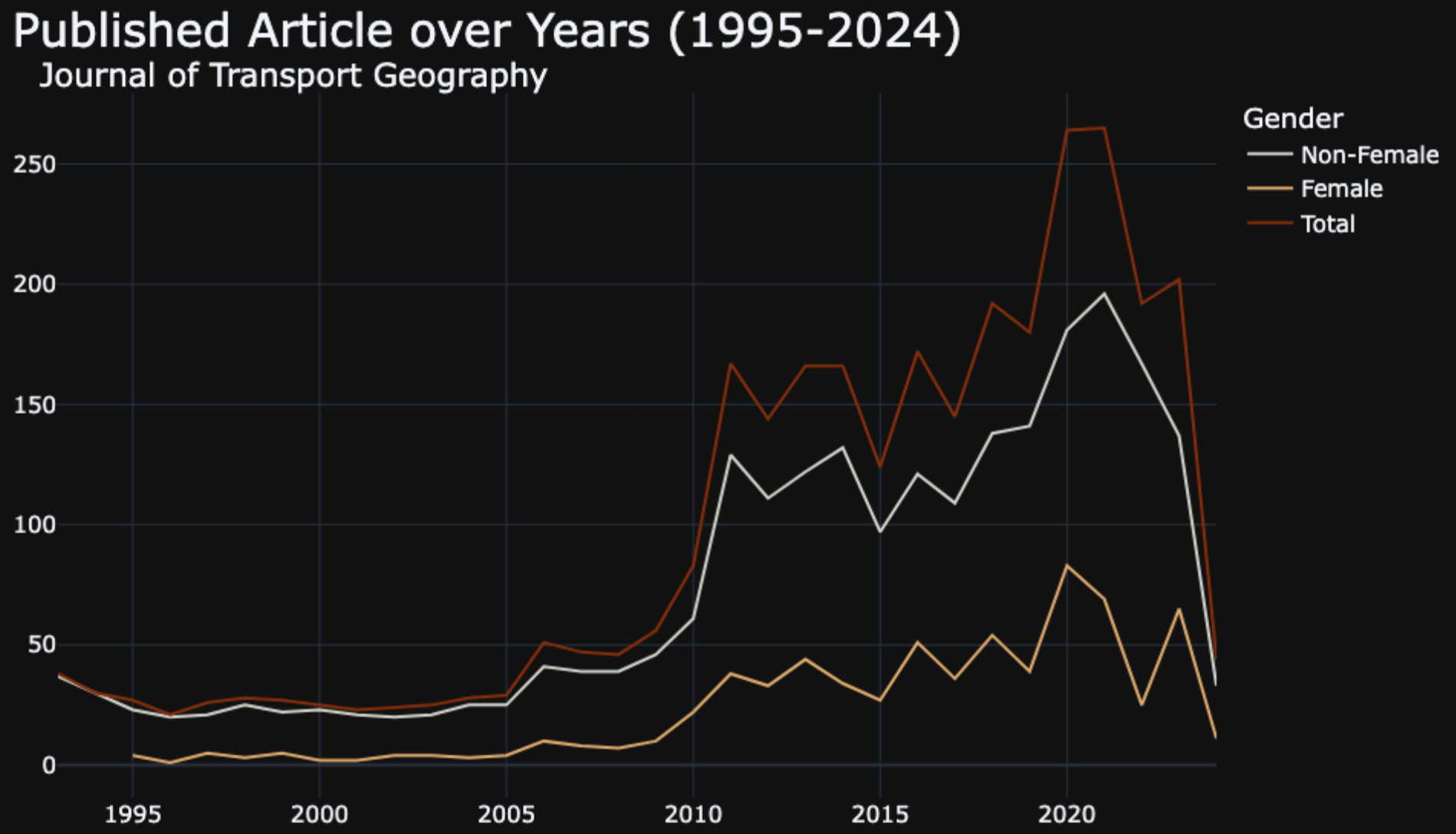
BACKGROUND

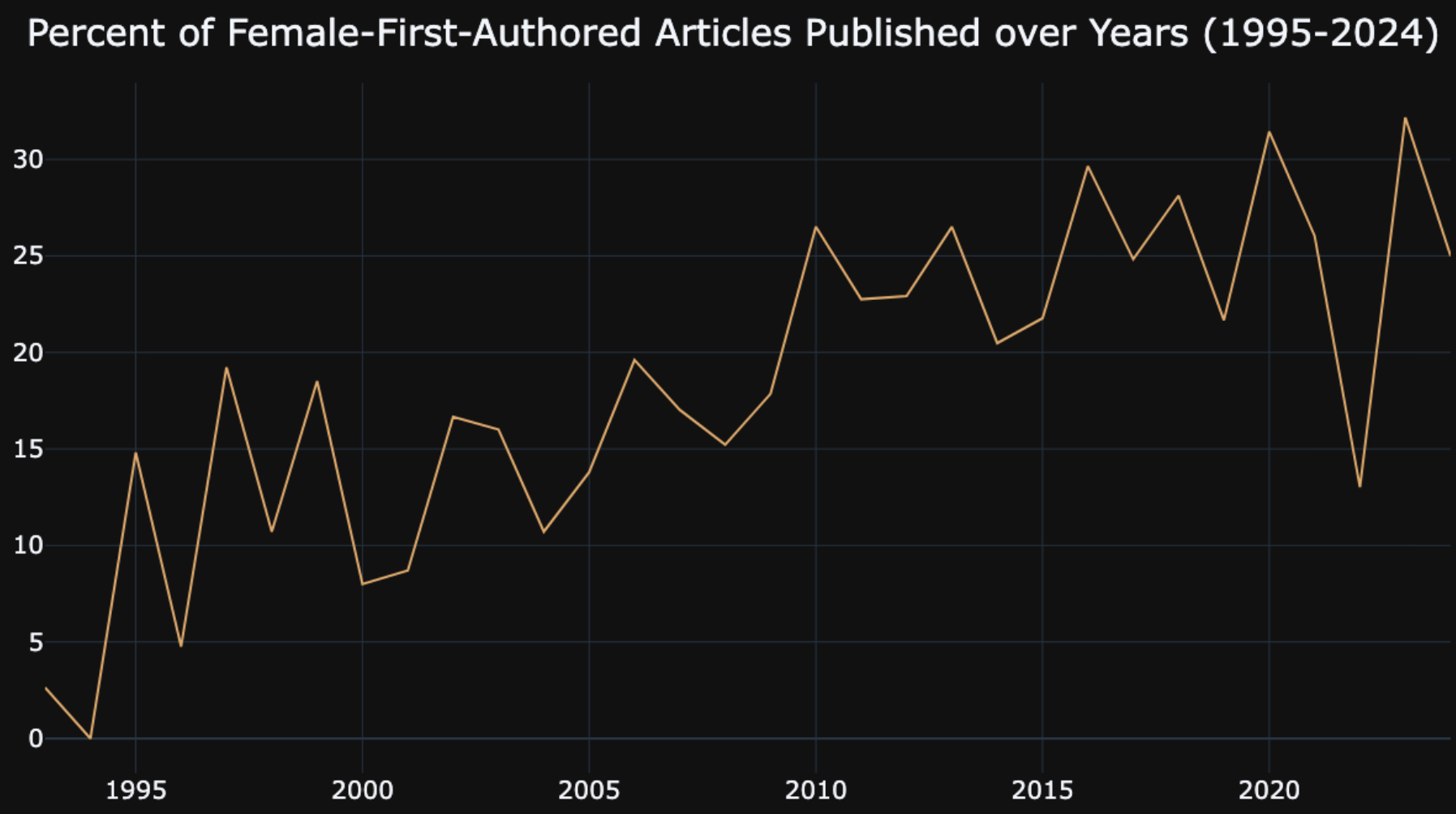
JOURNAL OF TRANSPORT GEOGARPHY




Published Articles Breakdown
Journal of Transport Geography
Category
Female
Non-Female
3000
2500
2000
1500
1000
500
0
2356
704


Published Articles Breakdown
Journal of Transport Geography
Category
Female
Non-Female
No Abstract
3000
2500
2000
1500
1000
500
0
306
2101
653


Published Article over Years (1995-2024)
Journal of Transport Geography
Gender
Non-Female
Female
Total
250
200
150
100
50
0
1995
2000
2005
2010
2015
2020

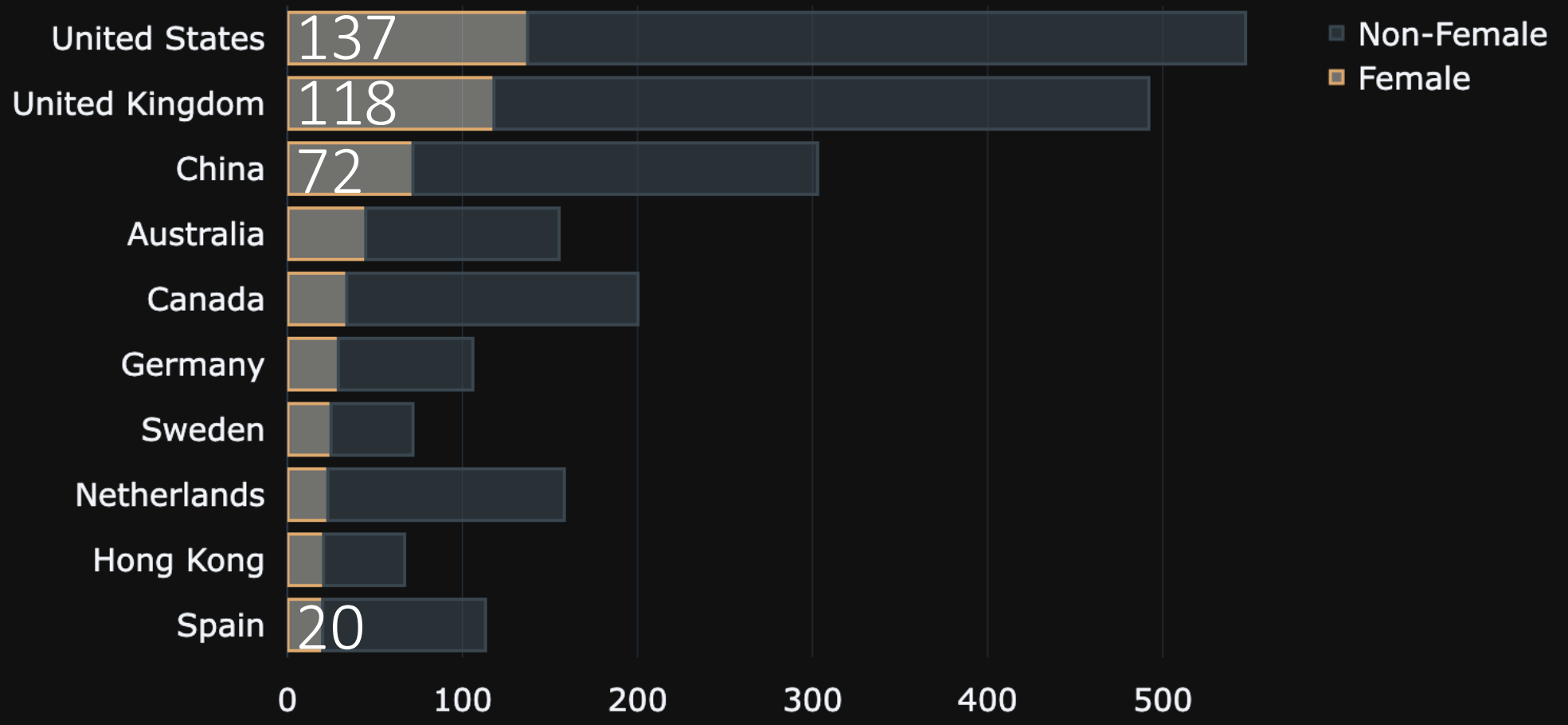

Percent of Female-First-Authored Articles Published over Years (1995-2024)
30
25
20
15
10
5
0
1995
2000
2005
2010
2015
2020

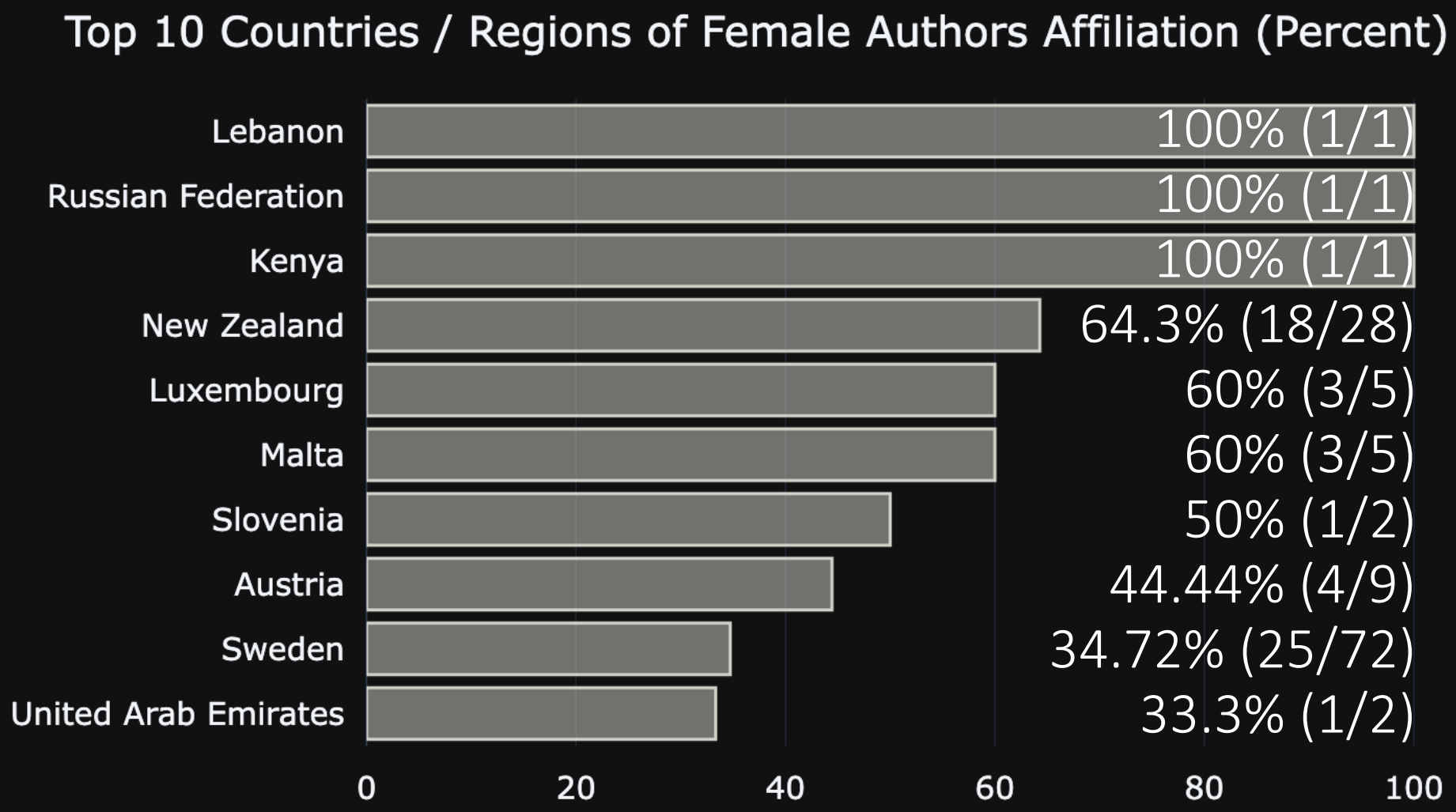
NON-FEMALE		FEMALE
9.13	PAGE COUNT	9.22
39.59	CITATIONS	40.99
2.65	CO-AUTHORS	2.89

FEMALE'S WORK AT A GLANCE

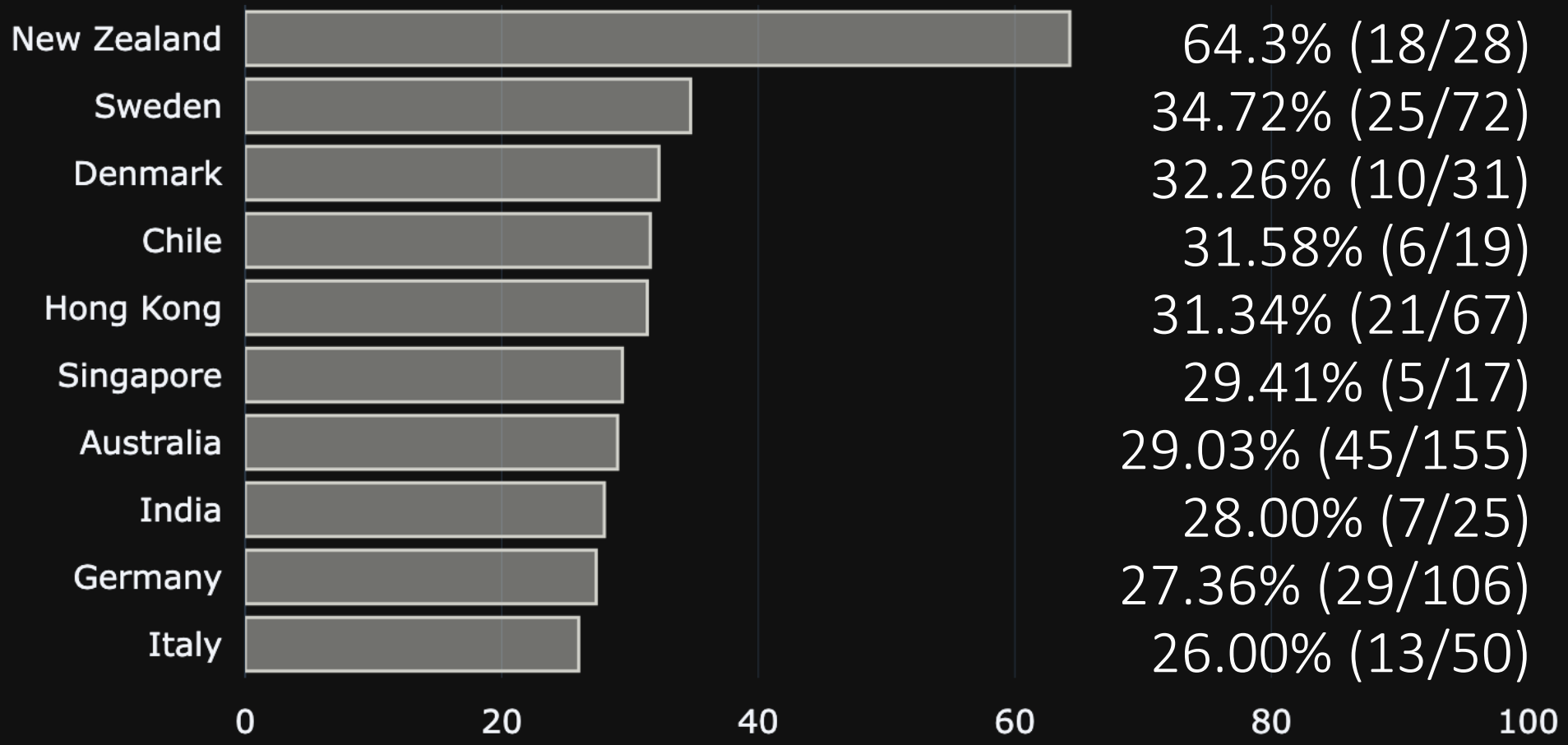
*Affiliation ≠ Nationality

Top 10 Countries / Regions of Female Authors Affiliation




Top 10 Countries / Regions of Female Authors Affiliation (Percent)
Lebanon
Russian Federation
Kenya
New Zealand
Luxembourg
Malta
Slovenia
Austria
Sweden
United Arab Emirates
100% (1/1)
100% (1/1)
100% (1/1)
64.3% (18/28)
60% (3/5)
60% (3/5)
50% (1/2)
44.44% (4/9)
34.72% (25/72)
33.3% (1/2)
0
20
40
60
80
100

Top 10 Countries / Regions of Female Authors Affiliation (Percent)



64.3% (18/28)
34.72% (25/72)
32.26% (10/31)
31.58% (6/19)
31.34% (21/67)
29.41% (5/17)
29.03% (45/155)
28.00% (7/25)
27.36% (29/106)
26.00% (13/50)

DATA SOURCES

- Scopus (with Elsevier)
  - Downloaded information about all published articles published on the Journal of Transport Geography
  - Info includes Authors, Year, Page Count, Citations, Abstracts, Funding Sources, Affiliations, Index Words, Author Keywords, etc.
- Gender Info – Manually Input from Online Search
  - Collaborative effort of Prof. Elizabeth Delmelle, Ann & Sofia

RESEARCH QUESTIONS

- How have female-identifying researchers been contributed to the field of Transport Geography?
- Is there anything special (regarding the topics) about female-first-authored articles?
- Goal – to advocate further for equality and open up more research funding and opportunities to female researchers

RESEARCH FRAMEWORK

PART I
Keyword Counts

PART II
BERT Topic Identification

PART III
Logistic Regression

PART I – Keyword Counts

METHOD

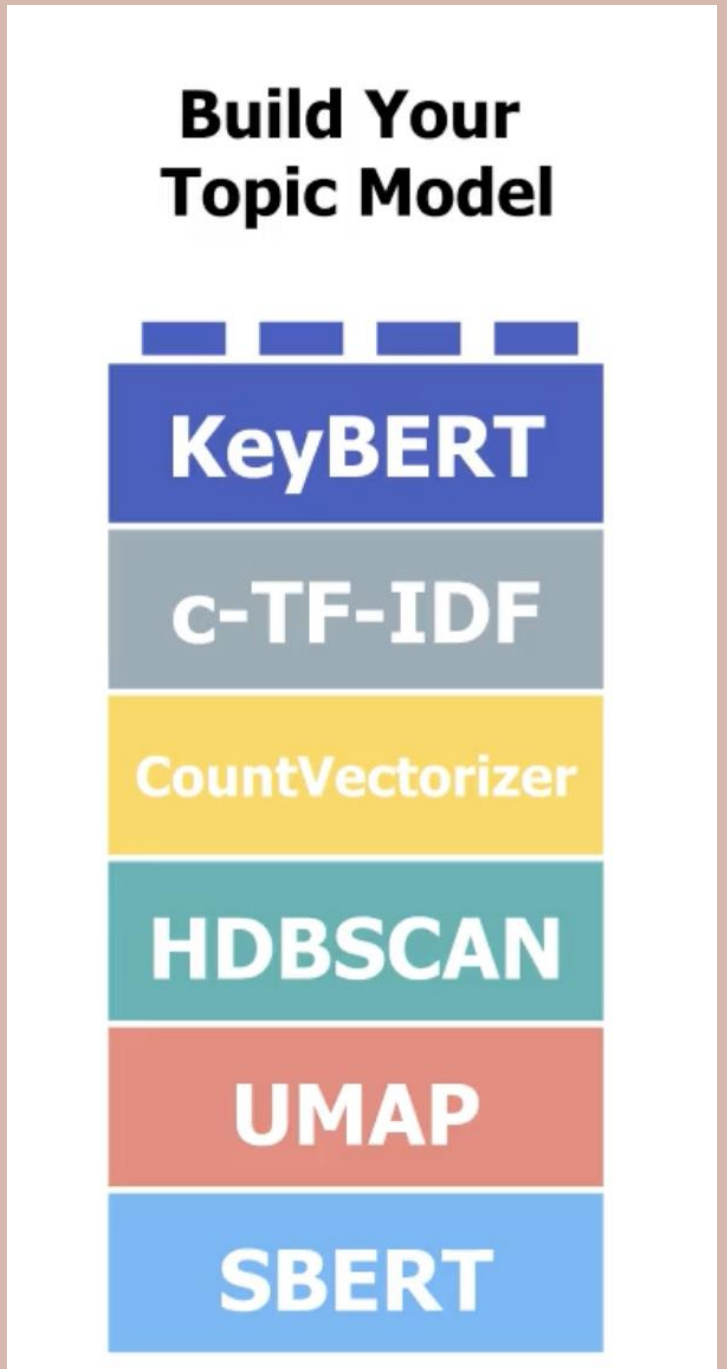
- Split dataset into female vs. non-female first author groups
- Splitting individual word in abstract
- Remove common stopper words
- Then count frequency of words
- Compare the Top 10 most frequently used words across 2 groups

PART I
RESULTS

FEMALE	NON-FEMALE
Mobility	Accessibility
Urban	Urban
Accessibility	Spatial
Public	Model
Analysis	Analysis
Areas	Public
Spatial	Network
Different	Time
Research	Areas
Network	Transportation

PART II - BERTOPIC
METHOD

6 TUNE TOPIC REPRESENTATION
5 WORD-WEIGHT SCHEME
4 TOKENIZE DOCUMENTS
3 CLUSTERING - REDUCE EMBEDDINGS
2 REDUCING DIMENSIONALITY
1 DOCUMENT EMBEDDINGS



INPUT

["Ride-pooling systems, despite being an appealing urban mobility mode, still struggle to gain momentum. While we know the significance of critical mass in reaching system sustainability, less is known about the spatiotemporal patterns of system performance. Here, we use 1.5 million NYC taxi trips (sampled over a six-month period) and experiment to understand how well they could be served with pooled services. We use an offline utility-driven ride-pooling algorithm and observe the pooling potential with six performance..... ",

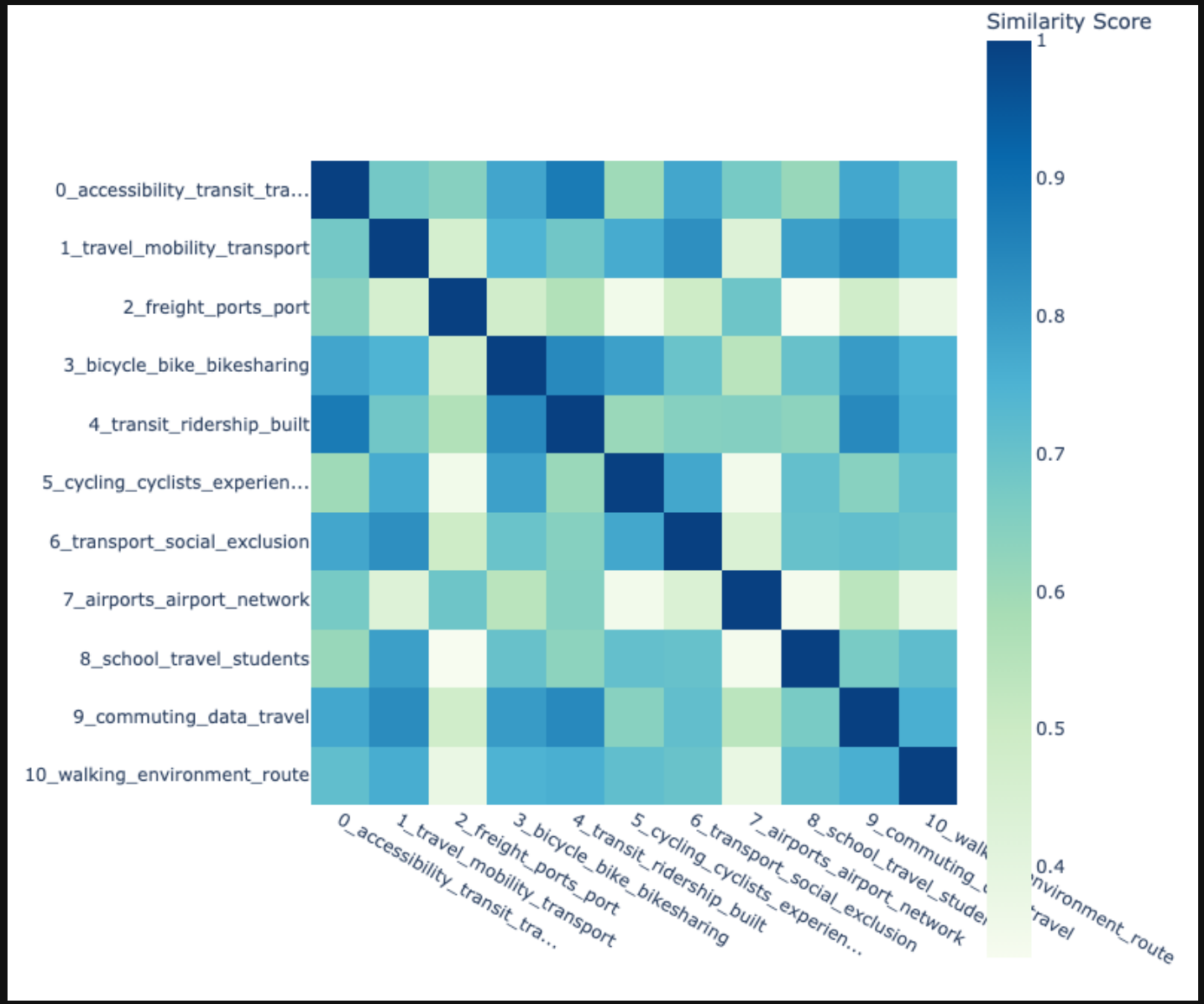
OUTPUT

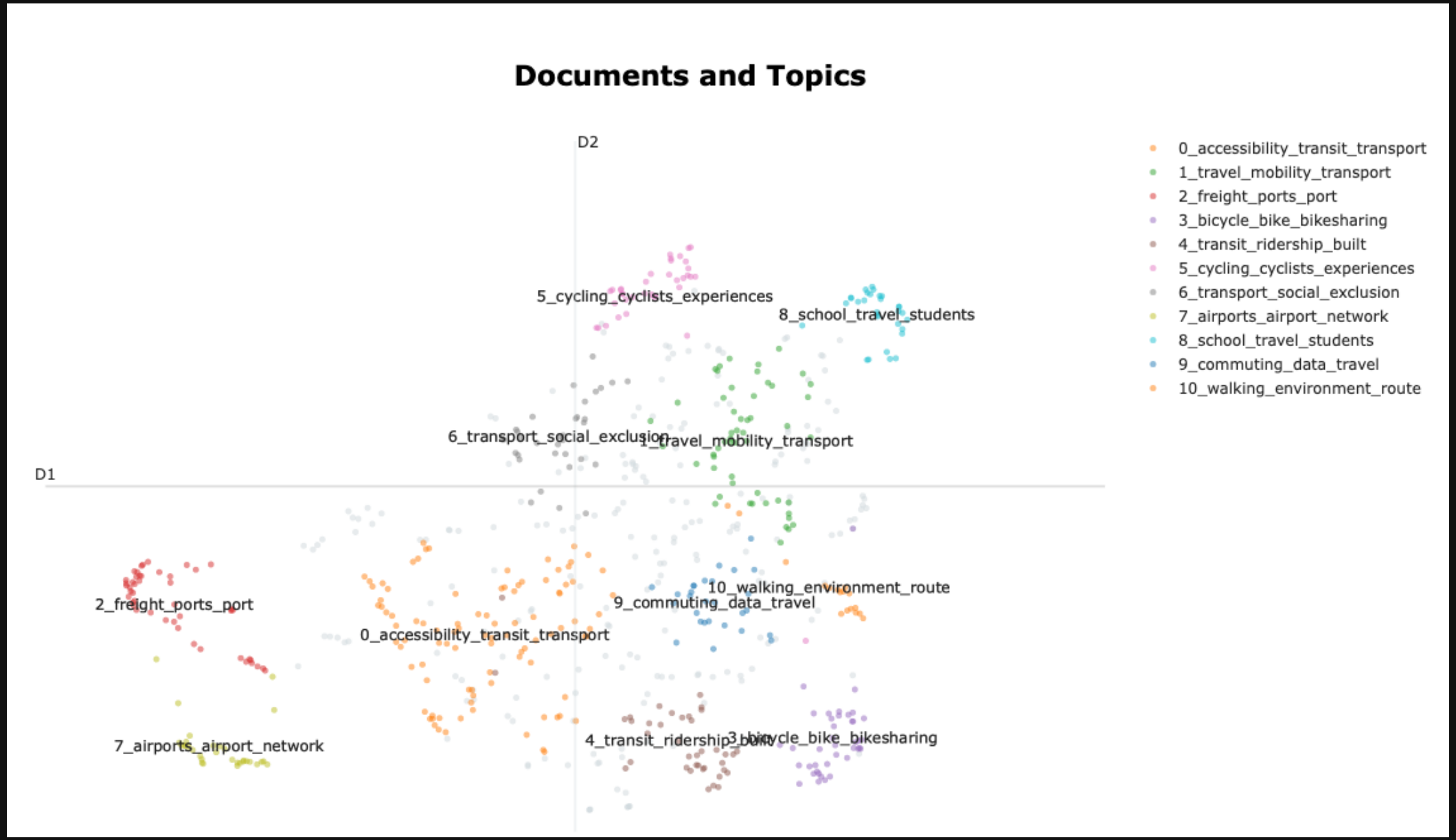
Topic	Count	Name	Representation
-1	211	-1_transport_travel_mobility_urban	[transport, travel, mobility, urban, use, pape...
0	95	0_accessibility_transit_transport_public	[accessibility, transit, transport, public, ne...
1	52	1_travel_mobility_transport_time	[travel, mobility, transport, time, car, life,...
2	46	2_freight_ports_port_logistics	[freight, ports, port, logistics, shipping, tr...
3	39	3_bicycle_bike_bikesharing_public	[bicycle, bike, bikesharing, public, data, tra...
4	38	4_transit_ridership_built_built environment	[transit, ridership, built, built environment,...
5	36	5_cycling_cyclists_experiences_transport	[cycling, cyclists, experiences, transport, ur...
6	33	6_transport_social_exclusion_mobility	[transport, social, exclusion, mobility, socia...
7	31	7_airports_airport_network_air	[airports, airport, network, air, hub, aviatio...
8	29	8_school_travel_students_children	[school, travel, students, children, parents, ...
9	27	9_commuting_data_travel_segregation	[commuting, data, travel, segregation, built, ...
10	16	10_walking_environment_route_choice	[walking, environment, route, choice, pedestri...

BERT (1)

TOPICS OF FEMALE FIRST AUTHOR ARTICLES

FEMALE TOPICS SIMILARITY MATRIX




Documents and Topics
D2
D1
5_cycling_cyclists_experiences
8_school_travel_students
6_transport_social_exclusion
1_travel_mobility_transport
2_freight_ports_port
10_walking_environment_route
9_commuting_data_travel
0_accessibility_transit_transport
7_airports_airport_network
4_transit_ridership_built
3_bicycle_bike_bikesharing
0_accessibility_transit_transport
1_travel_mobility_transport
2_freight_ports_port
3_bicycle_bike_bikesharing
4_transit_ridership_built
5_cycling_cyclists_experiences
6_transport_social_exclusion
7_airports_airport_network
8_school_travel_students
9_commuting_data_travel
10_walking_environment_route

BERT (2)

TOPICS OF ALL ARTICLES & COMPARISONS

FEMALE	NON-FEMALE
TOPIC 1	TOPIC 0
TOPIC 0	TOPIC 2
TOPIC 4	TOPIC 1
TOPIC 2	TOPIC 3
TOPIC 5	TOPIC 6
TOPIC 3	TOPIC 5
TOPIC 7	TOPIC 8
TOPIC 8	TOPIC 7
TOPIC 9	TOPIC 4
TOPIC 6	TOPIC 10

NON-FEMALE

TOPIC 0
TOPIC 2
TOPIC 1
TOPIC 3
TOPIC 6
TOPIC 5
TOPIC 8
TOPIC 7
TOPIC 4
TOPIC 10

FEMALE

TOPIC 1 [cycling, bicycle, bike, bikeshare]
TOPIC 0 [port, ports, logistics, freight]
TOPIC 4 [school, students, children, travel]
TOPIC 2 [air, airports, airport, airline]
TOPIC 5 [commuting, commute, workers, work]
TOPIC 3 [accessibility, economic, infrastructure, transport]
TOPIC 7 [accessibility, transport, social, research]
TOPIC 8 [older, older adults, adults, elderly]
TOPIC 9 [ridership, bus, transit, built]
TOPIC 6 [HSR, Cities, China, economic]

(10) [patterns, data, travel, information]

PART III (A) - REGRESSION
METHOD

- Logistic Regression
- Dependent Variable (y) = Female / Non-Female (1 / 0 dummy)
- Independent Variable = Unique Keywords (Convert into dummy)
  - BERT Topics – 47 Variables (337 if Exploded)
  - Index Words – 2573 Variables
  - Author Keywords – 6878 Variables
- Interpret Coefficient

REGRESSION on BERT TOPICS

RESULTS (Topics)

- Accuracy – 0.7554

1. Practices, Mobility, Transport, Private, Change [TOPIC 24]
2. Rail, Neighborhood, Neighborhoods, light rail [TOPIC 41]
3. School, students, children, travel, childrens [TOPIC 4]
4. Weather, ridership, weather conditions [TOPIC 22]
5. Gender, car, women, men, car use [TOPIC 23]

1. Railway, Europe, Baltic, States, Transport [TOPIC 35]
2. Rail, Freight, Closure, Intermodal [TOPIC 38]
3. Ride Hailing, Ride Sourcing, Transit, Ride Pooling [TOPIC 18]
4. Geography, Research, Transport, Geography [TOPIC 28]
5. Air, Airports, Airline, Airlines [TOPIC 2]

REGRESSION on BERT TOPICS

RESULTS (Explode Topics into Keywords)

- Accuracy – 0.7599

1. Change
2. Everyday
3. Mobility Practices
4. Behaviour
5. Modal
6. Private
7. Practices
8. Neighbourhoods
9. Gentrification
10. Weather

1. Railway
2. Pressures
3. Closure
4. Sea Level
5. States
6. Rail Freight
7. Rail / Air
8. Capacity
9. Railways
10. Demand

REGRESSION on INDEX WORDS

RESULTS

- Accuracy – 0.7664

Words Most Associated with Female-First Author

1. Secondary Education
2. Child
3. Women Status
4. Gender Issue
5. Charlotte
6. Meta-analysis
7. North Carolina
8. New Zealand
9. Qualitative Analysis
10. Poverty

1. Waterway Transport
2. Regional Economy
3. Planning Process
4. Honshu
5. Paris
6. Model Test
7. Tokyo (Kanto)
8. Infrastructure Planning
9. Africa
10. Sensitivity Analysis

REGRESSION on INDEX WORDS

RESULTS

1. Secondary Education
2. Child
3. Women Status
4. Gender Issue
5. Meta-analysis
6. Qualitative Analysis
7. Poverty
8. Survey Method
9. Attitudinal Survey
10. Social Justice

1. Waterway Transport
2. Regional Economy
3. Planning Process
4. Model Test
5. Infrastructure Planning
6. Sensitivity Analysis
7. Distribution System
8. Uncertainty Analysis
9. Classification
10. Service Sector

REGRESSION on AUTHOR KEYWORDS

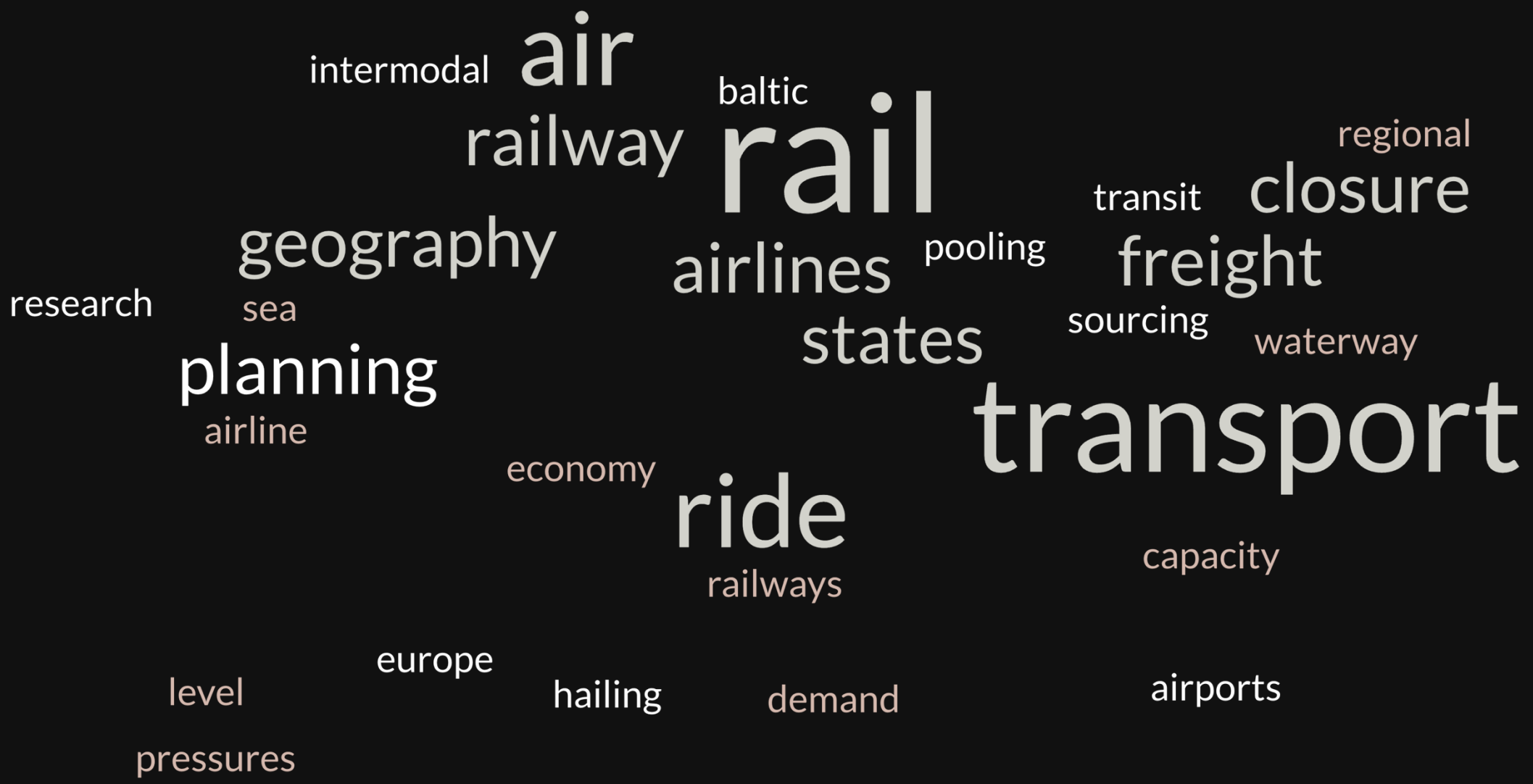
RESULTS

- Accuracy – 0.7698

Words Most Associated with Female-First Author

1. Walk
2. Ethnography
3. Young People
4. Content Analysis
5. Quality of Service
6. Information and Communication Technologies
7. Travel Diary
8. Motility
9. Mobile methods
10. Inequalities

1. Governance
2. Light Rail
3. Location
4. Airlines
5. Air Transport
6. High Speed Rail
7. Networks
8. Globalization
9. Developing Countries
10. Regulation


intermodal
air
baltic
railway
rail
regional
transit
closure
geography
pooling
freight
airlines
research
sea
sourcing
states
waterway
planning
transport
airline
economy
ride
capacity
railways
europe
level
hailing
demand
airports
pressures
NON-FEMALE


gentrification
everyday
walk
young
private
modal
travel
mobility
car
weather
change
attitudinal
ethnography
social
secondary
education
poverty
gender
women
meta-analysis
issue
neighbourhoods
status
rail
behaviour
qualitative
people
child
method
justice
FEMALE

GIRL POWER
Conclusion

Change
Everyday
Mobility Practices
Behaviour
Modal
Private
Practices
Neighbourhoods
Gentrification
Weather
Secondary Education
Child
Women Status
Gender Issue
Meta-analysis
Qualitative Analysis
Poverty
Survey Method
Attitudinal Survey
Social Justice
Walk
Ethnography
Young People
Content Analysis
Quality of Service
Info and Comm Technologies
Travel Diary
Motility
Mobile methods
Inequalities

Bibliography

J. Ignacio Conde- Ruiz, Juan-Jose Ganuza, Manu Garcia, and Luis A. Puch, "Gender distribution across topics in the top five economics journals: a machine learning approach," *Journal of the Spanish Economics Association*, no.13 (Nov 2022): 269 – 308.

"BERTopic," https://maartengr.github.io/BERTopic/index.html#fine-tune-topic-representations, accessed May 12, 2024.