A phantom gap? Center and region in Dutch political debates (c. 2013-2023)

A quantitative discourse analysis

Ruben Land

Student number: 6382797

Supervisors: Dr. Evert Meijers and Dr. Martijn Smit

Program: MSc Applied Data Science

Institution: Utrecht University

June 29, 2024

Contents

Abstract

The Netherlands is sometimes perceived as polarized between the urban center named the Randstad including the four main cities Amsterdam, Den Haag, Rotterdam and Utrecht, and the rest of the country, the *regio*. The gas extraction-induced earthquakes in Groningen for example pit the center against this province. People in Groningen feel that their province is exploited to fill the state coffers, while their concerns are not being taken seriously by the politicians in Den Haag. Survey research indeed finds a divide between the center and periphery, as the *regio* finds itself relatively neglected by the national politics. To date, no research has been conducted on the existence of this divide in the political debates themselves. This thesis aims therefore to map the extent to which the contexts of places in the Randstad and *regio* differ in political debates. Different contexts could hint at disparate treatment. This is done by using the word embedding model `word2vec` for mapping toponyms in vector space. Larger distances in vector space indicate more divergent contexts. Distances in vector space are then modelled in a gravity model. Results show that the division between Randstad and *regio* only effects the context around places in a minor way, indicating the nuance there is in the debates. Individual place characteristics, for example, are of a greater importance.

Key words: network analysis — politics — word embeddings — `word2vec` — The Netherlands — toponyms

1 Introduction

The Netherlands is often divided into a polarity of the main urbanized area in the West, the Randstad, and the rest of the country, the *regio*. Although not as prominent as in the United States, where there exists a clear division between the so-called 'Red' and 'Blue' America [29], one can still observe that voting patterns differ between the Randstad and *regio*. People in the periphery feel that their local concerns are not taken seriously by the center, fueling regional resentment and an increased populist vote [12]. Villages, towns, cities and regions constitute not only physical places, but also bear symbolic meaning to which people attach feelings and base their identity on [11, 23]. As such, toponyms are not neutral in politics. Politicians can use them to attract potential voters from local constituencies that feel disenfranchised. The aim of this thesis is to examine if a distinction exists in the context in which places from either the Randstad or *regio* generally appear in the parliamentary debates of the Dutch House of Representatives, the *Tweede Kamer der Staten-Generaal*. Research thus far focuses mainly on survey research to map the supposed divide between center and periphery. Although previous research already investigated Dutch city networks using co-occurrences [26], mapping places based on their common context is not yet done to my knowledge and opens new doors for discourse analysis of toponyms.

Parliamentary debates cover a wide range of topics, such as environment, energy and housing. Political debates thus offer a promising way of assessing the divide given this diversity of topics. In particular, I expect to find differences in the context surrounding place names from either the rural or strongly urbanized part of The Netherlands. This thesis thus revolves around the following research question: *To what extent are places from the Randstad and regio distinctly embedded in Dutch parliamentary debates*? So, evidently, the gap between Randstad and *regio* here is measured in terms of how different the contexts are in which locations of either category appears. These contexts do not need to be more positive or negative for the Randstad or *regio*. They can simply just be different.

The analysis centers around the period 2013-2023 and is chosen out of practical considerations, since the application programming interface (API) of the Open Data Portal of the TK only covers this period. Nonetheless, the time frame is also substantially relevant, as it marks the period in which the divide between center and region became more prominent in politics and media. One example is the nitrogen crisis, where the government and farmers were pitted against each other, for the reasons that the government had to substantially reduce the nitrogen emissions by livestock farmers. It was already known since the 1970's, but

the farmer's unions always opposed measures to reduce the emissions. Judicial decisions forced the Rutte III administration to finally act on the problem in 2019. One of its solutions was buying out farmers. On their turn, the farmers protested against these measures that would supposedly threaten their way of life by blocking high ways with their tractors, for example [36].

Another example is the course of action taken by the government in dealing with the damages caused by the gas extraction in the province of Groningen. The Dutch Petroleum Company (*Nederlandse Aardolie Maatschappij*, NAM) discovered in 1959 a large gas field in Slochteren, Groningen. The gas field turned out to be huge and the NAM, backed by the State, extracted gas from it ever since. Earthquakes were being measured since 1976. Small in the beginning, but increasing in severity ever since. Still, the NAM rejected the idea that the gas extraction caused the earthquakes. Only after a M3.6 earthquake on the Richter scale in 2012 below the Huizinge village the political landscape started to shift towards reducing gas extraction. Nevertheless, it took some years before the administration decided to stop the extraction all together. In the meanwhile, inhabitants faced difficulties receiving compensation for damages, because the NAM acted unhelpful. All of this made the population of Groningen feel that the province was treated as a colony (*wingewest*), where the central government took the profits from the gas in Groningen without returning a significant share of the profits to the province. [15, 37]. The West part of the country profited at the expense of the 'far away' Northern province.

In short, there are perceived tensions and differences between the center and the rest. The main question is if such a partition is also present in the parliamentary debates. A quantitative approach is appealing, given the large amount of texts. Underlying the methodology is the idea that different toponyms are used in varying contexts. Contexts are here words in the vicinity of the toponym. So, we would expect that a place in the Randstad like Amsterdam appears in quite different settings than for example Maastricht in the very South of the country. Groningen might well appear in even more distinct contexts, as the province grabs the headlines frequently due to gas extraction induced earthquakes. Word embeddings are one way of quantifying the meaning of a word given its context, as they represent a word in a multidimensional vector that captures the meaning of a word in diverse contexts. This investigation thus revolves around comparing these vectors based on if places belong to the Randstad or *regio*.

Investigating the hypothetical contextual gap between the Randstad and *regio* entails conducting a network analysis that allows to distinguish in network terminology the center and periphery, respectively. Periphery has a negative connotation, however, as it implies a dependency and subordination to the center [15]. Furthermore, these terms are too generalizing. Eindhoven, for example, lies in the so-called periphery but is a scientific hub. Nevertheless, these terms are often used in scientific work as they are of use in clarifying the power distribution [38].

This thesis aims to contribute to research on the geographic divide in two ways. First and foremost, it analyzes the supposed chasm between the Randstad and the periphery through political debates. In this way it contributes to studies about spatial polarization of Dutch society [12, 20]. Although the Randstad is not an official administrative unit, people often refer to it as a whole [7]. Moreover, investigates the potential appearance of a geographical divide in the political arena, thus moving from survey research to the domain of textual analysis of a large corpus by applying traditional as well as relatively new statistical methods. Hence, this research also evaluates with this case study the applicability of word embeddings in discourse analysis.[1]

[1]All code is made public. See the repository: [24]

2 Data

Debates of the House of Representatives are extracted from the Open Data Portal of the TK in April 2024, spanning from 17 September 2013 to 7 March 2024 [31]. These plenary debates are public, allowing anyone to consult them. In order to filter the relevant data, all available end publications of debates are retrieved using the application programming interface (API) in Extensible Markup Language (XML) format. In total there are 1120 debates in the period of June 2013 until March 2024, excluding a small number of debates without an end publication of the transcription. The XML files are rich in detail. For example, the file describes the time that a debater starts their argument. Not all of these details are relevant here. Instead, only the names of the speakers, their utterances and party, and the topic and date of the debate are extracted. The party is not named if the speaker's position is Minister or Staatssecretaris.

The data are filtered by preserving debates about nature and environment, agriculture, housing and spatial planning and infrastructure. Although the debates address many other topics, such as foreign affairs, the societal tensions between Randstad and *regio* come particularly to the fore with respect to the environment, as mentioned in the Introduction. The analysis thus limits itself to these topics as to have a particular focus.

Some preprocessing steps followed thereafter, namely removal of interpunction, numbers, stopwords and lemmatization. Capital letters were left in place to facilitate retrieving toponyms. Stop words were selected based on a standard list of Dutch stop words, supplemented by around the two hundred fifty most frequent words in the debates, such as *voorzitter* (chairperson) and the definitive article *de*. Words such as these appear so often that they do not carry meaningful information for the particular analysis at hand [28]. Furthermore, inflected words are lemmatized, because the inflections cause the model to treat words with the nearly identical meaning as different terms. Lemmatization entails bringing back words to their dictionary form. So, *werd* becomes *worden* and *steden* becomes *stad*. The `udpipe` tool lemmatizes tokens with an accuracy of approximately ninety percent for Dutch [35]. These preprocessing steps were necessary to retain meaningful words and standardize them, given the purposes of this analysis [8]. Table 1 displays an example of the cleaned version of the data. The `category` of the debate to which the utterance belongs is not present in this example to prevent the table from being too large.

Table 1: *Example rows of preprocessed subset*

speaker	party	date	preprocessed utterance
Dik-Faber	ChristenUnie	2019-06-20	lang Intussen zomer beginnen
Adema	ChristenUnie	2023-05-11	Wassenberg verbeterplan NVWA discussie Eersel
Ronnes	CDA	2016-12-22	corporatie weinig inzicht bieden verduurzaming
Adriaansens	VVD	2022-11-23	danken Amhaouch strategische autonomie maritiem
Weverling	VVD	2020-11-24	naïef bashen tuinbouw Futselaar Nederland
Boulakjar	D66	2023-02-22	vorig week tweet stof
Moorlag	PvdA	2019-11-19	blij Verhoeven detailhandel agenderen smaak
Ollongren	D66	2018-12-13	dingen stelselwijziging zodanig inhoudelijk kennis

Next, besides the debates themselves geographic data is required. I chose to only use toponyms of inhabited places, so administrative entities such as provinces are excluded to keep the analysis focused and the units of analysis comparable. Groningen and Utrecht are the names of both the provinces and their capitals. Groningen city is left out, because politicians refer more than ninety percent of the time to the province instead of the city, in contrast to Utrecht, where they mean most often the city.[2] A publicly available dataset containing variables such as place names, WGS84 coordinates and province was enriched with the number of inhabitants per place using the geographical database GeoNames [14, 19]. Unfortunately, GeoNames does not state when

[2]This is based on a hand coded random sample of fifty utterances each for Utrecht and Groningen.

Table 2: *Corpus and sample frequencies excluding stop words*

Corpus	Frequency
Debates	1,120
Utterances	535,841
Total tokens	69.981.918
Unique tokens	265.604
Subset	
Debates	521
Utterances	77,250
Total tokens	10,728,305
Unique tokens	89,607
Toponyms	
Count toponyms	7,152
Unique toponyms	142
Randstad	32
Regio	110

the population sizes were recorded, but the final updates to the selected places range from 2017 to 2022. Moreover, some ambiguous names were not apt for consideration. The analysis left toponyms out that could refer to other parts of speech such as nouns, adjectives and surnames [26]. Examples include Dieren (animals) and Bergen (mountains). As such, ambiguous place names are excluded to minimize the chance for false positives. See the Appendix for a list of included and excluded places (Table A2, Table A1). The largest place will be linked to that name in case that multiple places bear the same name. Hengelo is for instance a town in Gelderland and Overijssel, the one in Overijssel being larger in population size.

Table 2 displays the frequencies of the full corpus including all debates and those of the subset containing only the debates on nature and environment, agriculture, housing, and spatial planning and infrastructure. Less than half of the debates concern these subjects. Only around fourteen percent of the utterances is present in the subset, partly due to preprocessing. The final selection includes almost eleven million tokens, of which approximately ninety thousand unique.

Basic filtering of the corpus by means of matching returned 142 place names. The places are nicely spread over the country. Some regions are, however, less represented, the province of Zeeland being an example (Figure 1). Major cities as well as smaller villages are present in the data. Another feature that catches the eye is the high amount of unique places in the province of Groningen, which is to a large extent attributable to the earthquakes there. There exists a correlation between the size of the place and its number of mentions nonetheless ($r(140) = .75$, $p < .001$). Table 3 provides elementary descriptive statistics on the toponyms. Noticeable is the large range of population sizes, spanning from the small villages Krewerd and Saaksum in Groningen to the largest city Amsterdam. Distances between places retrieved in the text are also quite varied, as they are spread over the whole country.

The Randstad can have narrower or broader definitions. Narrowly it would only encompass the 'Big Four', namely Amsterdam, Den Haag, Rotterdam and Utrecht. Here, however, I follow Burger and colleagues (2014) by including the surrounding municipalities as well [7]. Places in municipalities that are border to border within a radius of ten kilometres from the Big Four municipalities are considered to belong to the Randstad, because these are well integrated with the Big Four. A lot of people live for instance in Almere or Amstelveen, but work in Amsterdam (Figure 2).

Table 3: *Toponym descriptives*

Variable	Mean	Sd	Min	Max
Population size	$57,047.60$	$98,695.87$	90	$741,64$
Physical distance	116.63	64.81	1.99	310.74

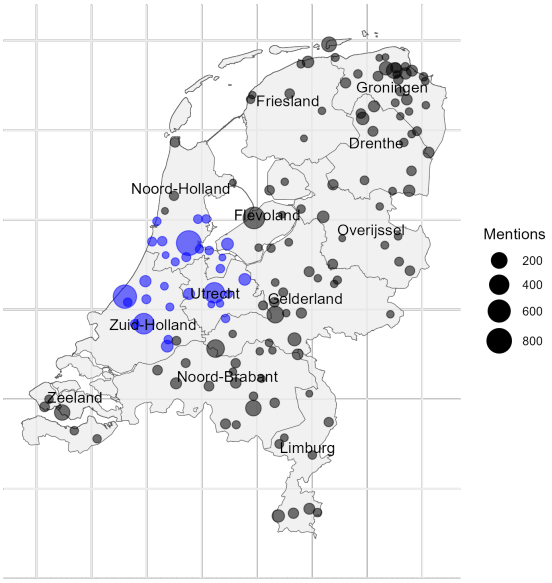

Figure 1: *Toponym frequencies in the subset, Randstad in blue* [17]

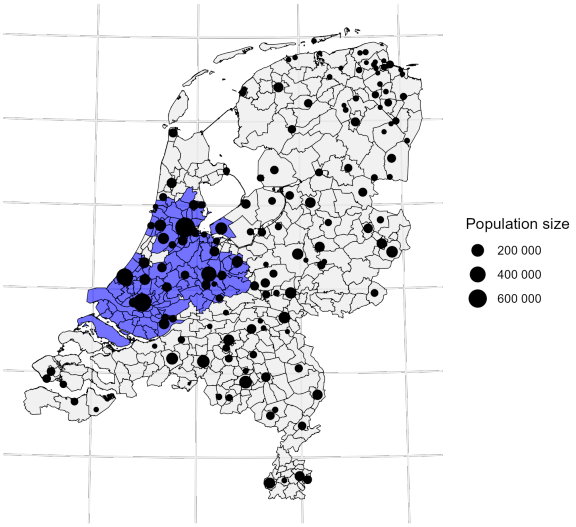

Figure 2: *Division of municipalities into Randstad (blue) and regio (grey) including found places* [10]

3 Methods

Statistical techniques exist to calculate the semantic distance between words, just like places in the real world can be told apart by their physical distance. So, the main aim here is then to express the meaning of a word in a theoretical space and compare these 'contextual' distances between pairs of toponyms regressed on relevant variables. In essence this is a specific kind of discourse analysis involves finding regular patterns in text and then proposing an interpretation of its meaning [9], because the meaning of a toponym takes the form of a vector and is compared to other toponyms. Of course, discourses change throughout time, as shown by Garg and colleagues (2017) [18]. Unfortunately, a decade is generally a rather small period to find such trends, especially if some places do not have a lot of mentions. Therefore I treat the debates as though they took place in one point of time.

Word embeddings offer an interesting quantitative way to conduct discourse analysis by expressing the meaning of words in vectors in order to make them comparable. In practice this entails predicting a target word based on its neighbors or vice versa, the reasoning being that one 'shall know a word by the company it keeps' [16, 22]. Conducting the analysis by using toponym frequencies could have been another way to investigate the same topic, but I consider it less interesting, given that it is only natural that some places appear less in national politics because of their small significance. Researching where they appear if they appear is more telling. Sentiment analysis was also an option, but a hypothetical schism between center and periphery does not need to manifest itself in sentiment per se. It does not mean that politicians talk relatively negatively about the *regio* if people feel their region is neglected by the politicians in The Hague. Politicians might simply converse differently about places in the center than the periphery. Therefore, word embeddings are the method of choice.

These vectors prove themselves useful to investigate how people use words. Dénigot and Burnett (2021) show how the same key concepts, such as freedom (*liberté*) can be applied differently by conservative and progressive parties in the French parliament in debates about gay marriage [13]. Similarly, Garg and colleagues (2021) use word embeddings to track changes in gender and ethnic stereotypes during the 20th and 21st centuries [18]. Although interesting, a comparison of different political strands like conservative versus progressive or administration versus opposition has to be cast aside here, as initial results indicated that the model specification was problematic, which might be due to slicing the data to create an embedding per place

per political color.

Different choices lead to different results. A multitude of choices can be made to create word embeddings. For one thing, word embeddings can be either static or dynamic. Static vectors such as created by `word2vec` have one vector per term, while for example the language model `BERT` creates dynamic representations, where the vector of a word can differ per context [22]. `BERT` has the useful feature that it gives the same word with different meanings distinct vectors, such as for *bergen* (mountains) and the toponym Bergen. Nonetheless, static embeddings are preferred here, because they are more transparent than the black box models to which `BERT` belongs.

There exist different algorithms to create static word embeddings, the most well known ones being `word2vec` and Global Vectors (`GloVe`). `word2vec` is a local model. In short, it takes the target and surrounding words as positive examples and randomly samples other words as negative examples. The prediction task is to predict the target word as accurately as possible, given the positive and negative samples. The predictions serve as the embeddings [22]. `GloVe`, on the other hand, can be characterized as a global model by using the co-occurrence matrix for modelling the embeddings. It models the ratio of probabilities that words co-appear in a corpus [32]. So, $P(\textit{Groningen} \mid \textit{natural gas})/P(\textit{Groningen} \mid \textit{traffic jam})$ and $P(\textit{Utrecht} \mid \textit{natural gas})/P(\textit{Utrecht} \mid \textit{traffic jam})$ would give a higher probability for Groningen, suggesting that natural gas is more strongly associated with Groningen than with Utrecht relative to their associations with traffic jam. Although the embedding models are mathematically similar, they can return quite different results [33]. There exists debate about which model performs better. Pennington and colleagues (2024) show that `GloVe` performs better on English texts [32], while `word2vec` arises as the best model in an Italian language experiment [3]. The results of the `word2vec` model are leading in this thesis given the conflicting results, but I run the regression models twice as a robustness check. Once with the cosine distances resulting from `word2vec` as the dependent variable and once with the cosine distances obtained by `GloVe`. This serves to verify that the embedding model does not drive the results.

Both algorithms allow for setting context windows. Shorter windows (circa two context words) return vectors that are semantically similar, while larger windows (circa ten words) return topically related words [22]. Therefore, a context window of ten is chosen here, since the primary goal of this thesis is to compare topically related toponyms. `word2vec` offers two ways of exploiting the context windows in calculating word embeddings, namely Skip-gram and Continuous Bag of Words (CBOW). Skip-gram predicts the context words given the target term, while CBOW predicts a target word based on the context words. Skip-gram proved to be better at semantic tasks than CBOW, so is therefore the chosen method [27].

Word embeddings possess two noteworthy qualities. First of all, these vectors can exhibit bias, as the input is human made text with all of its biases and stereotypes [22]. This feature might turn out to be undesirable in practical applications, where they would perpetuate these biases. Here, however, the embeddings serve the purpose of measuring differences words surrounding the Randstad and *regio*, so the bias actually forms a useful property [13].

A more pressing concern stems from the variability of the vector representations due to the relatively limited size of the sampled corpus of debates in comparison with the many millions or billions of tokens on which researchers usually train embedding models. The `word2vec` algorithm, for instance, initializes the vectors randomly, which means that separate runs over the same corpus can result in different vector values, especially for smaller sized corpora. Antoniak and Mimno (2018) recommend bootstrapping of the embeddings in order to quantify the variability and mean over bootstrapped samples [1]. Bootstrapping entails calculating a statistic many times on the same sample by sampling with replacement from that sample itself [21].

Here, in line with Antoniak and Mimno's recommendations, I bootstrapped the data twenty five times [1], calculating the cosine distance between focus toponyms multiple times (Algorithm 1). The cosine similarity

quantifies the distance between word vectors in vector space and is defined as the angle between two vectors:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

where *A* and *B* are vectors with i dimensions. The cosine distance is defined as one minus the cosine similarity. It takes on values between zero and two. A larges value indicates a larger distance. Although not a conventional type of distance, it nonetheless adheres to the general definition of 'distance' [6]. Another potential distance metric would be the Euclidean distance, but the cosine distance is more appropriate for high dimensional data, such as text [40]. After that, the average and the standard deviation were taken of these multiple cosine distance estimates per toponym pair.

The mean cosine distances between toponyms are then used as the dependent variable in a gravity model, a model commonly used in economic geography [4]. The slightly rewritten formula of Newton's Law of Gravitation forms the premise of the gravity model.

$$I_{ij} = G\frac{M_i^{\beta_1} M_j^{\beta_2}}{d_{ij}^{\beta_3}} \tag{2}$$

In essence, it states that the interaction I between two objects is correlated with the masses M_i and M_j of the objects and inversely correlated with the distance d_{ij} between two objects [4], meaning that increasing the physical distance between two objects decreases their interaction, while augmenting their mass increases their interaction. G is a constant of not much importance here. With regard to the cosine distances this would mean that the population size serves as the mass and the physical distance in kilometres as the distance. Equation 2 can be rewritten to fit a regression, after defining t_{ij} as $D_{ij} \cdot \exp(r_{ij})$, where r_{ij} is a dummy denoting if `neither`, `one` or `both` of the places in a dyad belong to the Randstad region. The regression formula is then deduced in the following way [39]:

$$\begin{aligned}
I_{ij} &= G\frac{M_i^{\beta_1} M_j^{\beta_2}}{t_{ij}^{\beta_3}} \\
\ln(I_{ij}) &= \ln(G\frac{M_i^{\beta_1} M_j^{\beta_2}}{t_{ij}^{\beta_3}}) \\
&= \ln(G) + \ln(M_i^{\beta_1}) + \ln(M_j^{\beta_2}) - \ln(t_{ij}^{\beta_3}) \\
&= \ln(G) + \ln(M_i^{\beta_1}) + \ln(M_j^{\beta_2}) - \ln((D \cdot \exp(r))_{ij}^{\beta_3}) \\
&= \ln(G) + \ln(M_i^{\beta_1}) + \ln(M_j^{\beta_2}) - \ln(D_{ij}^{\beta_3}) - r_{ij}^{\beta_4},
\end{aligned}$$

Algorithm 1 Bootstrapped cosine distance calculation

```
B ← value                                          ▷ Define B bootstrapped samples.
texts ← vector                                     ▷ Define vector with n texts.
dist ← vector                                      ▷ Define empty vector for distances.
for b = 1, ..., B do
    s ← sample(texts, replace = TRUE)              ▷ Sample with replacement.
    w2v ← word2vec(s)                              ▷ Estimate the model.
    dist ← append(dist, cos_dist(w2v[toponym_i], w2v[toponym_j]))   ▷ Append cosine distance.
end for
mean ← mean(dist)
sd ← sd(dist)
```

resulting in the following regression model:

$$\begin{aligned}\ln(\texttt{cos\_dist}_{ij}) = \beta_0 &+ \beta_1 \cdot \ln(\texttt{pop\_size}_i) + \beta_2 \cdot \ln(\texttt{pop\_size}_j) \\ &+ \beta_3 \cdot \ln(\texttt{phys\_dist}_{ij}) + \beta_4 \cdot \texttt{randstad}_{ij} + \varepsilon_{ij}\end{aligned}$$

$\texttt{randstad}_{ij}$ is a factor variable with the levels `both`, `one` and `neither`, referring to the Randstad membership of a given dyad. The model allows for the inclusion of additional covariates. Table 4 provides an exemplary snippet from the dataset.

Table 4: *Example rows of dataset*

$\texttt{place}_i$	$\texttt{place}_j$	$\texttt{cos\_dist}_{ij}$	$\texttt{phys\_dist}_{ij}$	$\texttt{pop}_i$	$\texttt{pop}_j$	randstad	groningen
Boxtel	Borgsweer	0.67	222.44	29,511	125	neither	one
Utrecht	Alphen_aan_den_Rijn	0.51	31.24	361,742	70,251	both	neither
Hoogeveen	Oss	0.49	124.35	38,754	76,430	neither	neither
Zeist	Rotterdam	0.48	55.81	60,949	598,199	both	neither
Eindhoven	Nijmegen	0.43	49.89	209,620	158,732	neither	neither
Den_Bosch	Alphen_aan_den_Rijn	0.49	65.23	134,520	70,251	one	neither

Note: Dependent variable: $\texttt{cos\_dist}_{ij}$; $\texttt{dist}_{ij}$ is the physical distance in kilometres.

4 Analysis and Results

First, a short discussion of the bootstrapped cosine distances is in order. The left panel of Figure 3 displays the distributions of the cosine distances between place name pairs. The cosine distances are approximately normally distributed. The variability of the estimates of cosine distances per toponym pair is considerable, with two high-density peaks of high and low standard deviations as observed in the right panel of the same Figure. Some pairs have approximately the same cosine distance for almost all bootstraps. The relatively high standard deviations hint that the cosine distances for other toponym pairs vary considerably due to fluctuations in the `word2vec` embeddings between bootstraps. This means that the inclusion of a particular set of utterances can determine the place of a toponym in vector space in these cases, which is not desirable. Although interesting to investigate, the variability of the distances is not the main aim of this thesis, so for now it suffices to keep in mind that some of the distances might be unreliable to a greater or lesser extent.

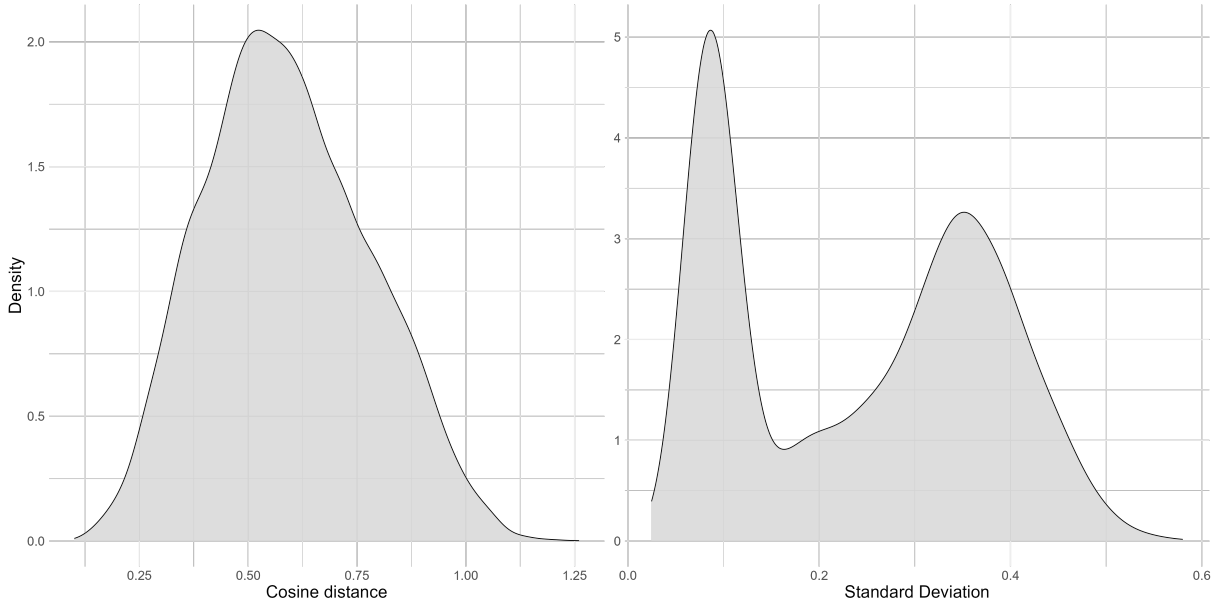

Figure 3: *Distributions of the mean cosine distances per pair and the inter-bootstrap variability*

Figure 4 shows the distribution in two dimensional space of the places found in the debate after conducting Multidimensional Scaling (MDS) on the mean cosine distance matrix. MDS is a dimension reduction technique geared towards proximity data by initializing random coordinates and calculating the distances between those in k-dimensional space (here $k = 2$). The algorithm then compares the synthetic distances to the input cosine distance matrix and adjusts the coordinates iteratively until the distances of the MDS procedure match most closely the input matrix's [5]. MDS thus deduces coordinates from the distance matrix. This is necessary, because the distance matrix does not contain coordinates themselves, prohibiting plotting them in two dimensional space. The resulting plot hints that there is no clear separation between the Randstand and *regio* in the debates, although the places in the Randstad lie relatively close to each other. There exists, however, a noticeable cluster of toponyms from the province of Groningen, with names such as Loppersum and Slochteren, resulting from the prominence of the discussions about the earthquake's in this province in the political arena.

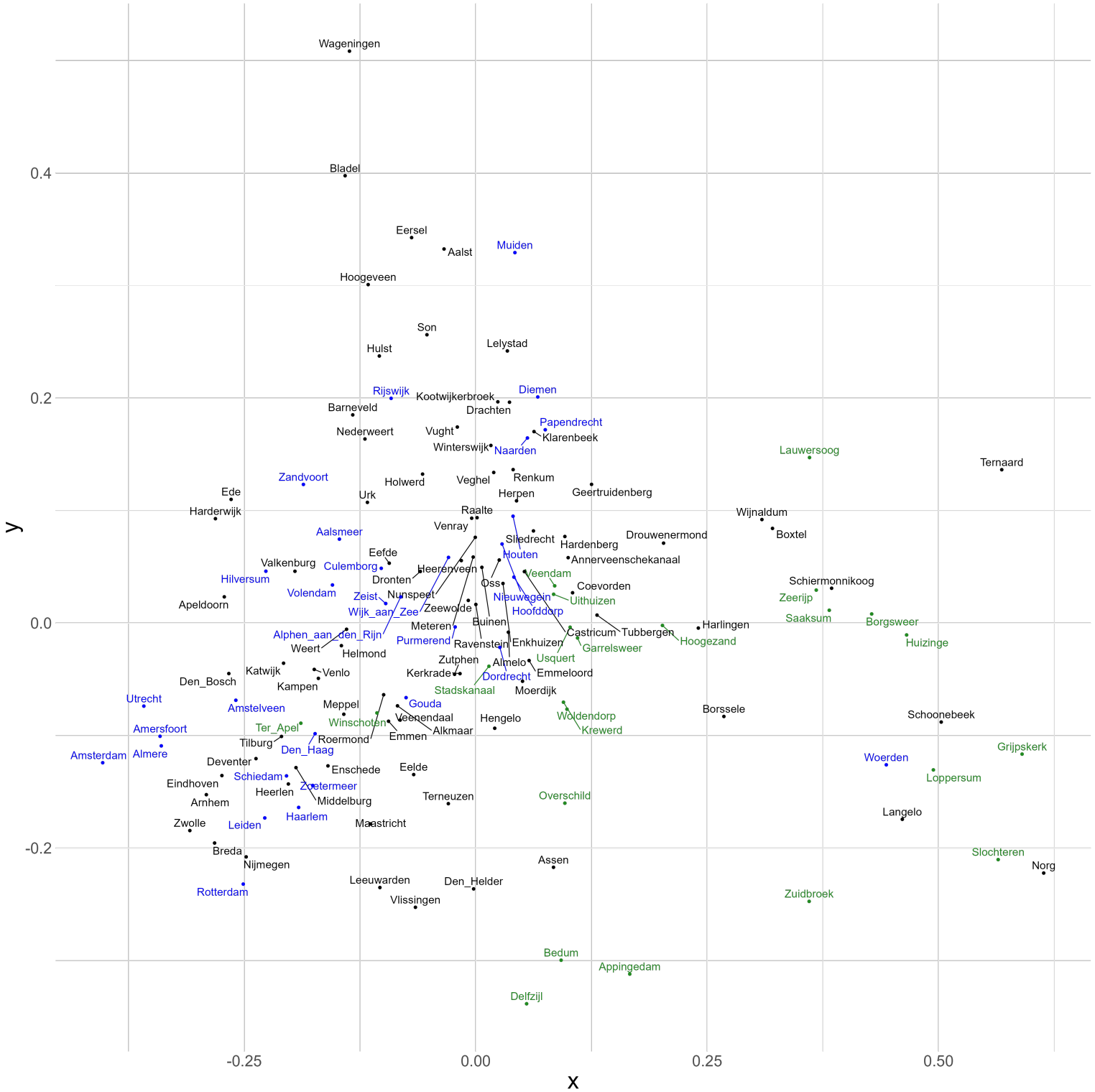

Figure 4: *Places in space after Multidimensional Scaling, Randstad in blue and Groningen in green*

Visualizing the cosine distances highlights some patterns, but it is just an oversimplified representation of vector space and does not tell a lot about why certain words inhabit their particular spots. Table 5 displays therefore the top ten most occurring terms of a selected eight toponyms within the skip-gram windows for illustrative purposes. Amsterdam, Den Haag and Utrecht are surrounded by the same words, like 'city' (*stad*). Also, other toponyms frequently accompany these places. This is different for Slochteren and Huizinge that have words pertaining to natural gas, 'gas extraction' (*gaswinning*) being one example. Wageningen is isolated in Figure 4 due to its university centered vocabulary. Interestingly, the most frequent words around Maastricht refer to the airport named Maastricht Aachen Airport. Another way of looking at which words are characteristic to a toponym is by calculating the Point Mutual Information (PMI) values, although this statistic is less closely related to the embedding algorithm. The interested reader is hence referred to Table A3 in the Appendix.

Moving on to the models, Table 6 shows the results of the gravity models with the logged cosine distance ($\texttt{cos\_dist}_{ij}$) between each possible pair as dependent variable. The first model only includes the variables of interest, resulting in a rather small *R*-squared. Moreover, indications of endogeneity exist, as the distribution of the residuals over the fitted values is heteroscedastic. The second model therefore includes unit level fixed effects, removing the variation due to individual places, hugely improving the model diagnostics (see Figure A1 in the Appendix). A doubling of the physical distance increases the cosine distance by a factor $\exp(0.07 \cdot \ln(2)) = 1.05$, keeping constant all other variables. The effect of the logged populations size variable is of a similar order, since the cosine distance between two places decreases by around five percent if the population of one of them doubles. Most importantly, the cosine distance decreases by 0.24 percent if one of them does not belong to the Randstad, relative to the case when both cities in a pair belong to the center. If neither belongs to the Randstad, the distance decreases even further by 0.44 percent.

Exploration of the embedding space has already uncovered that there is a distinct cluster for Groningen (Figure 4). Groningen places are found in relative proximity to each other at the outskirts. `groningen` was therefore added to the model, indicating if `both`, `one` or `neither` of the places is situated in the province of Groningen. This model is a significant improvement over the second model (FE 1) by decreasing the Residual Sum of

Table 5: *Top ten words (and their counts) according to co-occurrence frequencies*

Amsterdam	Huizinge	Slochteren	Maastricht
Utrecht (792)	aardbeving (349)	gaswinning (39)	Aachen (167)
stad (297)	beving (294)	gas (30)	Airport (117)
Rotterdam (420)	gaswinning (193)	Groninger (23)	Schiphol (106)
Nederland (338)	Groningen (178)	schade (22)	Eelde (106)
Den_Haag (297)	zwaar (103)	gasveld (22)	Groningen (80)
sociaal (228)	augustus (96)	Groningen (20)	luchthaven (63)
wonen (222)	Staatstoezicht (73)	Norg (18)	Rotterdam (47)
bouwen (216)	Mijnen (70)	gasvelden (18)	Lelystad (47)
wethouder (215)	veiligheid (59)	welvaart (16)	Eindhoven (44)
regio (203)	impact (56)	ontdekken (15)	vliegveld (42)
Utrecht	**Wageningen**	**Den Haag**	**Ter Apel**
Amsterdam (792)	Universiteit (872)	Groningen (354)	aanmeldcentrum (40)
Universiteit (318)	universiteit (242)	Amsterdam (297)	situatie (35)
stad (297)	Research (224)	Nederland (283)	Nederland (30)
bouwen (241)	rapport (161)	Rotterdam (277)	buschauffeur (30)
Nederland (203)	Nederland (154)	regio (271)	week (28)
gesprek (195)	University (146)	stad (248)	vervoer (25)
Rijnenburg (167)	kennis (100)	bepalen (225)	burgemeester (24)
regio (163)	blijken (98)	politiek (213)	openbaar (23)
Rotterdam (156)	Economic (90)	lokaal (180)	asielzoekerscentrum (23)
Kamer (134)	wereld (80)	Kamer (174)	asielzoeker (21)

Squares by 10.92 ($F(283.03)$, $p < .001$). The effect size of the logged physical distance variable does not change much in contrast to the population variables, which even change sign. A doubling of the population size in one of the toponyms of the pair goes hand in hand with an increase of the cosine distance by a factor $\exp(0.23 \cdot \ln(2)) = 1.17$. The effect if one or neither place belongs to the Randstad compared to when they have both Randstad membership is strongly reduced. The cosine distance is reduced by fourteen or eight percent if none or one of them is in the Randstad region, respectively. Much of the variation in `randstad` is absorbed by `groningen`. If one or both of the places in a pair lies in the province of Groningen, the cosine distance increases by 2.08 or 1.10 percent, respectively. All coefficients of the fixed effects models are highly significant.

See Table A4 in the Appendix for the results of the cosine distances based on the `GloVe` embeddings. Some differences exist as a result of a different estimation method of the dependent variable $\texttt{cos\_dist}_{ij}$. The sign and magnitude are to a great extend comparable, however, suggesting that the results that the embedding models provide are moderately robust.

Inspecting the fixed effects themselves also supports their inclusion, because there is no discernible pattern

Table 6: *Gravity models:* `word2vec`

	Model		
	OLS 1	FE 1	FE 2
$\ln(\texttt{phys\_dist}_{ij})$	0.09***	0.07***	0.05***
	(0.01)	(0.002)	(0.003)
$\ln(\texttt{pop}_i)$	−0.02***	−0.07***	0.23***
	(0.002)	(0.004)	(0.01)
$\ln(\texttt{pop}_j)$	−0.01***	−0.07***	0.23***
	(0.002)	(0.004)	(0.01)
`randstad` (both)	(reference)	(reference)	(reference)
`randstad` (neither)	−0.10***	−0.44***	−0.14***
	(0.02)	(0.03)	(0.03)
`randstad` (one)	−0.08***	−0.24***	−0.08***
	(0.02)	(0.02)	(0.02)
`groningen` (neither)			(reference)
`groningen` (both)			2.08***
			(0.10)
`groningen` (one)			1.10***
			(0.05)
Constant	−0.57***	0.91***	−5.97***
	(0.04)	(0.10)	(0.35)
Observations	9,870	9,870	9,870
R^2	0.05	0.83	0.84
Adjusted R^2	0.05	0.83	0.84
Residual Std. Error	0.34 (df = 9864)	0.14 (df = 9727)	0.14 (df = 9725)
F Statistic	95.26*** (df = 5; 9864)	341.50*** (df = 142; 9727)	358.92*** (df = 144; 9725)
Note:			*p<0.1; **p<0.05; ***p<0.01

in the sizes of the fixed effects (Figure 5). After taking into account the physical distance, population sizes and Randstad membership, there is still a lot of variation left. This variance can be explained by the individual places. For example, the fixed effects coefficient for Nieuwegein in the Utrecht province is -0.85, meaning that the cosine distance decreases by 0.85 percent when Nieuwegein is present in the pair. On the other hand, the cosine distance increases by 0.27 percent when Borssele is included. The reason for the large coefficient size of the latter could be the presence of a nuclear power plant.

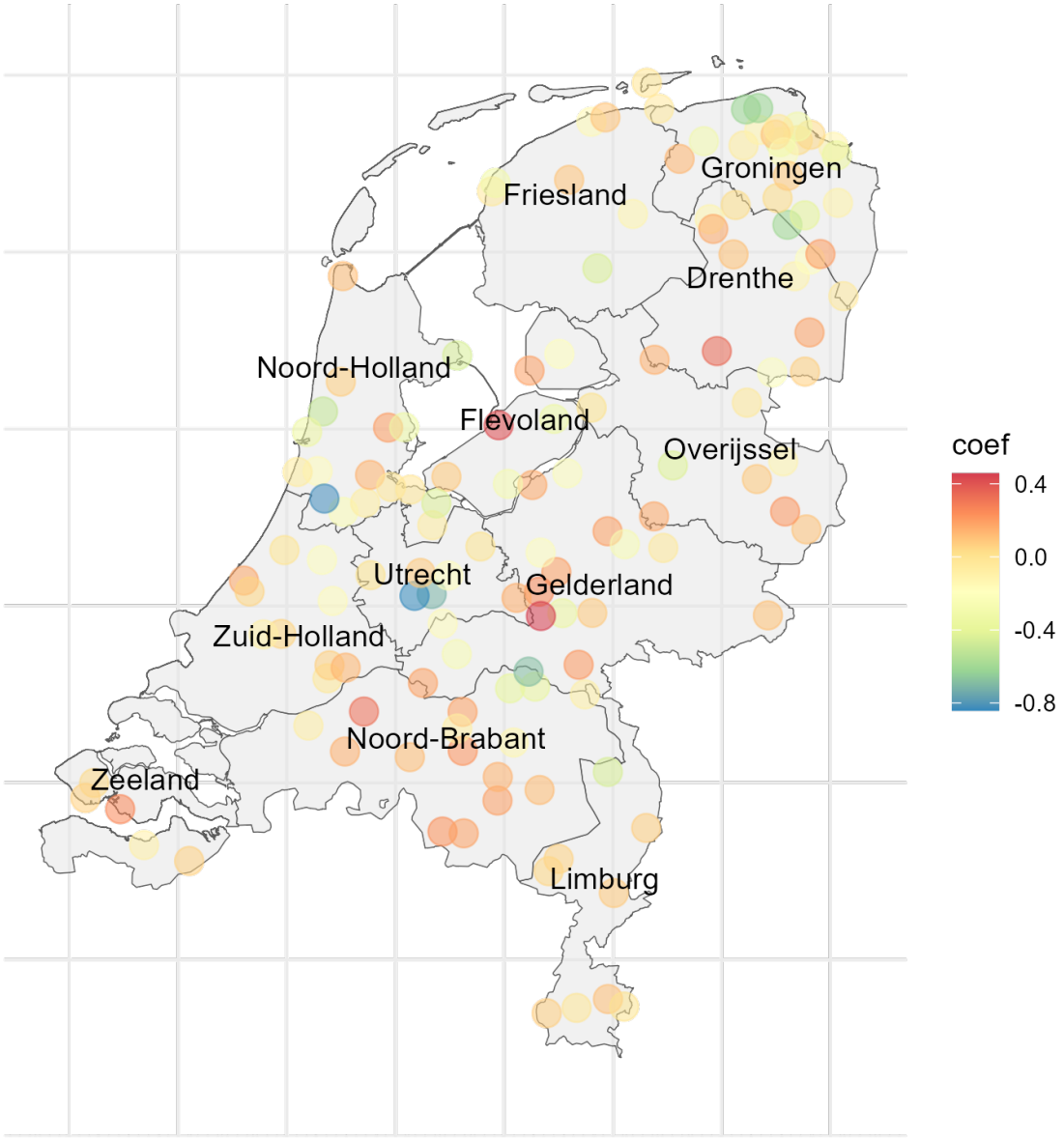

Figure 5: *Fixed effects coefficients obtained by FE 1 (Table 6), a higher fixed effect for a place in a pair indicates a larger cosine distance. [17]*

5 Discussion

These results are only one piece of the puzzle. As stated in the Introduction, residents of the *regio* feel that the Randstad receives a special treatment from the politicians. With the exception of Groningen, this thesis provides evidence against a differential treatment of the center and periphery, but other factors could be at play. For one thing, the fact that places from the Randstad and *regio* are not easily distinguishable might well contribute to the discontent. Apparently, politicians converse about places from either part of The Netherlands in relatively the same way, so issues specific to the *regio* may not receive the attention in national politics that locals might desire. Thus, similar treatment of Randstad and *regio* is not always equivalent good representation.

Even similar contexts do not translate directly into similar treatment. Consider for example rhetoric. Politicians might use a different rhetoric in the same context differently for two places based on if they are part of the Randstad. Previous research shows that a politician's rhetoric on a topic can change abruptly after an impactful event [2]. These subtle ways of speech are not capturable by word embeddings. Simply stated, static embeddings especially are a relatively rudimentary tool that leaves space for more fine grained methodologies. Subsequent research could delve into an more in depth analysis of differences in rhetoric surrounding toponyms.

Another related challenge is differences between politicians, parties, ideologies, and cabinet and opposition. Politics is a market place of ideas. Unfortunately, this thesis took the debates as a coherent whole due to practical limitations (see Methods). Further research can offer insight in the extent to which different political colors speak differently about the Randstad and *regio*. One might expect, for example, that a party with a large regional constituency such as the Farmer–Citizen Movement (*BoerBurgerBeweging*) has a different attitude towards the rural areas of The Netherlands than the liberal and cosmopolitan Democraten 66.

Methodologically there are also some suggestions for follow up research. For one thing, this analysis does not allow to differentiate between Groningen en Utrecht, toponyms that are the names of the province and their capitals. This issue generalizes to other cases where a word has more functions than just toponym, because the static embedding models treat every unique token as one unique word. I solved this by excluding ambiguous place names (Table A1), but this is certainly not the perfect solution. Moreover, the interpretation of the word embeddings is not straightforward, as the vectors are a numerical abstraction of the meaning of a word. Qualitative textual research could fruitfully contribute to contextualizing the findings of the presented results [25].

Finally, the political field changes by the day. Issues at play in The Netherlands ten years ago differ from those in the present. Still, this thesis treated a whole decade as if it was one point in time as to ensure the validity of the embeddings by pooling as much examples of a specific toponym as possible. Further research is encouraged to investigate if and how the context around toponyms changes throughout time. Especially Groningen could be an interesting case in this regard, as it is known that the political attitude towards gas extraction in that province changed quite drastically in a short time [37]. My analysis also showed the prevalence of this topic in Dutch politics.

6 Conclusion

The objective of the thesis was analyzing the extent to which localities from the Randstad and *regio* appear in different contexts in parliamentary debates of the Dutch House of Representatives using word embeddings. Taken together, the image arises that there is not as much a separation of contextual embeddings between Randstad and *regio*, as there exists between Groningen and the rest. Although interesting, it is not surprising, given the many debates surrounding the closure of the Groningen gas fields and compensation of the locals. Nonetheless, the contexts in which places appear are slightly more equivalent when one or both of the places are from the periphery. This results suggests that the contexts in which localities from the Randstad appear diverge more from each other than those from the *regio*. The effect is smaller when one of the places lies in the Randstad and the other in the *regio*. Nevertheless, the effects are quite minor.

Moreover, the effects of these variables are overshadowed by the physical distance between a pair and their population sizes, which have a far greater effect on the contextual embedding. Larger physical distances unsurprisingly go hand in hand with larger cosine distances. Places closer to each other have generally the same issues. The positive effect for the population variables signal that larger places inhabit their own characteristic spot in vector space. That is, increases in population sizes augments the differences in contexts where places appear. So, although discernible, the membership to either the Randstad or *regio* does not determine the context in which a place is named too much.

Judging by the stark increase in explained variance, fixed effects for each place play a more significant role in modeling the contexts in which places are embedded. This finding implies that a lot of factors are at play in determining the context of a particular place. Factors that are independent of the segmentation into Randstad and *regio* and are characteristic to that place.

This research also raises new questions about the ideological and rhetorical meanings of the pleas. Shedding light on the deeper meanings of the arguments in debates by using qualitative research is recommended, since it provides insight into nuances of the debates that are underexposed by a quantitative method. Nevertheless, with regard to this thesis one can speak of a phantom gap: the gap between Randstad and *regio* in society seems larger than it turns out to be in political debates.

References

1. Antoniak, M. & Mimno, D. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics* **6,** 107–119 (2018).
2. Bastardoz, N., Jacquart, P. & Antonakis, J. Effect of crises on charisma signaling: A regression discontinuity design. *The Leadership Quarterly,* 101590 (2022).
3. Berardi, G., Esuli, A. & Marcheggiani, D. *Word Embeddings Go to Italy: A Comparison of Models and Training Datasets* in *Italian Information Retrieval Workshop* (2015).
4. Broekel, T., Balland, P., Burger, M. & Van Oort, F. Modeling knowledge networks in economic geography: a discussion of four methods. *The annals of regional science* **53,** 423–452 (2014).
5. Buja, A. *et al.* Data visualization with multidimensional scaling. *Journal of computational and graphical statistics* **17,** 444–472 (2008).
6. Burago, D., Burago, Y., Ivanov, S., *et al. A course in metric geometry* (American Mathematical Society Providence, 2001).
7. Burger, M., Van Der Knaap, B. & Wall, R. Polycentricity and the multiplexity of urban networks. *European Planning Studies* **22,** 816–840 (2014).
8. Camacho-Collados, J. & Pilehvar, M. On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. *arXiv* (2017).
9. Cameron, D. *Working with spoken discourse* (Sage, 2001).
10. Centraal Bureau voor de Statistiek & Kadaster. *Wijk- en buurtkaart 2014* Accessed: 15-06-2024. 2016. `https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/wijk-en-buurtkaart-2014`.
11. Cuba, L. & Hummon, D. A place to call home: Identification with dwelling, community, and region. *Sociological quarterly* **34,** 111–131 (1993).
12. De Lange, S., Van der Brug, W. & Harteveld, E. Regional resentment in the Netherlands: A rural or peripheral phenomenon? *Regional Studies* **57,** 403–415 (2023).
13. Dénigot, Q. & Burnett, H. *Using word embeddings to uncover discourses* in *Proceedings of the Society for Computation in Linguistics 2021* (2021), 298–312.
14. Den Otter, B. *4pp* `https://github.com/bobdenotter/4pp`. Accessed: May 16, 2024. 2019.
15. Duijvendak, M. De perfecte periferie: Turf, gas en Oost-Groningen. *Groniek,* 272–285 (2021).
16. Firth, J. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis,* 10–32 (1957).
17. GADM. *GADM Maps and Data* Accessed: 15-06-2024. 2023. `https://gadm.org/`.
18. Garg, N., Schiebinger, L., Jurafsky, D. & Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* **115,** E3635–E3644 (2018).
19. *GeoNames* `http://www.geonames.org/`. Accessed: 10-05-2024. n.d.
20. Huijsmans, T., Harteveld, E., Van der Brug, W. & Lancee, B. Are cities ever more cosmopolitan? Studying trends in urban-rural divergence of cultural attitudes. *Political Geography* **86,** 102353 (2021).
21. James, G., Witten, D., Hastie, T., Tibshirani, R., *et al. An introduction to statistical learning* second (Springer, 2013).
22. Jurafsky, D. & Martin, J. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* third (2023).

23. Kenny, M. & Luca, D. The urban-rural polarisation of political disenchantment: an investigation of social and political attitudes in 30 European countries. *Cambridge Journal of Regions, Economy and Society* **14,** 565–582 (2021).

24. Land, R. *Phantom Gap* `https://gitfront.io/r/calinuslongus/LidHJt3RRVvd/phantom-gap/`. 2024.

25. Lange, M. *Comparative-historical methods* (SAGE Publications Ltd, 2012).

26. Meijers, E. & Peris, A. Using toponym co-occurrences to measure relationships between places: Review, application and evaluation. *International Journal of Urban Sciences* **23,** 246–268 (2019).

27. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. *arXiv* (2013).

28. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* **26** (2013).

29. Munis, B. Us over here versus them over there... literally: Measuring place resentment in American politics. *Political Behavior* **44,** 1057–1078 (2022).

30. Nijman, B. *The Effect of Space-Language Bias on Toponym Co-occurrence Derived Networks* MA thesis (Utrecht University, 2022).

31. *Open Data Portaal* `https://opendata.tweedekamer.nl/`. Accessed: 26-04-2024. Tweede Kamer der Staten-Generaal, n.d.

32. Pennington, J., Socher, R. & Manning, C. *Glove: Global vectors for word representation* in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (2014), 1532–1543.

33. Rodriguez, P. & Spirling, A. Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics* **84,** 101–115 (2022).

34. Silge, J. *Word vectors with tidy data principles* `https://juliasilge.com/blog/tidy-word-vectors/`. Accessed: 06-06-2024. 2017.

35. Straka, M. & Straková, J. *Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe* in *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies* (2017), 88–99.

36. Van der Ploeg, J. Farmers' upheaval, climate crisis and populism. *The Journal of Peasant Studies* **47,** 589–605 (2020).

37. Verdoes, A. & Boin, A. in *Understanding the Creeping Crisis* (eds Boin, A., Ekengren, M. & Rhinard, M.) 149–164 (Springer International Publishing, 2021).

38. Vulpen, B. v., Bock, B. & Vn den Berg, C. Does regional decline trigger discontent? Unravelling regional development in the Netherlands. *Regional studies* **58,** 938–951 (2024).

39. Wölwer, A., Breßlein, M. & Burgard, J. Gravity models in R. *Austrian Journal of Statistics* **47,** 16–35 (2018).

40. Zhu, Y. *et al.* Machine: the new art connoisseur. *arXiv* (2019).

Appendix

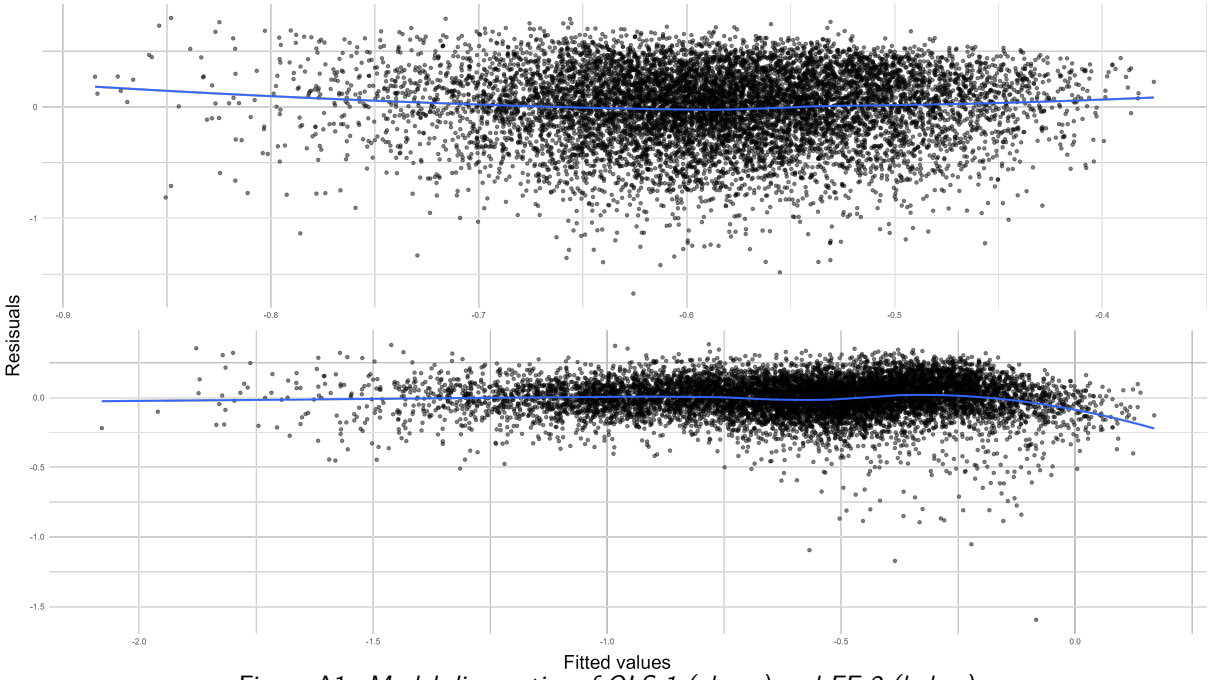

Figure A1: *Model diagnostics of OLS 1 (above) and FE 2 (below)*

Table A1: *Places excluded from analysis*

	Provincie	Place	Pop. size	Reason
1	Drenthe	Amen	80	interjection
2	Drenthe	Anderen	245	pronoun
3	Drenthe	De Wijk	2,495	noun
4	Drenthe	Een	535	numeral and article
5	Drenthe	Linde	110	noun and surname
6	Drenthe	Loon	265	noun
7	Drenthe	Oranje	395	color and surname
8	Drenthe	Vries	3,915	surname
9	Friesland	Boer	45	noun
10	Gelderland	Dieren	14,842	noun
11	Gelderland	Hengelo	102,773	smaller namesake
12	Gelderland	Rijswijk	1,862	surname
13	Gelderland	Stroe	695	surname
14	Gelderland	Voorst	1,505	surname
15	Gelderland	Zetten	2,985	verb and noun
16	Groningen	Den Ham	135	surname
17	Groningen	Groningen	233,218	province name
18	Limburg	Beek	8,415	noun and surname
19	Limburg	Heel	3,980	adverb and adjective
20	Limburg	Heide	455	noun
21	Limburg	Stein	11,290	name
22	Noord-Brabant	Alphen	3,160	surname
23	Noord-Brabant	Best	29,074	adjective
24	Noord-Brabant	Esch	1,820	surname
25	Noord-Brabant	Handel	1,310	noun
26	Noord-Brabant	Hoeven	4,895	verb
27	Noord-Brabant	Linden	225	surname
28	Noord-Brabant	Made	11,795	noun and English adjective
29	Noord-Brabant	Nispen	1,245	surname
30	Noord-Brabant	Schijf	770	noun
31	Noord-Brabant	Veldhoven	5,340	surname
32	Noord-Brabant	Zeeland	4,750	province name
33	Noord-Holland	De Waal	225	surname and river
34	Noord-Holland	Hem	1,035	pronoun
35	Noord-Holland	Huizen	41,273	noun
36	Overijssel	Goor	11,460	adjective
37	Overijssel	Nederland	10	country name
38	Overijssel	Zwolle	65	smaller namesake
39	Utrecht	Doorn	10,330	surname
40	Utrecht	Hoogland	10,587	surname
41	Zeeland	Waarde	1,080	noun
42	Zuid-Holland	Kaag	465	surname
43	Zuid-Holland	Valkenburg	3,925	smaller namesake

Inspiration for this table and Table A2: [30]

Table A2: *Places included in analysis, continued*

	Provincie	Place	Pop. size
1	Drenthe	Annerveenschekanaal	415
2	Drenthe	Assen	62,237
3	Drenthe	Buinen	750
4	Drenthe	Coevorden	14,600
5	Drenthe	Drouwenermond	530
6	Drenthe	Eelde	6,450
7	Drenthe	Emmen	57,010
8	Drenthe	Hoogeveen	38,754
9	Drenthe	Langelo	195
10	Drenthe	Meppel	30,697
11	Drenthe	Norg	3,290
12	Drenthe	Schoonebeek	3,845
13	Flevoland	Almere	176,432
14	Flevoland	Dronten	28,073
15	Flevoland	Emmeloord	46,409
16	Flevoland	Lelystad	70,741
17	Flevoland	Urk	17,345
18	Flevoland	Zeewolde	19,022
19	Friesland	Drachten	44,537
20	Friesland	Harlingen	16,119
21	Friesland	Heerenveen	43,094
22	Friesland	Holwerd	1,435
23	Friesland	Leeuwarden	91,424
24	Friesland	Schiermonnikoog	936
25	Friesland	Ternaard	1,165
26	Friesland	Wijnaldum	330
27	Gelderland	Aalst	1,580
28	Gelderland	Apeldoorn	136,670
29	Gelderland	Arnhem	141,674
30	Gelderland	Barneveld	35,095
31	Gelderland	Culemborg	26,826
32	Gelderland	Ede	67,670
33	Gelderland	Eefde	3,285
34	Gelderland	Harderwijk	40,516
35	Gelderland	Klarenbeek	565
36	Gelderland	Kootwijkerbroek	1,690
37	Gelderland	Meteren	3,855
38	Gelderland	Nijmegen	158,732
39	Gelderland	Nunspeet	19,496
40	Gelderland	Renkum	9,421
41	Gelderland	Wageningen	35,433
42	Gelderland	Winterswijk	29,623
43	Gelderland	Zutphen	36,188
44	Groningen	Appingedam	12,364
45	Groningen	Bedum	7,765
46	Groningen	Borgsweer	125
47	Groningen	Delfzijl	28,649
48	Groningen	Garrelsweer	495
49	Groningen	Grijpskerk	2,520
50	Groningen	Hoogezand	21,480
51	Groningen	Huizinge	105
52	Groningen	Krewerd	90
53	Groningen	Lauwersoog	130
54	Groningen	Loppersum	2,365
55	Groningen	Overschild	245
56	Groningen	Saaksum	90
57	Groningen	Slochteren	2,055
58	Groningen	Stadskanaal	20,000
59	Groningen	Ter Apel	3,880
60	Groningen	Uithuizen	4,935
61	Groningen	Usquert	1,275
62	Groningen	Veendam	28,155
63	Groningen	Winschoten	18,506
64	Groningen	Woldendorp	920
65	Groningen	Zeerijp	450
66	Groningen	Zuidbroek	2,255
67	Limburg	Heerlen	93,084
68	Limburg	Kerkrade	49,777
69	Limburg	Maastricht	122,378
70	Limburg	Nederweert	7,205
71	Limburg	Roermond	44,975
72	Limburg	Valkenburg	3,400
73	Limburg	Venlo	101,603
74	Limburg	Venray	39,047
75	Limburg	Weert	48,662
76	Noord-Brabant	Den Bosch ('s-Hertogenbosch)	134,520
77	Noord-Brabant	Bladel	9,785
78	Noord-Brabant	Boxtel	29,511
79	Noord-Brabant	Breda	167,673
80	Noord-Brabant	Eersel	18,185
81	Noord-Brabant	Eindhoven	209,620
82	Noord-Brabant	Geertruidenberg	20,941
83	Noord-Brabant	Helmond	74,740
84	Noord-Brabant	Herpen	2,360
85	Noord-Brabant	Katwijk	280
86	Noord-Brabant	Moerdijk	1,030
87	Noord-Brabant	Oss	76,430
88	Noord-Brabant	Ravenstein	1,465
89	Noord-Brabant	Son	4,930
90	Noord-Brabant	Tilburg	199,613
91	Noord-Brabant	Veghel	25,352
92	Noord-Brabant	Vught	25,043
93	Noord-Holland	Aalsmeer	22,991
94	Noord-Holland	Alkmaar	94,853
95	Noord-Holland	Amstelveen	79,639
96	Noord-Holland	Amsterdam	741,636
97	Noord-Holland	Castricum	35,256
98	Noord-Holland	Den Helder	59,569
99	Noord-Holland	Diemen	24,361
100	Noord-Holland	Enkhuizen	17,365
101	Noord-Holland	Haarlem	147,590
102	Noord-Holland	Hilversum	83,640
103	Noord-Holland	Hoofddorp	132,734
104	Noord-Holland	Muiden	6,223
105	Noord-Holland	Naarden	17,115
106	Noord-Holland	Purmerend	80,117
107	Noord-Holland	Volendam	22,000
108	Noord-Holland	Wijk aan Zee	2,095
109	Noord-Holland	Zandvoort	16,868
110	Overijssel	Almelo	72,725
111	Overijssel	Deventer	97,331
112	Overijssel	Enschede	153,655
113	Overijssel	Hardenberg	57,909
114	Overijssel	Hengelo	102,773
115	Overijssel	Kampen	48,980
116	Overijssel	Raalte	20,015
117	Overijssel	Tubbergen	20,334
118	Overijssel	Zwolle	111,805
119	Utrecht	Amersfoort	139,914
120	Utrecht	Houten	1,335
121	Utrecht	Nieuwegein	61,489
122	Utrecht	Utrecht	361,742
123	Utrecht	Veenendaal	61,271
124	Utrecht	Woerden	48,431
125	Utrecht	Zeist	60,949
126	Zeeland	Borssele	22,285
127	Zeeland	Hulst	1,405
128	Zeeland	Middelburg	46,485
129	Zeeland	Terneuzen	27,930
130	Zeeland	Vlissingen	45,273
131	Zuid-Holland	Alphen aan den Rijn	70,251
132	Zuid-Holland	Den Haag	474,292
133	Zuid-Holland	Dordrecht	119,260
134	Zuid-Holland	Gouda	71,952
135	Zuid-Holland	Leiden	119,713
136	Zuid-Holland	Papendrecht	31,621
137	Zuid-Holland	Rijswijk	47,299
138	Zuid-Holland	Rotterdam	598,199
139	Zuid-Holland	Schiedam	75,438
140	Zuid-Holland	Sliedrecht	23,854
141	Zuid-Holland	Wassenaar	25,353
142	Zuid-Holland	Zoetermeer	115,845

Table A3: *Top ten words according to PMI values*

Amsterdam	Huizinge	Slochteren	Maastricht
Hartje	Sociaalpsychologisch	afgevenen	A2passage
forenzen	recordgaswinning	deputeeren	eurotop
Beach	contourenbeeld	Kameronderzoek	ENCIgroeve
Castle	berge	Vpbplichtig	Ceramique
Bouwrecord	energiecluster	knipklei	RijkswaterstaatZuid
bouwrecords	opgevoeren	gaswinner	stedentrip
Declarations	hulpeloosheid	Usquert	Fontys
Drone	langstdurende	progressie	kalksteengroeve
Dordrecht	toedekken	roerig	Avenue
IJburg	OVVrapporten	weerzin	area
Utrecht	**Wageningen**	**Den Haag**	**Ter Apel**
oudSPcollega	UR	reisduur	nachtopvang
herindelingsprocedure	Landbouwuniversiteit	Hanzelijn	Velema
PacMan	kunstblader	nietSGPmanier	noodverzoek
sciencefictionachtig	Brinkschool	voetbalgebied	aankomsthal
huurcontrole	alumni	zwembad	socialewerkplaatsen
eigenaargerelateerd	Environmental	EnergieRijk	Veiligelanders
GreenCab	naten	Sacramentskerk	Acuut
stedendriehoek	Bioveterinary	forenzen	aanmeldproces
Werkende	oudbestuursvoorzitter	afscheiding	aanmeldcentrum
ongelijkvloers	Velthof	patiënt	radiostilte

Note: The top ten pointwise mutual information (PMI) values of eight purposefully selected places given their skip-gram windows are calculated for illustrative purposes. The PMI indicates 'which words occurred more often than expected based on how often they occurred on their own'. This translates to the following formula: $\text{PMI}(a, b) = \log \frac{P(a,b)}{P(a)P(b)}$, where $P(a)$ and $P(b)$ is the unigram probabilities of words a and b and $P(a)P(b)$ as their joint probability [34]. Most places are surrounded by words characteristic to them. Top terms for Slochteren and Huizinge in Groningen often pertain to the natural gas extraction, such as 'record gas production' (*recordgaswinning*). The lists for Amsterdam and Utrecht include terms concerning housing: 'building records' (*bouwrecords*). Terms in Ter Apel's list concern migration, for example 'registration center' (*aanmeldcentrum*). All values are of a similar order, ranging from 2.632 to 4.181. *naten* is a nonesense term created by the lemmatization model.

Table A4: *Gravity models:* `GloVe`

	Model		
	OLS 1	FE 1	FE 2
$\ln(\texttt{phys\_dist}_{ij})$	0.04^{***}	0.01^{***}	0.01^{***}
	(0.004)	(0.001)	(0.001)
$\ln(\texttt{pop}_i)$	0.01^{***}	-0.03^{***}	0.22^{***}
	(0.001)	(0.002)	(0.01)
$\ln(\texttt{pop}_j)$	0.01^{***}	-0.03^{***}	0.22^{***}
	(0.001)	(0.002)	(0.01)
`randstad` (both)	(reference)	(reference)	(reference)
`randstad` (neither)	0.003	-0.30^{***}	-0.06^{***}
	(0.02)	(0.01)	(0.01)
`randstad` (one)	-0.02	-0.15^{***}	-0.03^{***}
	(0.02)	(0.01)	(0.01)
`groningen` (neither)			(reference)
`groningen` (both)			1.78^{***}
			(0.04)
`groningen` (one)			0.89^{***}
			(0.02)
Constant	-0.68^{***}	0.51^{***}	-5.21^{***}
	(0.03)	(0.05)	(0.15)
Observations	9,870	9,870	9,870
R^2	0.02	0.95	0.96
Adjusted R^2	0.02	0.95	0.96
Residual Std. Error	0.29 (df = 9864)	0.06 (df = 9727)	0.06 (df = 9725)
F Statistic	34.37^{***} (df = 5; 9864)	$1{,}360.15^{***}$ (df = 142; 9727)	$1{,}575.88^{***}$ (df = 144; 9725)
Note:			*p<0.1; **p<0.05; ***p<0.01