

TPI MCA Innovation Quotient: using web scrapping and machine learning to estimate companies' and industries' likelihood of engagement with innovation practices

By Fatima Garcia Elena (TPI and The Data City) and Raquel Ortega-Argiles (TPI, The University of Manchester)

Practices like “innovation” are challenging to define using traditional sector data, as they often focus on productivity and business performance-related metrics. This context has driven research organisations to explore other ways to estimate a firm's or a sector's participation in innovation practices. Data science methods like web crawling and machine learning have been successfully applied to better understand an actor's or a sector's characteristics, highlighting the relevance of adopting new technologies to widen the available tools and metrics. The Productivity Data Lab has harnessed the power of these technologies to create a new indicator of innovation activity for the [Mayoral Combined Authority \(MCA\) Scorecards](#). The TPI Innovation Quotient indicates whether there is a significant concentration of companies using innovation-related language on their websites in an MCA compared to the UK. By providing a way to understand how regions score with the UK and among themselves, the TPI Productivity Data Lab further contributes to public and private efforts to diversify open data on the UK economy.

Data and Sources

The input data used to compute the Innovation Quotient is sourced from [The Data City](#). The Data City is a platform-as-a-service and data-as-a-service company that provides company data for UK companies. They match Company Registration Numbers as per Companies House with web domains, creating a library of company website text for approximately 1.6 million UK companies, which is combined with other company information data. The Data City has also developed a machine learning algorithm that makes it possible to classify companies using website text data following different approaches.

The Data City has used website text data and machine learning to create a likelihood measure of a company's innovation: the innovation score. They trained the machine learning algorithm with the website text of companies with high expenditures on R&D and produced a language model that defines the common, over-indexing language across all innovative companies. The resulting language model was used to score all companies' websites and identify those that shared linguistic patterns.

First, companies were classified as either innovative or non-innovative. Then and from those that have been classified as innovative, companies receive a scoring that represents how likely they are to be innovative. The scoring recognises that some companies share more linguistic patterns with the model than others, and those whose language aligns more closely with the model are more likely to be innovative. For the calculation of the Innovation Quotient, the Data Lab decided to use all companies with an Innovation Score of 1, meaning that all companies classified as innovative will be considered in the analysis.

Transforming company microdata on innovation into a sector-wide indicator is possible because The Data City provides both Innovation Scores and sectoral classification information for companies. These two data points for all companies made it possible to compute the Innovation Quotient featured in the MCA Scorecards.

Method

The calculation of the Innovation Quotient follows a location quotient approach. The metric indicates whether an MCA has a higher, lower, or the same concentration of innovative companies than the UK. Location quotients are an optimal strategy to convert microdata into region and sector-wide indicators as they negotiate great discrepancies between different groups' absolute demographics, like the absolute number of companies in a region. The indicator is calculated as follows:

$$LQ_{Innovation} = \frac{\frac{CIS, MCA}{CURL, MCA}}{\frac{CIS, UK}{CURL, UK}}$$

Where

CIS,MCA: Number of companies with one innovation star scoring in a MCA.

CURL,MCA: Number of companies with a URL matched in the same SIC section.

CIS,UK: Number of companies with one innovation star scoring in the UK.

CURL, UK: Number of companies with a URL matched in the UK.

It is important to highlight a few points about this index. First and since the input data consists of the company-level innovation score produced by The Data City and this is calculated by analysing companies' website text, the calculation of the location quotient will not consider the total number of companies in a region but the total number of companies in with a company-domain match in the region. This way, the calculation of the Innovation Quotient mirrors the Innovation Score's approach.

Second, all companies flagged as innovative (one-star rating) are considered in the analysis. Since Innovation Scores look for the over-indexing of innovation-related language, larger companies that engage in more activities and hence have larger website text tend to have lower innovation score values. On the other hand, smaller innovative companies often have higher innovation score values because their website text is more concise and contains a higher density of innovation language. Therefore, and in order not to exclude companies due to their website structure and activity, all companies flagged as innovative are taken into account.

Third, one company can have trading addresses in more than one region. In this case, we have calculated MCA's Innovation Quotient considering all trading addresses, meaning that all trading addresses for a company were included in counting the number of innovative companies in a region. This is important because key, innovative companies like Rolls Royce and Airbus have locations in more than one region, and therefore their innovation practices reach further than their registered address.

Impact

This indicator has the potential to disrupt existing ways of understanding research and innovation practices at large scales. The current availability of official data related to research and innovation does not facilitate structural, comparative analyses of innovative practices. Hence, the Innovation Quotient can be seen as a measure to fill the gap that makes this type of analysis possible. Considering the political and economic leverage of innovation practices in industrial settings, the TPI Productivity Lab considered it imperative to provide an open-source solution enabling better decision-making.