Rethinking City Relationships:
Moving from Frequency Analysis to Collocation Analysis

Wang Tongjing

Department of Human Geography and Spatial Planning, Faculty of Geosciences, Utrecht University, Utrecht, The Netherlands

1. Introduction

Cities benefit from each other through connections. Understanding the nature and extent of these connections is crucial for targeted policy-making and urban planning strategies.

The range of intercity connections is diverse and can be measured in many ways. These include tangible links such as commuting and material transport, quantifiable by the frequency of car and train services between cities. Moreover, collaborative efforts, evident in jointly authored scientific publications and shared patent registrations, add another dimension to intercity relationships. Additionally, the toponym co-occurrence analysis method offers a unique approach, quantifying the strength of relationships between places based on the occurrences of their names appearing together in texts. While these methods offer different perspectives on city relationships, they share a common assumption: the more frequent the interactions between two places (or co-occurrences, in the case of the toponym co-occurrence method), the stronger their relationship.

However, the reliance on interaction frequency as a primary indicator of relationship strength has its limitations. High frequencies of connection do not necessarily equal strong associations. For example, there might be more frequent train services, greater volumes of scientific collaboration, and higher occurrences in texts between large cities due to their size, but this does not inherently translate to a higher relative rate of interactions compared to those between smaller cities. However, on the other hand, relative frequency can disproportionately amplify the significance of interactions between smaller cities. In such cases, even sporadic connections may be misleadingly interpreted as important due to their relative infrequency. These two scenarios suggest the limitations of using frequency as the sole metric for determining relationship strength.

This complexity mirrors the challenges in linguistic analyses of word relationships, where linguists employ the concept of “collocation” to measure word associations. Collocation emphasizes the need to consider two crucial dimensions for accurate assessment: the effect size, indicating the degree of association between words, and the statistical confidence that their co-occurrence is beyond mere coincidence. Translating this principle to the realm of city relationships underscores that a comprehensive evaluation of city connections entails more than merely observing interactions with other cities.

This paper advocates for a transition from conventional frequency analysis to a more nuanced collocation analysis, with the aim to discern interdependencies between cities from interactions that might simply arise due to the individual prominence of the cities involved. To empirically validate this approach, the study conducts an examination of toponym co-occurrence among 100

European cities. The study will first demonstrate the variation of relationship strength based on different metrics. Then the study will proceed to categorize the city relationships into four groups that demonstrate genuine interdependency. Through this methodology, this study seeks to offer more accurate insights into the complexity of inter-city relationships.

2. Method

To understand why a high co-occurrence frequency does not automatically imply a strong association, 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.

To assess whether these co-occurrences indicate a strong association, we can start by assuming that the city pairs are independent of each other. This means the occurrence of one city does not affect the occurrence of the other. Under this assumption, we can calculate the estimated co-occurrence based on the probability of each city appearing independently. The formula for this calculation is as follows:

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

$E_{AB}$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 matches the expected frequency under the independent assumption (300), suggesting that the appearance of City A does not affect the appearance of City B, then the association between City A and City B may not be beyond random co-occurrence. On the other hand, the observed co-occurrence of City C and City D (300) is 50 times higher than the expected frequency (6). This substantial difference indicates a high dependent occurrence between City C and City D—the likelihood of City C appearing increases when City D is mentioned, and similarly, the appearance of City D is likely when City C is mentioned.

This comparison of actual to estimated co-occurrence, under the assumption that the two city names are independent, is called Mutual Information (MI). It derives from information theory for quantifying the “shared information” between two words.

However, besides measuring the association degree, it is also necessary to measure if the co-occurrence is actually due to coincidence. Considering this situation, City E and City F co-occur

3 times, but City E and City F only occur 10 and 30 times, respectively. Then using the previous analysis, the estimated frequency of E and F would be $\frac{10}{1,000,000} \times \frac{30}{1,000,000} \times 1,000,000 = 0.0003$, almost 0. Consequently, the actual observed co-occurrence (3) is 10,000 times greater than the expected frequency (0.0003), resulting in a significantly high mutual information value. But, this still doesn’t necessarily mean there's a strong association between City E and City F, considering the low absolute number of their co-occurrences in such a large corpus. Therefore, it is also necessary to account for a word’s absolute frequency of co-appearances and their distribution.

To determine whether the co-occurrence of two words is due to coincidence, statistical significance scores are often employed. This method begins with the assumption that all words co-located with a specific term are coincidental, and their frequencies should follow a normal distribution. Under this assumption, the probability of a word that has a much higher co-occurrence with the specific term than the rest of the collocated words is very low. It suggests that their co-occurrence is not due to coincidence. To evaluate this, statistical metrics such as the Z-score, T-score, and confidence levels are employed to determine if the observed frequency is statistically significant and unlikely to have occurred by random chance.

3. Case study

3.1 Case study description

To offer a multi-perspective on intercity relationships as reflected in the text, this study calculates five distinct collocation metrics, co-occurrence frequency, mutual information, statistical confidence, combined metric of mutual information and statistical confidence, and gravity model.

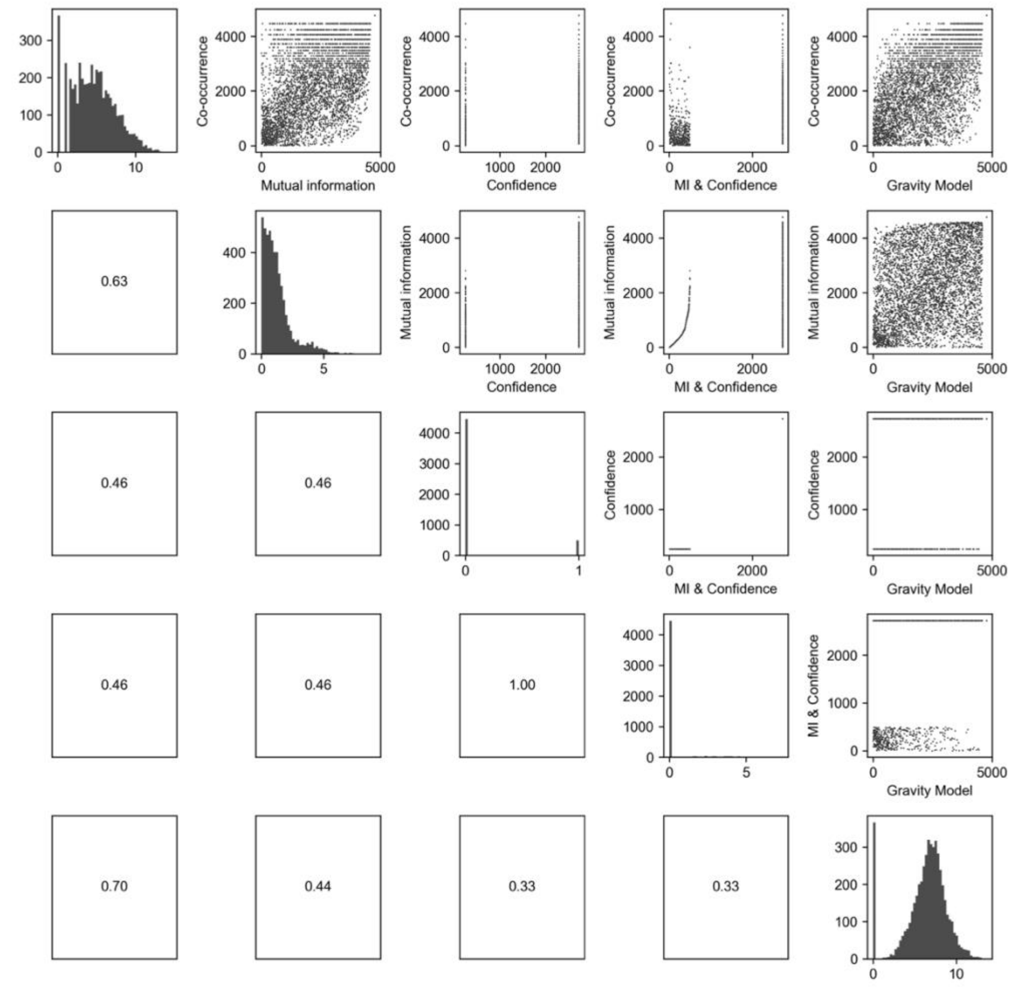
The data for this analysis is sourced from the English Wikipedia corpus as of January 1, 2023, which comprises 16,820,287 URLs (Wikipedia, 2023). The 100 largest European cities, as determined by the population of their urban units are selected, as reported by Eurostat in 2021.

3.2 Results of spearman rank correlation

Given the presence of many outlier values in the five types of relationships, the Spearman correlation is used to compare the relationships. This method focuses on rank correlations between the metrics rather than their absolute values, which helps mitigate the potential skewing effects of extreme values.

Figure 2 displays the results of the correlation. In the figure, the upper triangle consists of scatter plots illustrating the relationship between the rank of the city relationships in each type. The diagonal line represents the distribution of the value of each type of relationship strength, incremented by 1 and then log2-transformed. Meanwhile, the lower triangle of the figure details the Spearman correlation values for each pair of relationships.

Figure 2 Correlation between different metrics

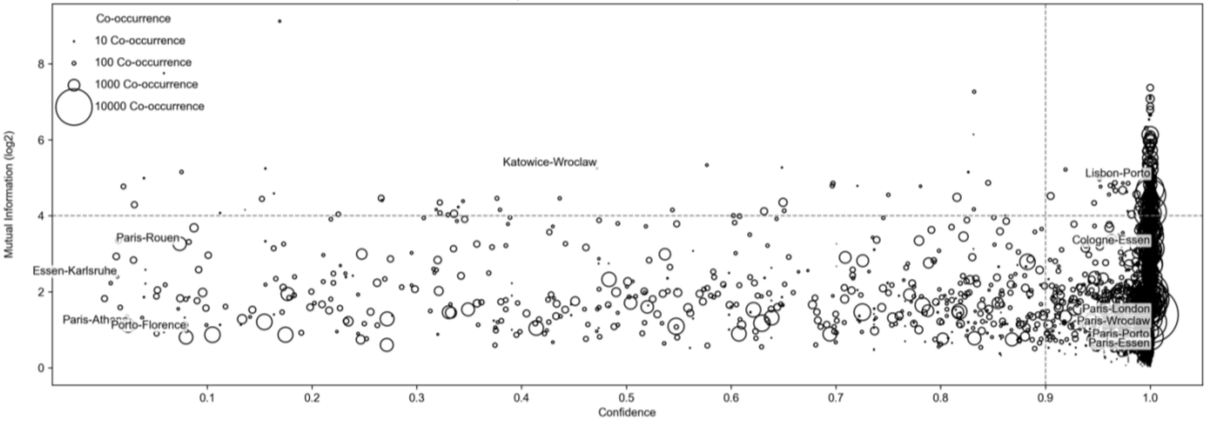


As depicted in Figure 2, the correlations between various association metrics are generally low except for a moderate positive correlation between co-occurrence and mutual information (0.63), as well as between co-occurrence and the gravity model (0.70). These correlations suggest that co-occurrence is more aligned with metrics that indicate effect size. However, it is noteworthy that mutual information and the gravity model, both metrics aimed at quantifying the strength of intercity relationships, do not exhibit a very high correlation (0.44). Additionally, the correlation between confidence and effect size metrics, as represented by mutual information and the gravity model, is relatively modest, standing at 0.45 and 0.33, respectively. The correlation between mutual information and the gravity model itself is also moderate at 0.44. This pattern suggests that different metrics can lead to varied outcomes, even when they aim to measure similar aspects of intercity relationships.

3.3 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 frequency of co-occurrence between the two city names. A few examples are labeled for illustration.

Figure 3 Relationship between mutual information and confidence



Based on the mutual information degree and confidence level, an intercity relationship can be classified into four types:

Low mutual information, low confidence: it means that the collocation patterns are infrequent and lack statistical dependence, suggesting their interactions are sparse, not dependent on each other, and their few co-occurrences are likely to be coincidental. Paris-Athens and Porto – Florence are in this category.

Low mutual information, high confidence: This situation suggests the relationships where the cities are frequently mentioned together, yet their co-occurrence does not imply a strong mutual dependence. The high confidence level confirms their frequent co-appearance, but the low mutual information suggests that this co-occurrence is more a reflection of each city's individual prominence rather than a direct interdependence between them. An example of this can be seen in the relationship between Paris and London. These cities are often mentioned together in various contexts due to their status as major European capitals, their economic significance, or as representative examples of large urban centers. However, the mention of one does not significantly influence the likelihood of the other being mentioned, indicating a lack of strong mutual dependence despite their frequent co-occurrence.

High mutual information, low confidence: In this case, two cities are not frequently co-mentioned, but when one city is mentioned, the other is likely to be mentioned as well. The relationship Katowice-Wroclaw falls into this category. There are two potential explanations for this occurrence. Firstly, their co-occurrence may be confined to specific texts, which are only a fraction of the overall text database. Secondly, the infrequent co-appearance of these cities might be just coincidental. To discern whether the co-occurrence of Katowice and Wroclaw is statistically significant requires a thorough analysis of the contextual data.

High mutual information, high confidence: This represents relationships that are both frequently mentioned and exhibit a strong, statistically significant contextual connection. Lisbon and Porto is an example. As Portugal’s largest and second-largest cities, their high co-occurrence is not solely due to their individual prominence, but also due to their shared characteristics and mutual dependence, reflecting a robust and meaningful relationship across various contexts.

4. Conclusion

The study underscores the inadequacies of frequency-based measures for accurately measuring the strength of city relationships. In place of this, it advocates for a collocation-based analytical approach, which takes into account both the likelihood of cities' interdependence and the statistical confidence that these occurrences extend beyond mere chance.

An empirical comparative analysis supports this methodological shift. The study reveals that frequent mentions of major cities like London and Paris in various texts are often attributed to their individual prominence rather than a strong inter-city relationship. In contrast, the co-mentions of second-tier cities such as Lisbon and Porto suggest a higher degree of interdependence. These findings highlight the need for a more nuanced approach to studying city interactions.