Technical Annex

NPIER Capabilities: Definition and Data Development

1. Introduction

This annex sets out the work undertaken by The Data City and Transport for the North (TfN) to examine the performance of the sectors within the North's economy. In collaboration with Cambridge Econometrics (CE), this work has explored the 24 sectors that make up the economy of the North, providing insight into the location, size, performance and specific activity of businesses within each sector.

A key part of TfN's evidence base is the Northern Powerhouse Independent Economic Review (NPIER). Published in 2016, the NPIER represented a unique collaboration with TfN partners, the North's Local Enterprise Partnerships (LEPs) and central government. As the first pan-Northern economic assessment, it established key economic indicators of the North, the Prime and Enabling Capabilities, and produced long-term economic projections that quantified the impact of closing the productivity gap between the North and the rest of England. The NPIER produced was the result of the extensive collaborative input and discussions between TfN and its partners.

The need for this work arose from research undertaken within TfN in Summer 2021. TfN sought to understand how the Prime and Enabling Capabilities, set out in 2016, had performed and adapted in the wake of major structural economic changes such as the vote to leave the EU and the Covid-19 pandemic over the past 5 years. However, it soon became clear that the current approach of using SIC codes did not provide a sufficient answer, and there were further limitations due to a lack of regionally disaggregated data. TfN understood that SIC codes are somewhat inflexible to the fast-paced changes to innovative sectors which are represented in the Prime and Enabling Capabilities. To this end, TfN commissioned The Data City[1] to provide a more comprehensive understanding of the Northern economic landscape, and how different sectors within the economy had adapted over recent years. The Data City were able to do this through their innovative Real Time Industrial Classifications (RTICs), which classifies businesses using language patterns within the website text of individual companies to understand their key activities and operations.

The work undertaken by The Data City has then fed into CE's work *NPIER: Capabilities, Local Data and Narratives research Workstream 1: Reviewing the North's Capabilities*. CE's analysis gives an overview of the Prime and Enabling capabilities using a traditional approach using ONS and other government data sources, and then extends this analysis using the Data City RTIC analysis.

[1] The Data City seek to map the UK's emerging economy and provide real time data on dynamic sectors and the companies within them.

2. Approach and Methodology

In order to align the NPIER Capabilities: Definition and Data Development work with the NPIER: Capabilities, Local Data and Narratives work programme, CE provided The Data City with SIC code definitions for 24 sectors spanning the UK economy.

Of these 24 sectors, 11 were able to be constructed using SIC data, as their descriptions aligned closely with those of specific groups of SIC codes (Method 1). A further eight sectors were constructed by adapting existing RTICs with the Data City's Data Explorer platform (Method 2), and the final five sectors were modelled as brand new RTICs (Method 3). The process for creating a new RTIC begins with creating taxonomies: these are frameworks that identify different pockets of activity (industry verticals) that share language patterns within a sector. Then, TfN and The Data City identified a series of keywords that represent the activities of each industry vertical. These keywords are central to the data-building process: they are used to find example companies for each vertical to train the machine learning algorithm. The platform then finds all companies that use similar language and groups them in individual datasets. The output is a list of companies that are similar to the example companies and use the keywords identified in the taxonomy. A summary of how each sector was constructed in this mixed methods approach can be found in Table One.

The data provided by The Data City was reviewed by both TfN and CE and has been embedded within CE's Workstream 1 report: *Reviewing the North's Capabilities*.

The Data City have built a unique platform, the Data Explorer, that combines company-level financial data from Companies House and CreditSafe with up to 75 pages of website text per company. This is then made usable through a Machine Learning classification process, which generates RTICs.

The Data City's RTIC classification method differs from the traditional hierarchical SIC classification approach as SIC codes are typically inflexible to the way in which modern businesses operate: nearly one-third of UK businesses are classified under a SIC code with a name containing the word 'Other'[2]. Furthermore, companies are only required to provide SIC codes for their activity at incorporation, and are not required to update them. This means that many for many companies, their SIC codes are not fully reflective of the activities they undertake as they diversify their operations. Both of these factors contribute to a landscape where, at least in the more innovative and emerging sectors of the economy, SIC codes are a poor indicator of actual business activity, meaning that key information would be missed by relying on a SIC code approach alone. Where SIC codes have been used in this analysis, TfN and The Data City have used The Data Explorer platform to verify that the SIC codes provide an accurate reflection of the activities going on within each sector.

However, it is also important to acknowledge the drawbacks of the Data City approach. For example, the RTICs are constructed using information from company websites, but not every company in the UK has a website. The UK business base consists of some five million companies, and all of these are available for analysis within the Data Explorer. The

[2] Office for National Statistics (2021), *UK business: activity, size and location 2021*, October 2021 [Accessible here]

Data City has matched 1.65 million companies to a website, and these are the companies that are included in the RTICs.

Table 1: Sectors by Construction Process

Method 1: SIC Data	Method 2: Existing RTIC	Method 3: New RTIC
Extraction and Processing of Basic Materials	Agriculture, Food and Drink Manufacturing, and Land Management	Advanced Manufacturing
Transport Equipment Manufacturing	Energy Generation and Storage	Electronics Manufacturing
International Transport; Ports and Airports	Pharmaceuticals and Advanced Chemical Products Manufacturing	Media and Publishing
Specialist Wholesale and Retail	Advanced or Offsite Construction	Research and Consulting – Physical Sciences and Engineering
Arts and Recreation	Specialist Logistics and Warehousing	Business Support Services
Accommodation and Hospitality	Software Development and Publishing	
Textiles, Clothing, Furniture and Traditional Manufacturing	Finance and Insurance	
Computing and Communications Technology Research, Consulting and Services	Life Sciences, Medicine and Human Health Research, Consulting and Services	
Higher and Further Education		
Real Estate Representation, Legal and Accounting		
Economics, Management and Social Sciences Research, Consulting and Services		

3. Links to the NPIER: Capabilities, Local Data, and Narratives Commission

Concurrent to the work presented here, TfN has commissioned CE and SQW to undertake a review of the original NPIER. This includes a data-driven review and appraisal of the North's key sector capabilities, including the NPIER's four Prime and three Enabling capabilities. As with the work presented here, the findings from this project will also support TfN's evidence base and analytical framework moving forwards.

Due to the related nature of the two projects, CE has worked closely and collaboratively with TfN and The Data City to ensure synergies and complementary outcomes between the two projects. In addition to this, both CE and The Data City have acted as a 'critical friend' for the other, providing constructive feedback and discussion to ensure the highest quality data and evidence.

Though CE's review has also prioritised and ensured a 'best-fit' approach to the use of SIC codes, CE is aware that they can still be somewhat inflexible to the fast-paced changes to innovative sectors which are included in the Prime and Enabling capabilities. Working with The Data City and their Data Explorer platform, CE has been able to define and explore additional, novel data for the Prime and Enabling capabilities.

This has a two-fold purpose: firstly, it allows CE to sense-check its assessment and classification of the Prime and Enabling capabilities in their review, particularly in terms of non-SIC code measures of economic specialisation, productivity, and growth. Second, it allows CE to consider emerging trends and innovation capabilities related to the Prime and Enabling capabilities that are not captured by conventional data or SIC codes, particularly in terms of three key modernising drivers: decarbonisation, automation and digitisation.

4. Project Outputs

The Data City's Data Explorer tool provides a comprehensive overview of sector performance. The platform's Compare tool allows users to compare the performance of sectors within different geographies: for the purpose of the below example, analysis has been undertaken on the 11 LEPs within the North compared to non-Northern LEPs. However, it is possible to disaggregate the data to a range of specific geographies, including a combination of LEPs, local authorities, cities, NUTS1 regions, and parliamentary constituencies.

The Data City Compare Tool

The Compare tool provides an overview of the RTIC being considered and the following information:

1. A Comparison Summary
2. Business counts by various geographical breakdowns
3. Employees by Local Authority
4. Sector keyword enrichment
5. Innovation keyword enrichment
6. Company size by employees
7. SIC Counts
8. Cumulative growth of companies
9. Net worth and Turnover

The following example demonstrates the data available on the compare tool and examines the Energy Generation and Storage industry, which comprises two RTICs: Energy Generation and Energy Storage. All of the following data has been taken from this RTIC only.

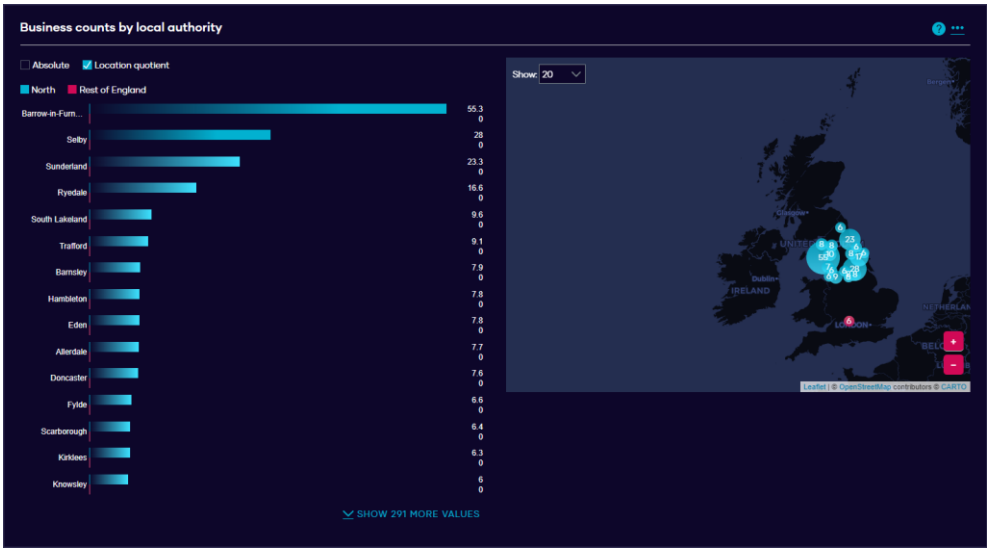
1. Comparison Summary

As can be seen below in Figure 1, the North is home to 22% of England's businesses within this industry, 40% of England's turnover within Energy Generation and Storage, and 30% of England's employees within the industry.

Figure 1- Comparison Summary

Figure 2 - Absolute Business Counts by Local Authority

Figure 3 - Relative Business Counts by Local Authority



2. Business counts by various geographical breakdowns

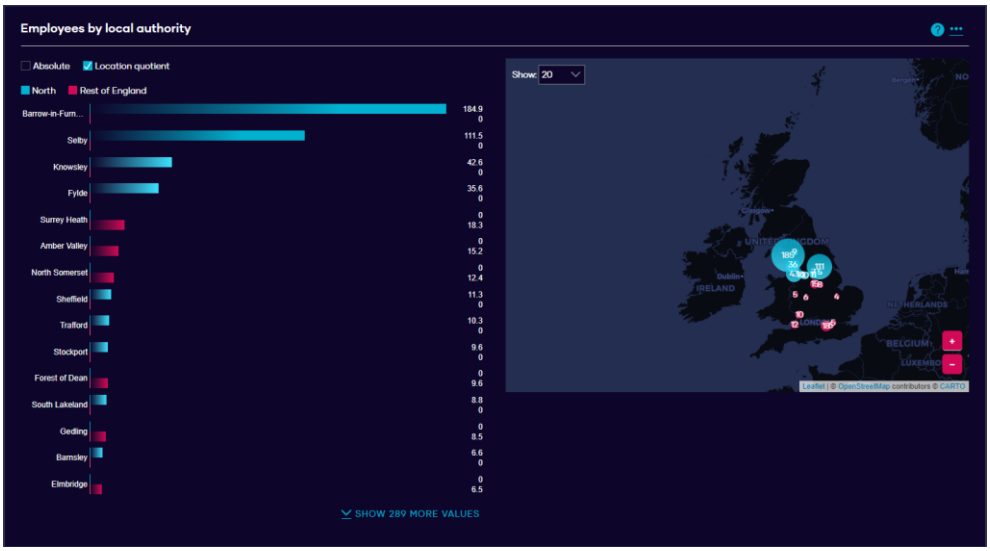
This data provides the number of businesses in each locality, including local authority, city, LEP, NUTS1 region and parliamentary constituency. The Data Explorer provides figures for each area, and also maps them. This information is available for both absolute figures and on a location quotient basis which provides the data relative to the rest of the UK. In the Energy Generation and Storage sector, we can see from Figure 2 and Figure 3 that the North is home to three of the fifteen areas with the largest absolute numbers of businesses, but is particularly dominant in the sector after controlling for the concentration of businesses in the areas of the north relative to the rest of the UK. Location quotient compares the proportion of businesses or employees in a given sector within an area to the proportion of businesses or employees in the same sector across the UK. For example, where the location quotient business count in Selby is 28 (as per Figure 3), the Energy Generation and Storage industry is 28 times more concentrated in Selby than in a typical local authority.

3. Employees by local authority

Data on the number of employees within the sector are only spatially disaggregated at a local authority level, unlike the number of businesses. The data is similarly available in absolute and location quotient terms, and is mapped within the Data Explorer. In the North, the data shows that Selby, Barrow-in-Furness, and Knowsley are employment hotspots within the Energy Generation and Storage sector.

Figure 4 - Absolute Employees by Local Authority

Figure 5 - Relative Employees by Local Authority



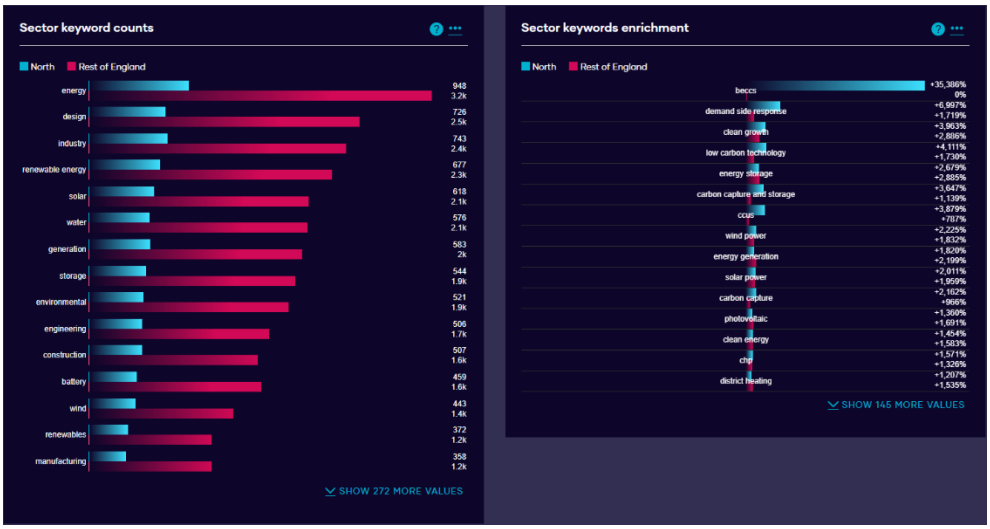
4. Sector keyword enrichment

During the machine learning process, The Data City analyse text from up to 75 web pages per company in order to determine whether or not a company belongs within a given RTIC. Through this process, The Data City obtain sector keywords that define the key activities within a sector. These keywords are in addition to the keyword lists developed as part of the Taxonomy stage, and so are led by the findings of the machine learning process, rather than being predetermined by the inputs to the process.

The sector keyword counts show the number of companies who use each word on their website, and these tend to be rather descriptive. For example, Figure 6 shows that 948 of the North's 1,012 businesses within the Energy Generation and Storage sector use the word 'energy' on their website.

The sector keyword enrichment, on the right-hand side, shows keywords that are over-represented and under-represented among companies within the Energy Generation and Storage sector compared to the average UK company. For example, the below shows that Energy Generation and Storage companies in the North are 35,386% more likely to mention 'BECCS' (Bio-energy with carbon capture and storage) than the average UK company. The fact that Energy Generation and Storage companies in the rest of England are no more likely to mention BECCS than the average UK firm means that this is a particular strength within the North. 'Clean growth' and 'low carbon technology' are also relative strengths for the North within the Energy Generation and Storage sector.

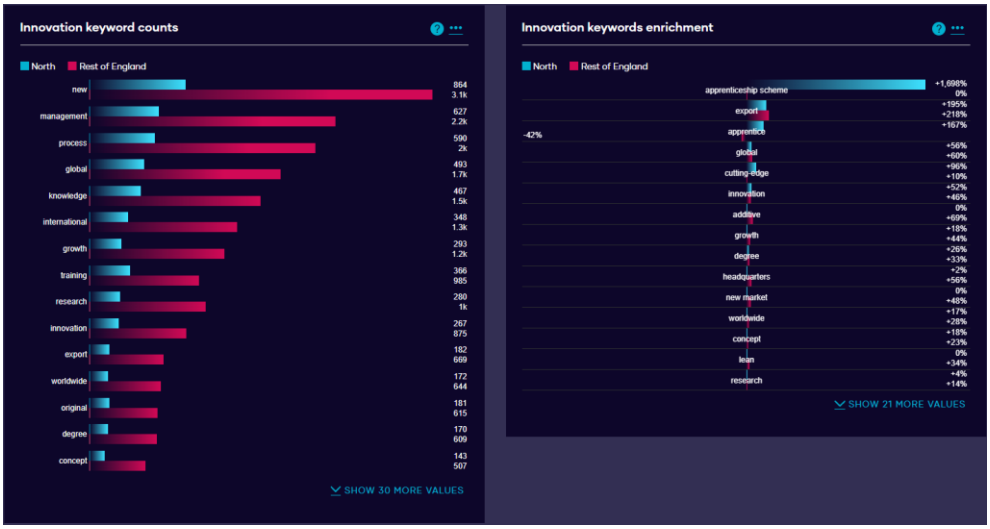
Figure 6 - Sector Keyword Counts and Sector Keyword Enrichment



5. Innovation keyword enrichment

The Data City also maintain a keyword list that identifies whether a firm may or may not be innovative. This works in a similar way to the sector keywords, in that they are identified during the machine learning process. Focussing on the innovation keywords enrichment, which shows the sector's strengths compared to the average UK firm, the data shows that the phrase 'cutting edge' is over-represented on the websites of Northern firms within the Energy Generation and Storage sector. This indicates that Northern firms within the sector perceive themselves to be highly innovative.

Figure 7 - Innovation Keyword Counts and Innovation Keyword Counts



6. Company size by employees

The Data Explorer can also provide information on the size of companies by employees, and how the number of employees within a sector has changed over time. From Figure 8 below, the Data Explorer shows that the three largest employers within the Energy Generation and Storage sector within the North employ around one-third (6,900 of

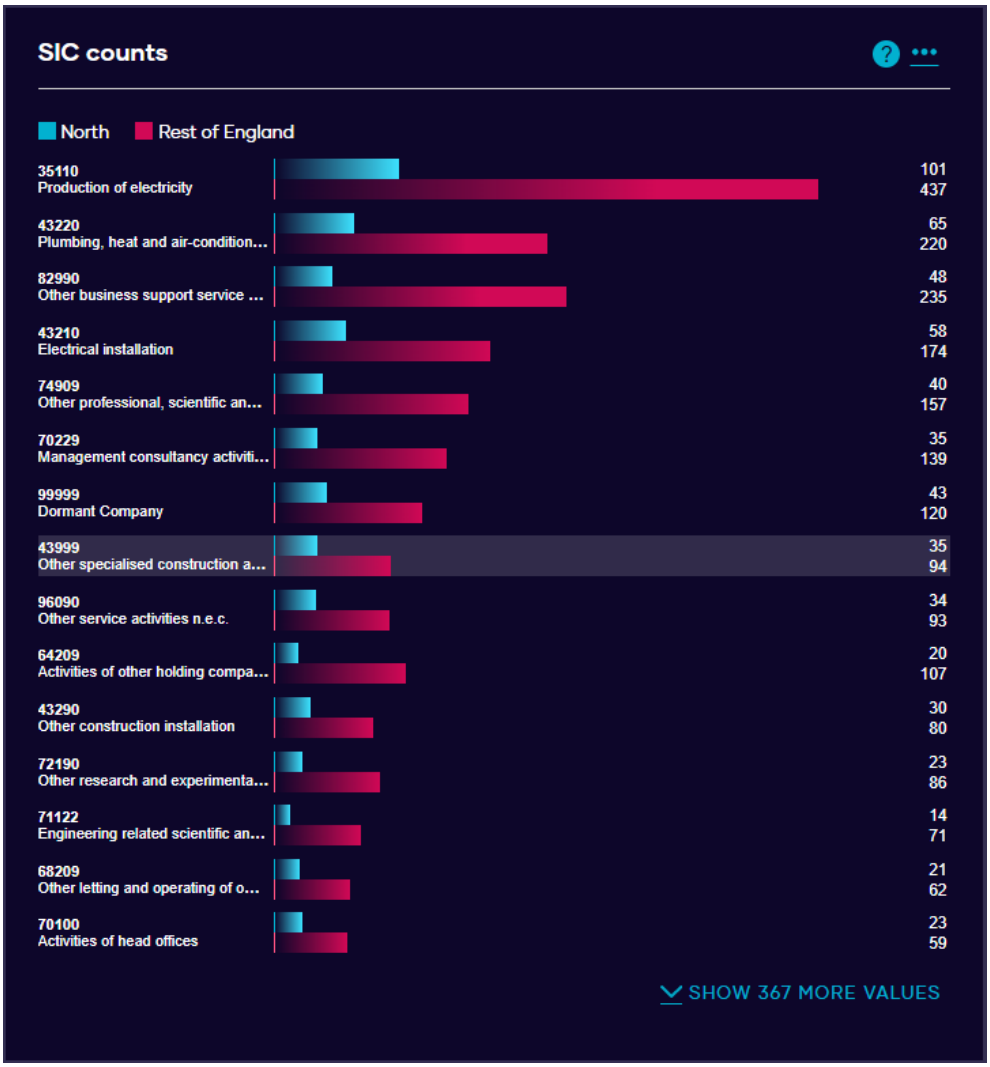
19,600) of those within the sector in the North. The Data Explorer shows projected figures for previous years where information has not yet been provided by companies that they know were in the industry at that time. The projections are made by extrapolating the changes observed within the firms that they do have data for in that year, and also in previous years.

Figure 8 - Company Size by Employees and Number of Employees

7. SIC Counts

The RTICs developed by The Data City can also be mapped back to the SIC codes that each firm has reported. This information is relevant to understand how firms would otherwise have been categorised without RTICs. In the 2016 NPIER, the Energy capability was defined by the total number of firms across 33 different 5-digit SIC codes. In this work undertaken by The Data City, Figure 9 shows that the Energy Generation and Storage sector has been defined by a range of businesses spanning 382 different 5-digit SIC codes. This clearly demonstrates the limitations of the SIC code approach, and how The Data City's machine learning approach addresses these limitations.

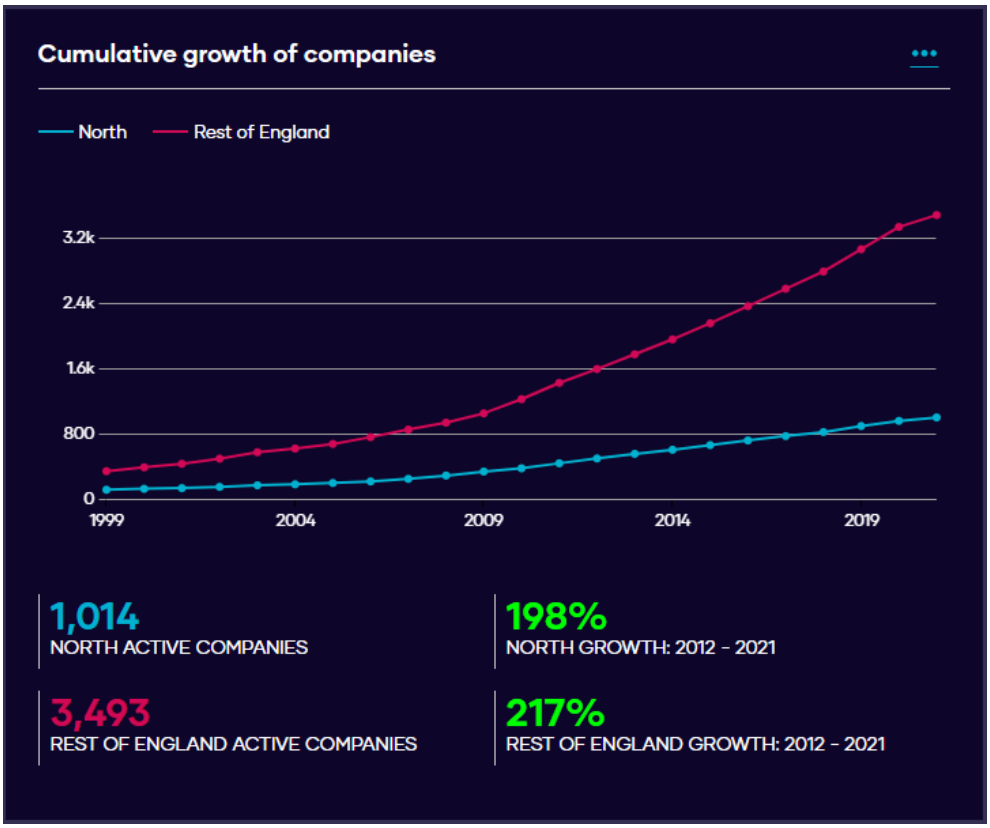
Figure 9 - SIC Counts within the Energy Generation and Storage Sector



8. Cumulative growth of companies

The Data Explorer also provides an illustration of the change in business counts across the two geographies within the sector. Figure 10 shows that the number of active companies within the North in the Energy Generation and Storage sector has grown by 198% between 2012 and 2021, slightly below the growth rate across the rest of England.

Figure 10 - Cumulative Growth of Companies



9. Net worth and Turnover

The Data City are able to provide financial data on firms through CreditSafe, who collate credit information on companies around the world. As shown by Figure 11, the Energy Generation and Storage sector had a net worth of £10.9 billion to the North in 2020, which is around 35% of the total Energy Generation and Storage sector in the whole of England.

Figure 11 - Net Worth and Turnover

5. Next Steps

This project completed in March 2022, but TfN retain access to the Data Explorer tool until November 2022. In this time, it is anticipated that the data will be utilised in the 2022/23 NPIER workstreams, providing further insight into the sectors of the Northern economy. One way in which the data may evolve over this period is the emerging impact of the Covid-19 pandemic across the North, as more companies return information on their 2021 performance.

TfN are also currently scoping the possibility of using the Data Explorer to expand our understanding of highly innovative firms in the North: what are the most innovative firms in the North doing? What differentiates them from less innovative firms in the North, and the most innovative firms in the rest of England?