Do more interactions mean stronger relationships? Measuring city relationship strength beyond total counts

Wang Tongjing[*]

Department of Human Geography and Spatial Planning, Faculty of Geosciences, Utrecht University, Utrecht, The Netherlands
Address: Princetonlaan 8a, 3584 CB Utrecht, The Netherlands.
Email: t.wang1@uu.nl

*Corresponding author

Abstract:

As two places can interact without specific relationships, measuring city relationship strength based solely on the total amount of interactions fails to fully capture the essence of 'relatedness.' Drawing on probability theory, this study proposes that a strong relationship between two cities is characterized by interaction levels beyond expectation. This expectation has two dimensions: effects and certainty. Building on this premise, the study develops a classification framework to delineate various types of relationships. This framework is empirically tested by examining the co-appearance of placenames for 100 European cities in Wikipedia articles. The findings indicate that while high co-occurrence frequencies are common among European capitals, these do not necessarily reflect stronger-than-expected relationships. Additionally, the results found significant variations across different relationship metrics, highlighting the need for a more comprehensive analysis in city network studies.

Keywords: city relationship, mutual information, collocation analysis, probability

1. Introduction

As Tobler's first law states (1970, p236), "*everything is related to everything else, but near things are more related than distant things*"—cities are embedded within broader urban systems and engage in various degrees and types of relationships (Castells, 1996; Neal, 2013). These intercity relationships foster synergies that drive urban and regional socio-economic development (Gordon and McCann, 2000; Meijers et al., 2018; Taylor and Derudder, 2022; Bathelt and Storper, 2023). Understanding how cities are related facilitates developing targeted policies that enhance city collaborations and optimize resource allocations.

There are various proxies used to represent city relationship strength, including the number of people or trains moving between them, jointly authored scientific publications, and shared patent registrations (e.g. Lai and Pan, 2020; Zhu et al., 2022; Cao et al., 2022; Castaldi and Drivas, 2023). Additionally, there is also an emerging method, toponym co-occurrence, which quantifies city relationship strength through the co-occurrence of city names appearing together in texts (e.g. Liu et al., 2014; Salvini and Fabrikant, 2016; Meijers and Peris, 2019; Wu et al., 2019). This method assumes that a higher occurrence of city names mentioned together indicates a stronger relationship. A common feature of all these measurement approaches is that they are total amount-based; that is, they imply that a higher amount of interactions suggests stronger relationships.

Yet, as Neal et al., (2022, p623) noted, "*larger weights do not necessarily indicate stronger or more important connections.*"—simply counting the total volume of interaction may not fully capture the relationship strength between two cities. For instance, if two large cities and two small cities each have one million people moving between them and the distances between both pairs are the same, the relationship strength would then be considered the same based on the total amount of people movement alone. However, as the classical gravity model of urban interaction implies (Zipf, 1941), large cities and shorter distances typically lead to more interactions. This would mean that relative to their size, the two small cities actually demonstrate a stronger tendency for people movement compared to the two large cities. This suggests that using the total amount alone as a proxy of relationship strength fails to reflect the genuine association level between the two cities.

To address this issue, a common strategy is to measure the relative strength of relationships. A common approach is to account for city size and distance through gravity model estimation (Tongjing et al., 2023). However, this approach introduces its own bias. One notable issue is the potential bias introduced when accounting for smaller cities. Due to their generally fewer interactions, even sporadic interactions can appear disproportionately significant, possibly exaggerating the perceived strength of their relationships (Neal, 2022; Tongjign et al., 2024a). Consequently, there remains a challenge in ensuring that the strong connections identified are not merely the result of coincidence.

The dilemma of association measurement extends beyond intercity connections and is equally pertinent in linguistic studies measuring word associations—Common words appear more frequently in texts, which can lead to the frequent co-occurrence of two commonly used words. Relying solely on co-occurrence frequency to assess relationships can overshadow the strength of associations between less common words (Baker et al., 2008; 2013).

In response to this challenge, linguists have identified that association measurement has two dimensions (Evert, 2004): effect size and certainty. Effect size suggests that a strong relationship should be 'unusually' higher than what is typically expected, while certainty implies that this unusual frequency of interaction is not merely coincidental. These two dimensions count not just the high volume of interactions but also the overall sample size, the frequency of each word's appearance, and the statistical distribution of occurrences of all associated words of each word.

Leveraging this duo-dimensional perspective from linguistics, this paper proposes shifting from a traditional total amount-based city relationship measurement to a more comprehensive multidimensional approach by introducing a relationship classification framework. This framework is designed from the perspective of statistics and probability theory to categorize different types of city relationships more distinctly. It aims to provide insights that can be instrumental for city relationship analysis.

This paper will empirically demonstrate this approach using toponym co-occurrence-based city relationships as an example, evaluating such relationships among 100 European cities. The study will measure, compare, and classify these city relationships across multiple dimensions, including the total amount of placename co-occurrence, effect size, statistical confidence, and gravity model estimation. Through this methodology, the study seeks to illustrate that no single metric can fully capture the multidimensions of city relationships.

This paper is an initial step and, given its exploratory nature, it uses Wikipedia as a starting point for its database. As the largest and most visited encyclopedic resource on the Web, Wikipedia offers a comprehensive and well-structured knowledge base that is suitable for examining city relationships in a general context (Stvilia et al., 2008; Lewoniewski, 2022; Shenoy, 2022). While the case study in this research focuses on placename co-appearance in texts, the methodology is equally applicable to other contexts where city relationships are manifested. These include joint scientific collaborations (Dong et al., 2020), patent-based knowledge networks (Balland et al., 2022), trade networks (Straka et al., 2017), corporate location-based interlocking models (Taylor and Derudder, 2016), and traffic volume-based transportation networks (Guo et al., 2020). This adaptability allows for a broader application of the methodology to various city network analyses.

The structure of this paper is organized as follows: Section 2 evaluates common city relationship measurement approaches, and then introduces concepts from linguistics for measuring word associations, which can be adapted to study city relationships. Section 3 introduces two distinct relationship measurement approaches and proposes a classification framework for city relationships. Section 4 conducts a comparative analysis of various metrics and presents the findings. Finally, Section 5 discusses the implications of these findings for future research and explores how these methodologies can be applied to other types of city network analyses.

2. Literature review

2.1 City relationships

In empirical research, intercity relationships are commonly conceptualized as 'spatial interactions (Miller, 2004).' These interactions are normally measured through the exchange of goods, movement of people, and collaboration in research and development, and the strength of these relationships is quantified by aggregating the total volume of these spatial interactions. Additionally, the spatial interactions can be indirect as "shared attributes" between places (Neal et al, 2022). The underlying assumption here is that stronger shared attributes between two cities facilitate interaction between them. A notable example of this is the 'interlocking world city network model' proposed by Taylor (2001), which evaluates city relationships based on the presence of branch offices of the same advanced producer services firms, providing a well-structured approach to understanding global city networks (e.g. Taylor and Derudder, 2015; Derudder et al., 2010).

A consistent observation from such empirical research is that city interactions align with the classical gravity model of urban interaction (e.g. Zipf, 1946; Neal, 2010; van Oort et al., 2010; Zhang, 2020; Li and Neal, 2023), which suggests that the level of interaction between two cities is well positively related to their populations. This relationship is intuitively plausible, as larger cities, with their extensive agglomerations, facilitate more interactions, but this does not necessarily mean these interactions are more 'intensive' per unit (Neal et al., 2014 and 2022; Cortinovis and van Oort, 2022; Steijn et al., 2022).

Consider the example of the high volume of people moving between London and Paris. Traditional approaches might interpret this heavy traffic as indicative of a strong relationship, but it may not necessarily indicate a positive correlation between the two cities. Instead, such interactions could predominantly be a function of their large populations, reflecting their roles as major, well-connected urban centers, rather than indicative of deeper socio-economic links.

Thereby, relying solely on the total volume of interactions between cities to represent their relationship strength, while useful, may fall short of capturing the nuanced essence of "relatedness" as described in Tobler's first law. Miller (2004, p. 285) emphasizes that recognizing a relationship between cities means that "*At the very least, we are claiming that there is a positive or negative correlation between these entities.*" And there are fruitful research have identified various factors that affect city relationship strength (). However, this definition might be too restrictive or hard to capture the complexity of city relationships, as city associations often do not exhibit straightforward linear correlations and are highly dependent on specific contexts (Storper, 2009; Gong and Hassink, 2020; Bathelt and Storper, 2023).

In this regard, statistics and probability theory offer valuable insights. From a probability perspective, a strong relationship is indicated when the likelihood of interaction between two entities exceeds what would typically be expected by chance (Sinclair, 1966 and 1991). It acknowledges that relationships are often elusive and indirect correlations—whether positive or negative—represent just one of the many possible forms of association (Evert, 2008).

This association measurement approach shifts the focus from measuring direct pairwise interaction levels to identifying outlier cases that warrant further detailed investigation (Baker and McEnery, 2005; Baker et al., 2008 and 2013). The underlying assumption is that if most cities interact in a typical manner, outliers that deviate from this norm suggest the presence of hidden factors influencing their interaction levels. To identify these outlier cases, it is necessary to examine the overall interactions of cities within the broader network, including the internal activities of all cities.

Following the previously mentioned example of the people flow. Despite significant movement between London and Paris, this traffic may represent only a small portion of each city's overall activities. Consequently, the probability of interaction between London and Paris might not be greater than that between two smaller cities. Conversely, if the interaction between Rotterdam and Antwerp is disproportionately high compared to what chance would predict, it raises a signal for targeted analysis to uncover the hidden factors driving this unusual pattern.

This possibility-based perspective has demonstrated success across various disciplines, including communication (Ouyang et al., 2023), stock trade (e.g. Fiedor, 2014), gene selection (e.g. Cai et al., 2009). A notable example in linguistics is the shift of word association measurements from frequency-based to effect and certainty-based, in which the interdependency between words is revealed (e.g. Baker et al., 2008; Brookes and McEnery, 2020). By applying similar principles, this approach can enrich our analysis of urban systems, enabling a deeper insight into the complexity and interdependency between cities beyond mere total volume-based analysis.

2.2 Association measurement

Linguist J.R Firth in 1957 introduced the notion of "collocation." Firth posited that the meaning and usage of a word can be largely characterized by its most typical collocates: "*You shall know a word by the company it keeps* (Firth 1957, p179). " Expanding upon Firth's contributions, collocation is defined as "*a combination of two words that exhibit a tendency to occur near each other in natural language* (Evert, 2008, p4). " This notion is further emphasized by Hunston (2011, p14), who asserts that "*the meaning of any word cannot be identified reliably if the word is encountered in isolation.*"

When measuring word associations, linguists often encounter the issue that common words frequently appear together not due to meaningful connections but simply because of their individual high usage rates (Stubbs, 1994 and 2001; Widdowson, 2000 and 2008; Kang, 2018). This phenomenon parallels the analysis of city relationships, where larger cities—akin to common words—tend to have more interactions simply due to their own size and activity levels.

To address this issue, linguists refine their focus to distinguish between word pairs that are routine co-occurrences and those that appear together more frequently than chance would suggest. This distinction is essential because 'higher-by-chance' co-occurrences tend to provide more valuable clues for further exploration, indicating the possible existence of significant, hidden factors or specific contexts strengthening the likelihood of co-appearance.

The concept of "higher-by-chance" encompasses two dimensions (Evert, 2008): first, the effect, which examines the extent to which the co-appearance of two words exceeds what would typically be expected by chance; and second, the certainty, which assesses the statistical confidence level that their co-appearance is not merely due to coincidence.

The following example illustrates the first dimension, the effect. Consider the following case: In a corpus of 1,000,000 articles, we have two pairs of cities: City A and City B, and City C and City D. Both pairs co-occur 300 times. However, the individual appearances of these cities are: City A appears 10,000 times, City B 30,000 times, City C 1,000 times, and City D 6,000 times.

Despite the same number of co-occurrences (300 times) for both pairs, the individual appearance of each city varies significantly. To assess whether these co-occurrences indicate a strong association, we can start by assuming that there is no inherent relationship between cities —each city is independent of the other, which means their joint appearances in the text are purely coincidental. This baseline assumption would allow us to calculate the expected number of co-occurrences between two cities using the formula for the probability of independent events:

$$E_{AB} = \frac{O_A}{N} \times \frac{O_B}{N} \times N$$

$E_{AB}$is the estimated co-occurrence of city A and B in paragraphs under the assumption that city A and City B are independent. $\frac{O_A}{N}$ indicates the probability of City A occurring, and $\frac{O_B}{N}$ is the probability of City B occurring.

Under this assumption, the estimated co-occurrence of City A and City B would be $\frac{10,000}{1,000,000} \times \frac{30,000}{1,000,000} \times 1,000,000 = 300$. For City C and City D, the expected frequency would be $\frac{1,000}{1,000,000} \times \frac{6,000}{1,000,000} \times 1,000,000$ =6.

The actual observed co-occurrence (300) of City A and City B aligns with the expected frequency under the assumption of independence (both observed and expected are 300), suggesting their co-appearance is no higher than coincidence. On the other hand, the observed co-occurrence of City C and City D (300) is 50 times higher than the expected frequency (6), indicating a relationship that is significantly beyond mere coincidence.

Regarding the second issue, certainty, consider the case of City E and City F in the same corpus of 1,000,000 articles. These cities co-occur 3 times and their individual appearances are 10 and 30 times respectively. Based on the previous analysis of effects, the expected co-appearance would be $\frac{10}{1,000,000} \times \frac{30}{1,000,000} \times 1,000,000 = 0.0003$, making the actual co-appearance (3) is 10,000 times greater than the expected co-appearance (0.0003). This would indicate a strong association. However, the very low absolute number of their co-occurrences in such a vase corpus introduces a degree of uncertainty about this result's reliability.

3. Method

This section will detail how to quantitatively assess two dimensions of associations (effects and certainty) using statistical and probability theory. Both dimensions are based on a comparison between actual observations and a benchmark estimation model. Specifically, the effect dimension evaluates how observed values compare to the model's predictions for the same cases, and the certainty dimension assesses whether the observed values conform to a benchmark that is representative of general cases.

3.1 Measurement of effects

Several geographers have developed methods that are aimed at measuring the effect dimension of city relationships. For instance, Tobler pioneered a method that used the co-occurrence of 119 pre-Hittite towns on cuneiform tablets made almost 4000 years ago in Cappadocia and implemented the gravity model to reconstruct urban systems (Tobler and Wineburg, 1971). Building on this foundation, Meijers and Peris (2019) and Tongjing et al. (2024a and 2024b) adopted the gravity model as a benchmark, comparing it against actual placename co-appearance to estimate the "intensity" of city relationships. The key assumption is that a ratio exceeding 1 indicates the presence of additional factors influencing the relationship strength, besides population and distance. However, these studies often encounter difficulties in clearly distinguishing the effects of agglomeration from genuine associative relationships between cities, tending to favor larger cities or those at greater distances.

In linguistic studies of word association, researchers adopt the mutual information metric from information theory, originally from information theory (Shannon, 1948). The advantage of this method is that it can capture a broader range of associations, not limited to the linear or monotonic relationships that traditional correlation coefficients (such as Pearson, Spearman, and Kendall) are designed to measure.

This approach is the same as the analysis described in Section 2.2, where the actual observed data are compared against the expected data under the assumption that the two entities are independent. Specifically, this involves the use of pointwise mutual information (Church and Hanks, 1990), which is formulated as:

$$MI_{AB} = \frac{O_{AB}}{N} / E_{AB}$$

$$E_{AB} = P_A \times P_B \times N$$

where MI_{AB} is the mutual information of AB, O_{AB} is the actual co-occurrence of A and B, $E_{AB}$is the estimated co-occurrence of city A and B in paragraphs under the assumption that city A and city B are actually independent, and N is the size of the corpus.

A mutual information value greater than 1 suggests an association is stronger than what would be expected if they were independent of each other. This means that knowing information about one entity provides significant information about the other, demonstrating interdependence. A value of 1 signifies independence, and a value less than 1 indicates an association that is weaker than estimated under conditions of independence, suggesting the two entities actually repel each other.

3.2 Measurement of certainty

Regarding the dimension of certainty, the underlying benchmark assumption is that all city interactions are due to coincidence. In geography, similar to the measurement of effects, researchers have also acknowledged this dimension. A remarkable example is Neal (2014 and 2022), who introduced the Stochastic Degree Sequence Model (SDSM) to identify the backbone of city relationships, which considers word co-occurrence following a stochastic process.

In addition to SDSM, other statistical significance metrics are also often employed. Linguists commonly start with the assumption that the co-occurrence of any two words is coincidental. In other words, if one were to collect all instances where a word co-occurs with others, the distribution of these co-occurrences is expected to follow a normal distribution. Thereby, it would generally be statistically unlikely for a word to exhibit an unusually high co-occurrence with a specific term, compared to its overall pattern of co-occurrences. This methodology and its implications are further illustrated in Figure 1.

Figure 1 Illustration of the statistical confidence

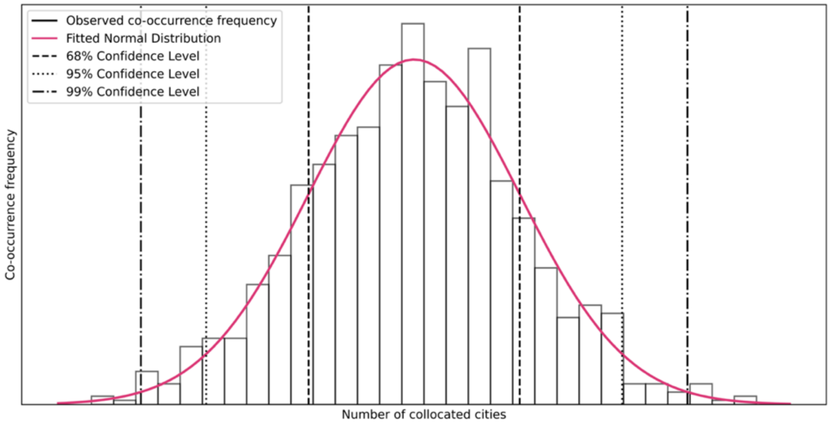


Figure 1 illustrates a benchmark scenario where the observed co-occurrences for a particular city are assumed to follow a normal distribution, depicted by the red curve. In this scenario, the majority of the co-occurrence numbers are clustered around the average value, with very few deviating significantly from this average. This suggests a low probability for the co-occurrence frequency to be either substantially higher or lower than the average.

Under this framework, we set the hypothesis that co-occurrences significantly above or below the mean are not likely due to random chance. To quantify this assertion, a confidence level can be established. For instance, if we set a confidence level at 90%, then co-occurrences that exceed what is predicted by this threshold can be considered non-coincidental. This implies that there is a 10% chance that labeling these co-occurrences as non-coincidental is incorrect.

3.3 Relationship classification

Through analyzing the relationship's dual dimensions (effects and certainty), this study introduces a classification framework designed to distinguish the varied associative natures of city relationships. By leveraging two widely used metrics—mutual information to measure effects, and confidence level to gauge certainty—city relationships are categorized into four types:

1. Low Mutual Information, Low Confidence: This scenario is typical for most city relationships, where interactions occur occasionally without significant underlying implications.

2. Low Mutual Information, High Confidence: This classification characterizes relationships where cities interact frequently, but the interactions do not significantly exceed what might be expected by chance. This pattern can be expected in large cities—while frequently engaging with each other, each also maintains broad interactions with many others, thereby diluting the significance of this seemingly strong relationship.

3. High Mutual Information, Low Confidence: In this category, the interactions between two cities account for a significant portion of their total interactions, although neither city has many interactions. This situation could arise under two different scenarios:

The interactions are within specific contexts but measured under a broad sampling. For instance, the interactions occur within specific contexts but are assessed as part of a larger dataset. For example, two small university towns collaborate intensively in medical research, yet their general level of collaboration across all fields remains low.

Alternatively, these interactions might still be coincidental. To differentiate between the two scenarios would require an in-depth analysis of the contexts in which these cities interact.

4. High Mutual Information, High Confidence: relationships where cities not only frequently interact but also share a strong, statistically significant connection, suggesting a deep and meaningful relationship, indicating a true strong relatedness between two cities.

3.4 Case study description

Different association metrics can lead to markedly different outcomes. Mutual Information tends to identify rare pairs of words that occur together, thus shedding light on unique or specialized relationships. Conversely, confidence level measurement favors more common pairings, providing insights into the predominant patterns. Given these differences, to fully grasp the complexities of city relationships—or any set of relationships—it is crucial to assess them from multiple dimensions.

To better illustrate the multi-dimensional nature of associations, this study will use the toponym co-occurrence model, which uses the placename co-appearance to represent city relationship strength, as it closely aligns with the collocation approach previously introduced in this paper. Specifically, this study will calculate five relationship measurement metrics: the total number of placename co-occurrences, mutual information, confidence level, combined Metric of mutual information and statistical confidence, and the compared Gravity Model (the ratio of actual co-occurrence frequency against the gravity model estimation).

The combined metric of mutual information and confidence integrated the two dimensions. In this case, relationships that surpass this confidence level are considered significant and their corresponding mutual information is set as the relationship weight. Conversely, relationships that do not meet this threshold are considered insignificant and are assigned a value of 0. The confidence level is set at 90% in this study.

The data for this analysis is sourced from the English Wikipedia corpus as of January 1, 2023, which comprises 16,820,287 URLs (Wikipedia, 2023). The 100 largest European cities, as determined by the population of their urban units are selected, as reported by Eurostat in 2021.

In this study, “co-occurrence frequency” is defined as the total number of URLs where the names of two cities appear together within a single paragraph, a textual collocation. While this definition might not capture as strong a relationship as those within a specific span, it is particularly appropriate for contexts with a free word order. This consideration is based on the understanding that related words may not always be adjacent at the surface level.

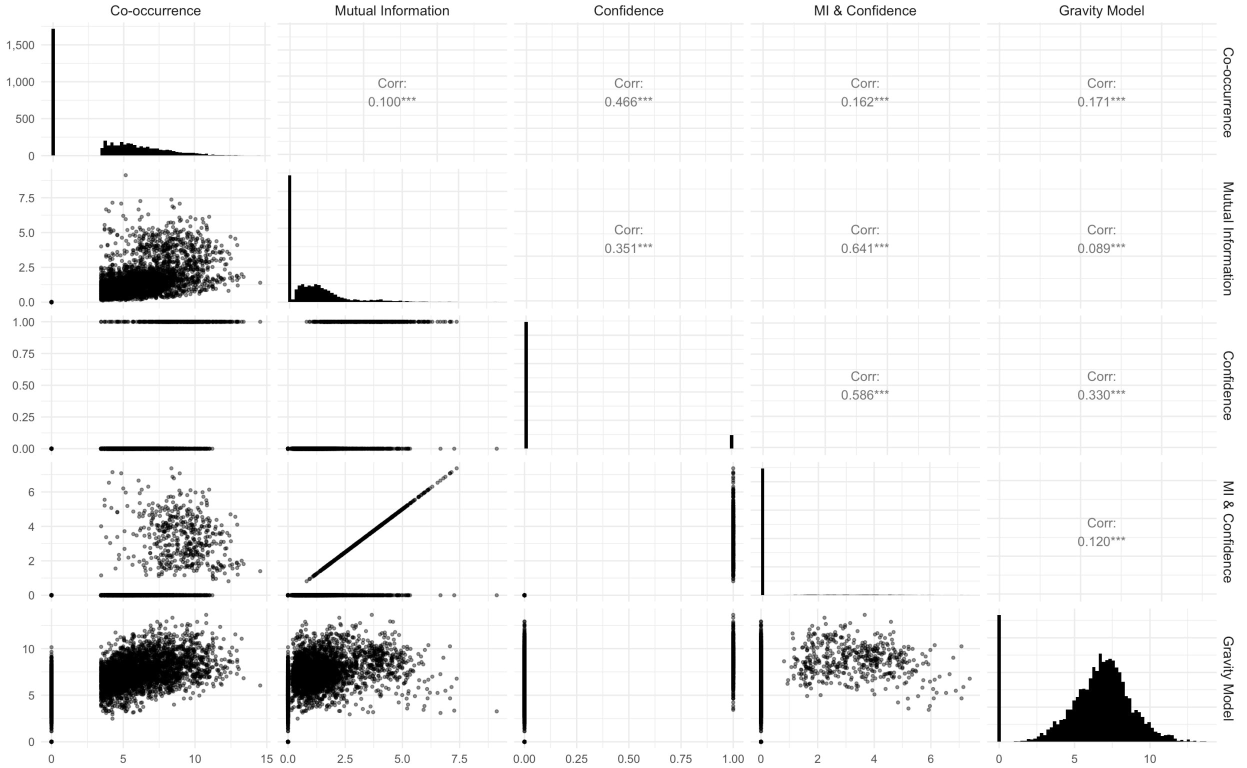
4. Results

This section will begin by comparing the variation between different relationship metrics to indicate that no single metric can comprehensively capture the nature of relationships. Following this, we will visualize the top relationships, which will help reveal patterns that are not immediately apparent. Finally, based on the score of effects and certainty, relationships will be classified to gain a deeper understanding of the nature of city relationships.

4.1 Results of Spearman rank correlation

The Person correlation results are presented in Figure 2.

Figure 2 Correlation between different metrics



In Figure 2, the lower triangle consists of scatter plots is the relationship between the city relationships in each metric. The diagonal line is the histogram of the corresponding relationship strength, and the upper triangle is the correlation value between each pair of metrics. To enhance visualization, data in the lower triangle and along the diagonal are adjusted by adding 1 to prevent calculation errors, followed by a logarithmic transformation to base 2.

As depicted in Figure 2, the relatively low correlations between all association metrics highlight that no single metric can comprehensively reflect city associations. The generally low correlation between co-occurrence and any of the association metric scores suggests that measurements based solely on the total amount of interactions fail to accurately reflect the true nature of placename associations. Additionally, the low correlation score between mutual information and confidence (0.351) highlights that metrics assessing different dimensions of relationships can produce quite varied results. Notably, the correlation between mutual information and the gravity model is also low (0.089), despite both metrics attempting to quantify the effect dimension of relationships.

Only the correlation between the confidence metric and the combined metric (0.586), and the mutual information metric and the combined metric (0.641) showed moderately positive relationships. However, this is because the confidence metric uses a binary classification where the majority of relationships are valued in 0, and consequently, it turned the majority of the combined metric to 0 as well. This alignment is largely due to the methodological overlap where both metrics dismiss weaker relationships.

This correlation result suggests that different metrics can lead to varied outcomes, even when they aim to measure similar aspects of intercity relationships, therefore, it is essential to employ multiple metrics to accurately capture the multi-dimensional nature of relationships.

4.2 Analysis of top 100 relationships in each metrics

The top 100 relationships as identified in each metric are visualized in Figure 3.

Figure 3 City network by various collocation metrics.

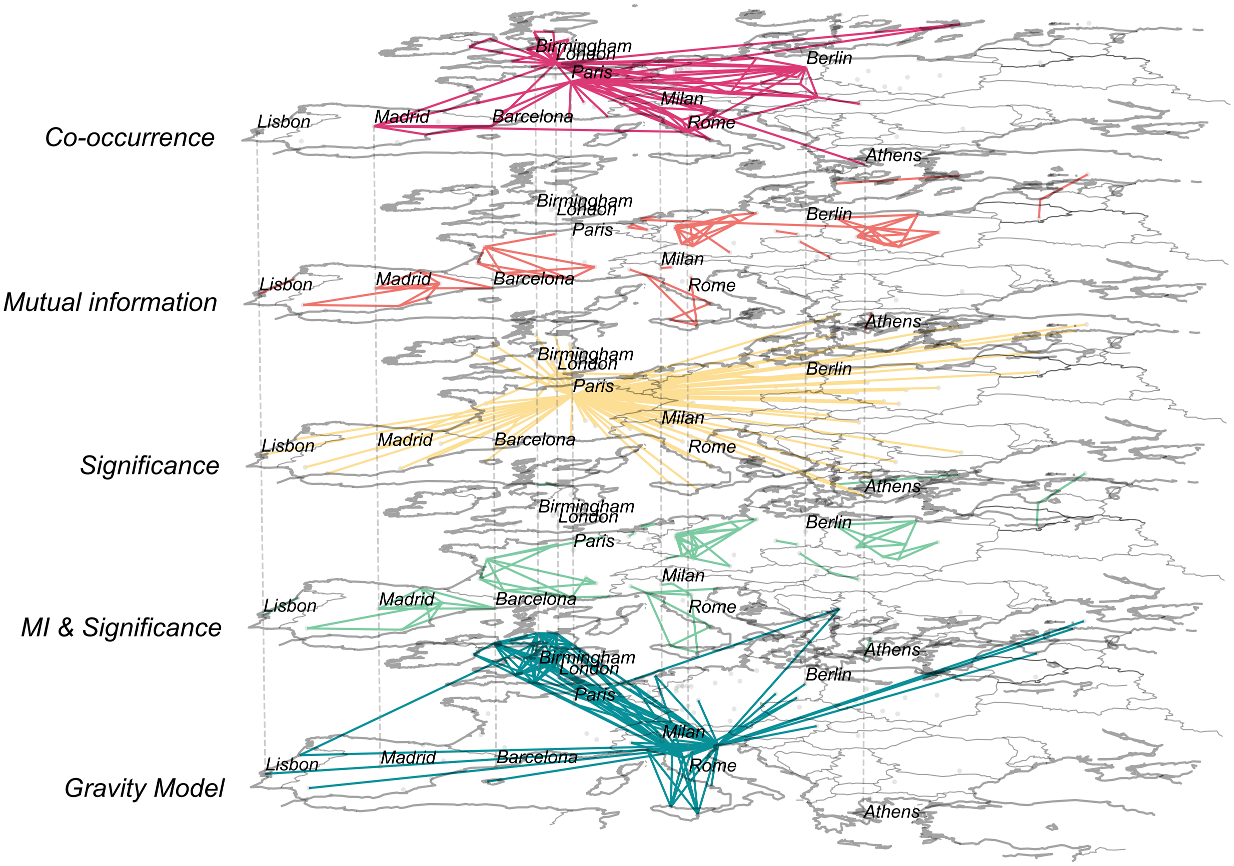


Figure 3 shows, generally, the strong relationships identified by co-occurrence number and confidence tend to be between larger cities. This is likely reflective of the higher visibility and interaction frequency among major cities. In contrast, the gravity model tends to emphasize relationships involving smaller cities or those separated by long distances because the gravity model's results are estimated by population size and distance. Meanwhile, the top relationships identified through mutual information are predominantly focused on smaller cities within a single country, often clustered within the same region. This pattern suggests that there is a risk of overlooking important regional connections if focusing solely on the total volume of interactions.

The top 10 relationships with the highest value are listed in Table 1 for comparison.

Table 1 Top 10 relationships with the highest value

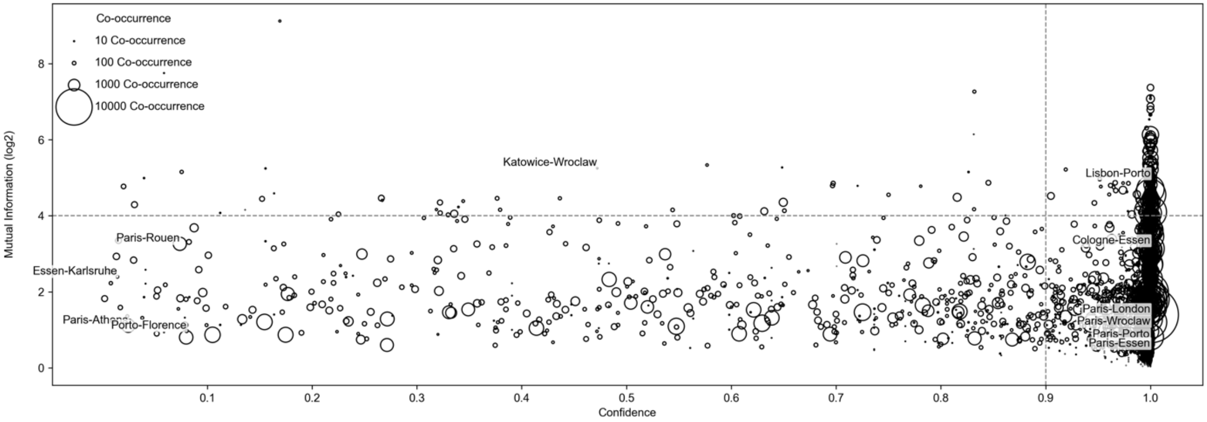
Relationship	Frequency	Relationship	MI	Relationship	Confidence	Relationship	MI& Confidence	Relationship	Gravity model
Paris-London	23842	Duisburg-Bottrop	556.13	Paris-London	100%	Duisburg-Dortmund	164.57	Florence-Venice	12621.00
London-Manchester	10692	Duisburg-Dortmund	164.57	Paris-Berlin	100%	Wroclaw-Gdansk	140.72	Newcastle-Portsmouth	10186.62
London-Edinburgh	9741	Essen-Bottrop	152.87	London-Berlin	100%	Poznan-Wroclaw	134.55	Venice-Bergamo	7745.78
Paris-Berlin	8478	Wroclaw-Gdansk	140.72	Paris-Madrid	100%	Palermo-Catania	134.19	Venice-Edinburgh	7554.66
London-Berlin	8051	Poznan-Wroclaw	134.55	London-Madrid	100%	Düsseldorf-Duisburg	116.72	Genoa-Venice	7525.75
Manchester-Liverpool	7861	Palermo-Catania	134.19	Madrid-Berlin	100%	Mannheim-Karlsruhe	109.68	Liverpool-Newcastle	6857.60
Paris-Rome	7829	Düsseldorf-Duisburg	116.72	Paris-Milan	100%	Duisburg-Bielefeld	100.11	Newcastle-Bristol	6409.36
London-Birmingham	7731	Mannheim-Karlsruhe	109.68	London-Milan	100%	Poznan-Gdansk	91.76	Florence-Edinburgh	6272.79
London-Liverpool	7668	Dortmund-Bottrop	100.48	Madrid-Milan	100%	Dortmund-Bielefeld	78.18	Newcastle-Nottingham	5859.16
Glasgow-Edinburgh	7167	Duisburg-Bielefeld	100.11	Paris-Barcelona	100%	Lodz-Poznan	72.64	Liverpool-Portsmouth	5851.96

4.3 Results of intercity relationships classification

Figure 3 presents the relationship between confidence and mutual information for city pairs. Each point on the figure represents a city pair. The x-axis is the confidence level, while the y-axis indicates the degree of mutual information. Additionally, the size of each point corresponds to the number of placename co-occurrences. Then it is split into four categories, whether the confidence level is above 90%, indicating whether the relationships are coincidental, and whether the mutual information is higher than 4, indicating the relationship effect is significantly interdependent.

A few examples are labeled for illustration.

Figure 3 Relationship between mutual information and confidence



Low Mutual Information and Low Confidence: Most city relationships, such as those between Essen-Karlsruhe and Porto-Florence, fall into this category. This observation suggests that their co-occurrences are no more frequent than random chance, indicating that these few instances are likely coincidental.

Low Mutual Information but High Confidence: Large or well-known cities often fall into this category. For example, the relationship between Paris and London, while frequently mentioned together in various texts, does not show co-occurrences that are statistically significant beyond random chance. This suggests that their high co-occurrence number is primarily due to their significant individual socioeconomic roles, rather than any substantial, meaningful intercity relationship.

High Mutual Information and Low Confidence: In contrast to the previous category, relationships in this group typically have only a smaller total number of co-occurrences. This low total reduces the confidence level, but paradoxically, the significance of serendipitous relationships may seem exaggerated. An example is the pairing of Katowice and Wroclaw. Determining the statistical significance of their co-occurrence requires a thorough analysis of the contextual data surrounding their mentions.

High Mutual Information and High Confidence: This category represents relationships that are both frequently mentioned and exhibit a strong, statistically significant contextual connection. An example is the relationship between Lisbon and Porto. As Portugal's largest and second-largest cities, their frequent co-occurrence not only underscores their individual prominence but also reflects their shared characteristics and mutual dependencies. This indicates a robust and meaningful relationship across various contexts, highlighting a significant and relevant connection between the two cities.

5. Conclusion

Grounded in statistics and probability theory, this study highlights that merely counting interaction volumes fails to accurately capture the true level of association between two places, as interactions can also occur between places that are independent of each other (Neal 2014a; Neal et al., 2022).

To discern genuine city associations, this study focuses on 'higher-than-by-chance,' from statistics and probability theory. This concept consists of two dimensions: effects and certainty: 'Effects' refers to the degree to which interactions are beyond what random chance would explain, while 'certainty' assesses the likelihood that these associations are not merely coincidental. To comprehensively illustrate the multi-dimensional nature of relationships between cities, this study adopts approaches from linguistic studies of word associations, applying five distinct metrics to measure city relationship strength based on the placename co-occurrence model. The results are compared and categorized to delineate the varying natures of these associations.

The findings reveal substantial variations in the results of relationship metrics, indicating that no single metric can fully capture the level of association between places. This complexity manifests in three ways: first, a high co-occurrence number of two cities does not necessarily equate to a high association score; often, it may be attributable to each city's own large appearance. Second, there is a lack of correlation between dimensions; a high score in effects, which suggests a strong association, does not consistently align with a high confidence score, and vice versa. Third, even when focusing on the same dimension, effects, different models like mutual information and the gravity model produce varied outcomes. Mutual information typically emphasizes relationships within the same region, while the gravity model tends to highlight connections between cities that are farther apart.

The classification framework proposed in this study distinguishes between seemingly strong city relationships and those that are genuinely associative. Specifically, mutual information can be a potentially suitable indicator for measuring the level of interdependency between cities, providing crucial insights for policies designed to enhance regional cohesion. Additionally, the confidence level is able to extract the most significant city relationships by transforming the weighted city network into a binary one. As cities vary only in the degree of their relationships, maintaining all weights often distorts the overall network structure (Watts, 2008; Serrano et al., 2009; Neal, 2014).

This study used the toponym co-occurrence model for ease of demonstration, due to its similarity to the adopted collocation approach in linguistic studies. Admittedly, the meanings of relationships extracted from Wikipedia are less obvious compared to relationship strength measured by scientific collaborations or people movements. Despite this, the method is versatile and can be adapted to these and other contexts, extending its applicability to assess associations between broader entities such as authors and articles, or products and consumers (Newman, 2001a, b; Zweig, and Kaufmann, 2011).

Additionally, there is room left for advancing this method. This paper adopts a rather restrictive way to measure association, in which only co-occurrence between placenames is recorded. A more comprehensive approach would involve examining the co-occurrence of all words associated with each city. However, pursuing such an approach would necessitate significantly more extensive data processing, presenting both a substantial challenge and a valuable opportunity for further exploration in the study of relationships.

Reference

Baker, P., & McEnery, T. (2005). A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. Journal of language and politics, 4(2), 197-226.

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. Discourse & society, 19(3), 273-306.

Baker, P., Gabrielatos, C., & McEnery, T. (2013). Sketching Muslims: A corpus driven analysis of representations around the word 'Muslim'in the British press 1998–2009. Applied linguistics, 34(3), 255-278.

Balland, P. A., Boschma, R., & Frenken, K. (2022). Proximity, innovation and networks: A concise review and some next steps. Handbook of proximity relations, 70-80.

Bathelt, H., & Storper, M. (2023). Related variety and regional development: a critique. Economic Geography, 99(5), 441-470.

Brookes, G., & McEnery, T. (2020). Correlation, collocation and cohesion: A corpus-based critical analysis of violent jihadist discourse. Discourse & Society, 31(4), 351-373.

Cai, R., Hao, Z., Yang, X., & Wen, W. (2009). An efficient gene selection algorithm based on mutual information. Neurocomputing, 72(4-6), 991-999. Cambridge Journal of Regions, Economy and Society 13 (3): 475–90.

Cao, Z., Derudder, B., & Peng, Z. (2019). Interaction between different forms of proximity in inter‐organizational scientific collaboration: The case of medical sciences research network in the Yangtze River Delta region. Papers in Regional Science, 98(5), 1903-1924.

Cao, Z., Derudder, B., Dai, L., & Peng, Z. (2022). 'Buzz-and-pipeline'dynamics in Chinese science: the impact of interurban collaboration linkages on cities' innovation capacity. Regional Studies, 56(2), 290-306.

Capello, R., & Caragliu, A. (2018). Proximities and the intensity of scientific relations: synergies and nonlinearities. International Regional Science Review, 41(1), 7-44.

Castaldi, C., & Drivas, K. (2023). Relatedness, cross-relatedness and regional innovation specializations: An analysis of technology, design, and market activities in Europe and the US. Economic Geography, 99(3), 253-284.

Castells, M. (1996), The Rise of the Network Society (vol. 1). Malden, MA: Blackwell.

Cortinovis, N., & van Oort, F. (2022). Economic networks, innovation and proximity. In Handbook of proximity relations(pp. 292-306). Edward Elgar Publishing.

Derudder, B. (2021). Network analysis of “urban systems”: Potential, challenges, and pitfalls. Tijdschrift voor Economische en Sociale Geografie 112 (4):404–20.

Derudder, B., Devriendt, L., & Witlox, F. (2007). Flying where you don't want to go: An empirical analysis of hubs in the global airline network. Tijdschrift voor economische en sociale geografie, 98(3), 307-324.

Derudder, B., Taylor, P., Ni, P., De Vos, A., Hoyler, M., Hanssens, H., ... & Yang, X. (2010). Pathways of change: Shifting connectivities in the world city network, 2000—08. Urban studies, 47(9), 1861-1877.

Dong, X., Zheng, S., & Kahn, M. E. (2020). The role of transportation speed in facilitating high skilled teamwork across cities. Journal of Urban Economics, 115, 103212.

Evert, S. (2008). Corpora and collocations. Corpus linguistics. An international handbook, 2, 1212-1248.

Fiedor, P. (2014). Networks in financial markets based on the mutual information rate. Physical Review E, 89(5), 052801.

Gong, H., and Hassink, R. 2020. Context sensitivity and economic-geographic (re)theorising.

Guo, Y., Li, B., & Han, Y. (2020). Dynamic network coupling between high-speed rail development and urban growth in emerging economies: Evidence from China. Cities, 105, 102845.

He, C., Wu, J., & Zhang, Q. (2020). Research leadership flow determinants and the role of proximity in research collaborations. Journal of the Association for Information Science and Technology, 71(11), 1341-1356.

Hunston, S. (2002) Corpora in Applied Linguistics. Cambridge: Cambridge University Press.

Kang, B. M. (2018). Collocation and word association: Comparing collocation measuring methods. International journal of corpus linguistics, 23(1), 85-113.

Lai, J., & Pan, J. (2020). China's city network structural characteristics based on population flow during spring festival travel rush: Empirical analysis of “tencent migration” big data. Journal of Urban Planning and Development, 146(2), 04020018.

Lewoniewski, W. (2022). Identification of important web sources of information on Wikipedia across various topics and languages. Procedia Computer Science, 207, 3290-3299.

Li, X., & Neal, Z. P. (2024). Are larger cities more central in urban networks: A meta-analysis. Global Networks, 24(2), e12467.

Liu, Y., Wang, F., Kang, C., Gao, Y., & Lu, Y. (2014). Analyzing Relatedness by Toponym Co-O ccurrences on Web Pages. Transactions in GIS, 18(1), 89-107.

Meijers, E. J., & Burger, M. J. (2017). Stretching the concept of ‘borrowed size’. Urban studies, 54(1), 269-291.

Meijers, E., & Peris, A. (2019). Using toponym co-occurrences to measure relationships between places: Review, application and evaluation. International Journal of Urban Sciences, 23(2), 246-268.

Miller, H. J. 2004. Tobler’s first law and spatial analysis. Annals of the Association of American Geographers 94 (2):284–89.

Neal, Z. P. (2010). Refining the air traffic approach to city networks. Urban Studies, 47(10), 2195–2215.

Neal, Z. P., Domagalski, R., & Sagan, B. (2022). Analysis of spatial networks from bipartite projections using the R backbone package. Geographical Analysis, 54(3), 623-647.

Newman ME (2001a) Scientific collaboration networks I. Phys Rev E 64:016,131

Newman ME (2001b) Scientific collaboration networks II: shortest paths, weighted networks, and centrality. Phys Rev E 64:016,132

Ouyang, C., Liu, Y., Yang, H., & Al-Dhahir, N. (2023). Integrated sensing and communications: A mutual information-based framework. IEEE Communications Magazine, 61(5), 26-32.

Salvini, M. M., & Fabrikant, S. I. (2016). Spatialization of user-generated content to uncover the multirelational world city network. Environment and Planning B: Planning and Design, 43(1), 228-248.

Shannon, C. E. (1948). A mathematical theory of communication. The Bell system technical journal, 27(3), 379-423.

Shenoy, K., Ilievski, F., Garijo, D., Schwabe, D., & Szekely, P. (2022). A study of the quality of Wikidata. Journal of Web Semantics, 72, 100679.

Sinclair, J. (1991) Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Sinclair, J. M. (1966). Beginning the study of lexis. memory of JR Firth. In C. E. Bazell, J. C. Catford, M. A. K. Halliday, and R. H. Robins (eds.), In Memory of J. R. Firth, pages 410–430. Longmans, London.

Storper, M. (2009). Regional context and global trade. Roepke lecture in economic geography. Economic Geography, 85(1), 1-21.

Straka, M. J., Caldarelli, G., & Saracco, F. (2017). Grand canonical validation of the bipartite international trade network. Physical Review E, 96(2), 022306.

Steijn, M. P., Koster, H. R., & van Oort, F. G. (2022). The dynamics of industry agglomeration: Evidence from 44 years of coagglomeration patterns. Journal of Urban Economics, 130, 103456.

Stvilia, B., Twidale, M. B., Smith, L. C., & Gasser, L. (2008). Information quality work organization in Wikipedia. Journal of the American society for information science and technology, 59(6), 983-1001.

Stubbs, M. (1994) 'Grammar, Text, and Ideology: Computer-assisted Methods in theLinguistics of Representation', Applied Linguistics 15(2): 201–23.

Stubbs, M. (2001) Words and Phrases: Corpus Studies of Lexical Semantics. Oxford:Blackwell.

Taylor, P. J., and B. Derudder (2016). World City Network: A Global Urban Analysis, 2nd ed. New York, NY: Routledge.Liu, X., Derudder, B., & Wu, K. (2016). Measuring polycentric urban development in China: An intercity transportation network perspective. Regional Studies, 50(8), 1302-1315.

Taylor, P., & Derudder, B. (2015). World city network: a global urban analysis. Routledge.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. Economic geography, 46(sup1), 234-240.

Tobler, W., & Wineburg, S. (1971). A cappadocian speculation. Nature, 231(5297), 39–41.

Tongjing, W., Meijers, E., & Wang, H. (2023). The multiplex relations between cities: a lexicon-based approach to detect urban systems. Regional Studies, 57(8), 1592-1604.

Tongjing, W., Meijers, E., Bao, Z., & Wang, H. (2024a). Intercity networks and urban performance: a geographical text mining approach. International Journal of Urban Sciences, 28(2), 262-283.

Tongjing, W., Yin, Z., Bao, Z., & Meijers, E. (2024b). Intercity relationships between 293 Chinese cities quantified based on toponym co-occurrence. Cybergeo: European Journal of Geography.

van Oort, F., Burger,M., & Raspe, O. (2010). On the economic foundation of the urban network paradigm: Spatial integration,

Widdowson, H. G. (2000). On the limitations of linguistics applied. Applied linguistics, 21(1), 3-25.

Widdowson, H. G. (2008). Text, context, pretext: Critical issues in discourse analysis. John Wiley & Sons.

Wikipedia. (2023). https://dumps.wikimedia.org, accessed on September 3, 2023

Wu, J., Feng, Z., Zhang, X., Xu, Y., & Peng, J. (2020). Delineating urban hinterland boundaries in the Pearl River Delta: An approach integrating toponym co-occurrence with field strength model. Cities, 96, 102457.

Zhang,W., Derudder, B.,Wang, J., &Witlox, F. (2020). An analysis of the determinants of the multiplex urban networks in the Yangtze River Delta. Tijdschrift voor Economische en Sociale Geografie, 111(2), 117–133.

Zhu, B., Pain, K., Taylor, P. J., & Derudder, B. (2022). Exploring external urban relational processes: Inter-city financial flows complementing global city-regions. Regional Studies, 56(5), 737-750.

Zipf, G. K. (1941). National unity and disunity: The nation as a bio-social organism. The Principia Press, Inc.

Zweig, K. A., and M. Kaufmann (2011). "A systematic approach to the one-mode projection of bipartite graphs." Social Network Analysis and Mining 1(3), 187–218.