

TPI MCA Digitalisation Quotient: using web scrapping and machine learning to estimate the prevalence of digital companies

Fatima Garcia Elena and Raquel Ortega-Argiles

TPI Productivity Lsb and The Data City

Defining “digitalisation” through traditional sector data proves difficult, as such data often prioritises productivity and business performance metrics. To address this limitation, research organisations have turned to innovative approaches for assessing how firms or sectors engage in digitalisation. The TPI Productivity Data Lab has developed a novel measure for the [Mayoral Combined Authority \(MCA\) Scorecards](#): the Digitalisation Quotient. This indicator identifies areas with a notable concentration of digital companies compared to the UK average. The TPI Productivity Lab supports broader efforts to enrich open data and advance understanding of the UK’s digital economy by enabling comparisons across regions and providing a clearer picture of digital activity.

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 into industry sectors using website text data following different approaches.

The Data City has coined this methodology and the output data [Real-Time Industrial Classifications \(RTICs\)](#). In short, RTICs are developed by training the machine learning algorithm with sets of company websites that represent industry verticals. The algorithm creates a model that defines the training set companies’ shared language and scores the rest of the company websites against it. This makes it possible to find all the companies that describe their activity similarly to those in the training set. Then, we selected a list of RTICs that map the digital ecosystem in the UK:

Artificial Intelligence, Cryptocurrency Economy, Cyber, Data Intermediaries, Design and Modelling Tech, Digital Creative Industries, E-commerce, Fintech, Gaming, Immersive Technologies, Internet of Things, SaaS, Software Development, Streaming Economy

Method

The Digitalisation Quotient is calculated using a location quotient approach. The metric indicates whether an MCA has a higher, lower, or the same concentration of digital companies as the UK. Location quotients are optimal for converting microdata into region—and sector-wide indicators as they negotiate great discrepancies between different groups’ absolute demographics, like the absolute number of companies in an area. The indicator is calculated as follows.

$$LQ_{Digitalisation} = \frac{\frac{CDIG, MCA}{CURL, MCA}}{\frac{CDIG, UK}{CURL, UK}}$$

Where

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

CURL,MCA: Number of companies with a URL matched in an MCA.

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, since the input data consists of the company-level classification produced by The Data City and relies on the availability of company website text data, the calculation of the location quotient will not consider the total number of companies in a region but the total number of companies with a company-domain match in the region.

Second, one company can have trading addresses in more than one region. In this case, we have calculated MCA's Digitalisation Quotient considering all trading addresses, meaning that all trading addresses for a company were included in counting the number of digital companies in a region. This is important because key digital companies have locations in more than one region, so their innovation practices reach further than their registered address.

Impact

This indicator could disrupt existing ways of investigating the prevalence of digital companies on a large scale. The current availability of official data related to research and innovation does not facilitate structural, comparative analyses of the digital ecosystem. The Digitalisation Quotient is a measure that fills the gap that makes this analysis possible. Digital technologies are a key part of current production processes, and it is critical to investigate the geography of digital companies.