Data Explained

Spatial and sectoral analysis of NI trade from 2014 to 2020

Author: Anne Marie Ward, Stuart Henderson, Esmond Birnie, Thomas Bell and Fred Booth

Ulster University

Date: November 2023

This project forms part of the beta testing of the Northern Ireland Statistics and Research Agency (NISRA) de-identified Business Data for Research (BDR) database. It was carried out in advance of the data being made available for wider use by approved researchers within the Administrative Data Research Centre Northern Ireland (ADRC NI).

The project was an exploratory study of the useability of the Northern Ireland Annual Business Enquiry (NIABI) and Broad Economy Sales and Exports Statistics (BESES) datasets. These datasets contain data that assists in assessing the performance and trade behaviour of Northern Irish businesses in their local operating environment and in international markets. The data can also facilitate insights into issues such as the performance of the 10X priority sectors[1], the role of

[1] DfE 2021, "A 10X Economy- A Summary of the Economic Vision for a Decade of Innovation".

geography in business success, and the impact of Covid-19 and the UK's exit from the European Union (EU).

Initial research questions

The study addressed several key research questions:

1. Is the data provided by NISRA (BESES and NIABI) useable for researching the performance and trading behaviour of Northern Ireland businesses?
2. How has the relative performance and trading behaviour of business sectors in Northern Ireland changed over the period 2014–2020, with particular emphasis on the 10X priority clusters?
3. Are there regional differences in the relative performance and trajectory of business sectors in Northern Ireland over the period 2014–2020, with particular emphasis on the 10X priority clusters?
4. Is it possible to design a model for Northern Ireland trade distinct from the United Kingdom?

Dataset and key variables

The project used an extract of data from the Northern Ireland Annual Business Inquiry (NIABI) 2014–2020, including the Broad Economy Sales and Exports Statistics (BESES). Download a table of the variables used:

ADR 262 variables list_RAP.xlsx

Data limitations encountered

Sampling limitations: Since not all businesses are surveyed in the NIABI, there is the potential for sampling error. The sample is taken from the Inter Departmental Business Register (IDBR). NIABI covers most of the economy in Northern Ireland, but does not include:[2]

- public sector activities (for the most part)
- public administration and defence (section O)
- farming (groups 01.1, 01.2, 01.3, 01.4 and 01.5 within section A)
- local authority and central government bodies in education (section P)
- human health and social work activities (section Q)
- medical and dental practice activities (Section Q, 86.2)
- financial and insurance activities (section K)

[2] As indicated in: https://www.nisra.gov.uk/statistics/annual-business-inquiry/abi-sample-coverage and https://www.nisra.gov.uk/system/files/statistics/NI-Annual-Business-Inquiry-Reporting-Unit-2021.pdf.

- businesses that are not registered for either Pay As You Earn (PAYE) or Value Added Tax (VAT).

NIABI also faces general issues common to survey methods such as respondent bias, human error and omission error.

Gaps and imputation: The dataset is presented with no missing data. Gaps in the data would be expected due to incomplete returns. This limits the possibility of imputing data based on previous returns as there is no indication of where a business has been selected but has not responded.

As non-responding businesses are not identified as such or included in the dataset, it is impossible to know whether a business has been selected in a particular year but has not completed the return. For example, data for business A is included for 2015, 2017 and 2019. If researchers knew that the business was selected for 2014, 2016, 2018 and 2020, but did not respond, then figures could be imputed for these years. As is, it may be assumed (rightly or wrongly) that the business was surveyed in 2016 and 2018 and did not respond, but no assumption can be made regarding 2014 or 2020 (the business may have entered and exited the market). It would be beneficial if the dataset explicitly identified business start-ups and failures.

There is also no clarification as to whether the data presented is actual returned data or imputed. It would be more transparent if imputed data could be marked as such and allow researchers to recognise where data has been imputed for a given business. It would enable researchers to decipher the extent of imputation and to determine whether imputed numbers are being used in subsequent imputation calculations.

It seems that zeros are entered when a business does not complete an entry in the survey. In many instances, a zero may be the appropriate entry. However, our investigations suggest that, in some instances, a zero entry is incorrect: for example, variables like total turnover (WQ550) and employment (WQ059). This is problematic for making inferences about businesses, sectors and regions, and may influence the weighting allocation calculation. A basic solution may be to require the business to enter a value or an explanation before permitting progress in the survey, when a zero is an unlikely entry (for example, total turnover WQ550).

Weightings: The dataset includes precalculated weightings to be used in grossing up to Northern Ireland totals. These weights are set to one if the returned data is considered to be an outlier, with all other weights being recalculated. Analysis of these weights shows that up to 32.2% (2017) of observations have been marked as outliers, with an average over the 2014–2020 period of 23.5%. With no means to check or recalculate these weights, they may be subject to suspicion by data users.

A fundamental determinant of the weighting process for size is employment numbers. Our investigations raise concerns about the accuracy of WQ059 (employment numbers). Of the 44,864 original business-by-year returns, 1,781 reported zero employment. Analysis of the outliers for turnover per employment and gross value added (GVA) per employment suggests underreporting of employment numbers. Indeed, we ended up removing the extreme observations.

Particular attention should be directed to ensuring accuracy in the recording of this variable at both the input and checking phases of the data collection process. This variable plays a major role in weighting values and is also used to derive several performance metrics, such as GVA per employment or turnover per employment.

Regional analysis: As NIABI/BESES data is provided at reporting unit level, returned data for large businesses is attributed to one location (normally their headquarters). This distorts the overall analysis of geographic effects on trade. It would be beneficial if these businesses provided data at the regional level to enable the total values to be apportioned by, for example, employee numbers or sales value per region. In addition, for 5,585 observations out of the original 44,864, location is unknown. To reduce the potential for returns to be categorised as unknown, the package used to collect the data should be pre-programmed to recognise Northern Ireland postcodes and to prompt the business to enter the postcode correctly if it is not recognisable by the system. The survey should capture when businesses move location and identify this in the dataset.

Covid-19: Subsidies (WQ414) reported in 2020 increased dramatically, suggesting that Covid-19 related payouts were not distinguished from trade subsidies when collecting data in that year.

Negative values: Some values should not be negative at source, for example, stock. The system used to input the data should be designed to flag unusual negative values and to seek reassurance, and perhaps an explanation for a negative number when this arises. In a similar vein, some of the variables are computed by NISRA such as GVA. GVA can be negative, however, a negative value may also indicate an error in the completion of the return. To provide a check on the relevant entries, the online survey should be designed to calculate the GVA values when the source information is being input. It should seek reassurance from the person completing the form, with knowledge of the business, that if a negative GVA is being implied this is a reasonable outcome.

Range of variables: Analysis of trade performance was limited due to the exclusion of capital expenditure, and a lack of (country specific) detail on export destinations and indicators of innovative activity such as research and development. The amount of capital expenditure invested by a business would likely contribute to its growth in exports. Specific export destination data would also enable distance (a proxy for cost) to be determined. This type of data would be important for the design of a Northern Ireland trade model, especially in terms of gravity modelling. We recognise that some of these variables are available in NIABI. Their inclusion in this study would perhaps have identified further insights into the drivers of trade performance beyond that reported in our findings.

GDPR: Location data is only available at Local Government District (LGD) level for GDPR reasons. This restricts the ability to analyse the potential effects of the UK's exit from the EU on trade in border regions. The addition of Super Output Area or postcode data would allow this analysis.

Necessary modifications to initial research questions or research design

Our analysis focused on exporting when we investigated trade at the regional level due to time constraints. However, our analysis included performance and the effects of Covid-19 on performance and trade behaviour.

We had planned to use video visualisations to show movements in trade by sector over time at the regional level. However, we elected to analyse each region separately and display the area-specific patterns.

Necessary modifications to the data

Deletions for possible errors: The dataset supplied by NISRA had 44,654 firm-level observations for 16,260 businesses located in Northern Ireland covering the period 2014–2020. Our initial work involved analysing the descriptive statistics including the total, mean, standard deviation, minimum, maximum and number of zeros for each variable across time, using both the reported data for the sample of businesses and the weighted data. This process uncovered several issues with the data.

First, a visual examination of the maximum values highlighted outliers. To determine if these were errors, we sorted the data by unique identifier code over time and checked their consistency/plausibility. Where we determined data to be erroneous, we elected to:

(1) keep the observation in the dataset and substitute a zero value for that cell when the other returns reported for that business in other years were zero, or;
(2) where the business had returned data in the other years (before and after the questionable entry), use imputation to derive a value.

This imputation calculated a value for the year by adjusting the prior year's data by growth in the median value of that variable for that particular sector for that year. A check of the resultant value was undertaken relative to the period after the questionable entry to determine if the use of the median approach provided a reasonable response. If not, then the average of the entries in the period before and after the 'questionable entry' was used instead.

We also counted all zeros for each variable. We recognise it is normal for businesses to have no value for several variables, such as subsidies (WQ414), employee costs (WQ450) or certain sub-categories of sales or purchases. However, we deem it highly unlikely that total sales (WQ550) is zero. Indeed, several observations reporting zero sales but making purchases resulted in an unexpectedly high number of businesses with negative GVAs. Given the importance of sales to evaluate trade, we decided to delete these observations. This resulted in 3,450 deletions. In addition, while a business may have negative GVA, we considered it highly unlikely that several businesses would have exactly zero GVA, and so deleted these observations. This resulted in 195 deletions (WQ2030). Finally, while businesses may have no employees, all businesses should have employment as an owner must be involved to generate the trade. This is captured by variable WQ059, employment. An analysis of the variable identified a further 100 businesses with a zero value in their return. These observations were deleted.

As a result of these adjustments, the dataset used for our analyses of trade behaviour is made up of 41,036 observations. Trade behaviour is examined using ratios of sales to specific markets relative to overall sales: Great Britain (GB) sales as a percentage of total sales; Republic of Ireland (ROI) sales as a percentage of total sales; Rest of the European Union (REU) sales as a percentage of total sales; and Rest of the world (ROW) as a percentage of total sales.

A final adjustment was made when we investigated performance measured by sales per employment and GVA per employment. An analysis of the extremes of the resultant ratios identified many anomalies. Due to time constraints, and consistent with Wales (2018, 2019), we elected to remove the extreme 1% of observations from the dataset. This resulted in a further 1,368 observations being removed resulting in a final dataset with 39,668 observations. We used this dataset in the regressions that investigated the performance metrics (sales per employment and GVA per employment). A cursory examination of the extreme outliers identified that the problem seemed to arise with the WQ059 employment variable (understated) and not with the

reported sales variables. Therefore, we elected to keep the dataset with 41,036 observations in the regressions examining external trade behaviour.

Data transformation for regression analysis: Finally, even with the removal of these extremes, the distribution of the data continued to suggest errors/outliers. Therefore, to reduce the impact of these errors/outliers, we transformed the data using the hyperbolic sine function. This caters for negative and zero values, which are prevalent in the dataset and relevant for several of the variables being examined, such as GVA per employee (which may be negative), sales to the rest of the world (ROW), or subsidies received in the period (both may be zero).

Inflation adjustment: We used regional, sector-specific GDP deflators supplied by the ONS (2022) to adjust our dataset for inflation. As a result, the underlying data has been adjusted to 2019 prices. This enables us to examine changes over time after adjusting for a general rising price level.

Attempting a SIC-based definition of the DfE 10X "priority clusters": The DfE 10X economic vision for Northern Ireland,[3] as of the start of 2023, highlighted seven key priority clusters:

1. Agri-tech
2. Life and health sciences
3. Advanced manufacturing and engineering
4. Fintech/financial services
5. Software (including cyber)
6. Screen
7. Low carbon

This list of seven represents an extension and elaboration of the thinking and choice of clusters identified in earlier 10X documents. In 2021, five clusters were highlighted:[4]

1. Agri-tech
2. Life and health sciences (e.g. personalised medicine)
3. Advanced manufacturing and engineering (e.g. composites)
4. Fintech/financial services
5. Digital, ICT and creative industries (e.g. cyber security)

It seems that the final cluster in the 2021 list, Digital, ICT and creative industries, has been disaggregated into two parts (i.e. software and screen). Low carbon has also been added (although parts of it would have been contained under agri-tech and advanced manufacturing and engineering).

At present, neither the NIABI/BESES dataset nor the 10X documentation identifies individual businesses according to 10X priority clusters. In the absence of such classification, we chose to align the priority clusters to the Standard Industrial Classification (SIC) 2007 to facilitate analysis.

The task of translating from the Department for the Economy (DfE) key industries or clusters to the SIC is subjective. A particular challenge is that the SIC is a product-based approach - i.e. it classifies businesses according to what is being made or which service is being provided. By contrast, the 10X clusters are more associated with *how* goods or services are made/provided, e.g. food products with an application of technology. In some instances, technologies are specified

[3] DfE October 2022, "10X Vision- Next Steps for Implementation", and October 2022, "A 10X Economy- An Open Call for Research Proposals".
[4] DfE 2021, "A 10X Economy- A Summary of the Economic Vision for a Decade of Innovation".

with specific clusters, e.g. AI and data analytics, but this did not necessarily help with SIC alignment.[5]

To align 10X with SIC codes, we proceeded as follows:

- First, Companies House provides an online search facility whereby keywords can be entered to see if they register at any point within the SIC (this service is presumably for use by businesses as they attempt to self-identify their main activity when filling in the Annual Business Inquiry (ABI) return forms). We took the keywords used by DfE in their 2021 summary document (*A 10X Economy: A Summary of the Economic Vision for a Decade of Innovation*) and entered them (or various combinations thereof) into the search box to see which, if any, SIC codes were suggested. For example, in terms of software, we tried "cyber security", "AI", "artificial intelligence", "data analytics" etc. Some judgement was necessary when matching similar but not identical terms such as "data analytics", as it did not appear but "6310 data processing" did. We noted that very few of the exact terms used in the DfE document appear in the SIC. A small number can probably be matched up to other sector names which are reasonably close.

- Second, as a complement and addition to the first method, we went through the actual SIC, at the 4- or 5-digit level, to see if this suggested what might or might not be included within the 10X priority clusters.

Finally, there are no fintech businesses in the NIABI and BESES datasets available to us as such financial services businesses are not surveyed. A small number of businesses included in the dataset are categorised as screen and low carbon. Therefore, to enable data presentation that complies with ADRC NI rules on identification, these classifications are merged with others: screen with software and low carbon with advanced manufacturing and engineering. Our final classification has four categories: agri-tech; health and life sciences; advanced manufacturing, engineering and low carbon; and software and screen.

Defining regions of interest: A focus of our study is to perform a spatial analysis of the data. This is of interest given the changing administrative burden and physical checks imposed by the trading environment following the UK's exit from the EU.

We consider that the endowments for trade vary across Northern Ireland, e.g. proximity to Northern Irish ports and dual carriageways. A further dimension of interest is proximity to the border as this may influence the propensity to engage in cross-border trade. This is of particular interest given the UK's exit from the EU. Furthermore, Belfast has favourable endowments (such as proximity to ports, airports, universities, knowledge hubs, peer businesses and talent pool) that merit distinguishing a Belfast regional location from other business locations.

Due to GDPR concerns, data was only available at the LGD level. This restricted our ability to create regions of interest based on, for example, proximity to the border. In addition, when the data is sliced into the eleven LGD regions, the outputs breach ONS Secure Research Service (SRS)

[5] In addition to the "problem" that the DfE listing of technologies was more complete for some clusters rather than others, it should be noted that this information relates only to the five clusters identified in 2021 rather than the later and longer list of seven. Data sources, other than the NIABI, may shed further light on which businesses are engaged in the type of activities which would imply they are part of a cluster, e.g., company statements/reports on business websites. This is subject to the caveat that self-promotional statements may exaggerate technological and other performance.

minimum counts necessary for anonymity. Therefore, we created our own regions of interest, influenced by the Nomenclature of territorial units for statistics (NUTS) 3 classification.[6]

1. We merged the LGDs: North Down; Ards; and Mid and East Antrim. This was justified in terms of roughly similar distance from Belfast.[7]
2. We merged the LGDs: Fermanagh and Omagh; and Mid Ulster. Both areas interface with the border and are considered to have limited infrastructure (no port, train, public airport and limited dual carriageways/motorways).
3. We merged the LGDs: Newry, Mourne and Banbridge; and Armagh. Both areas interface with the border and are distant from Belfast but are the closest counties in Northern Ireland to Dublin port and airport. These counties are also supported by dual carriageways and the Enterprise train route (Belfast–Dublin), making travel between businesses and Belfast and Dublin easier for the physical movement of goods and people.
4. We kept separate the LGDs: Derry and Strabane; and Causeway Coast and Glens. Both are relatively far from the ports in Belfast and Dublin, however, Derry and Strabane interfaces with the border and has a public airport.

Our classification yielded eight regions which we call:

1. Belfast
2. Antrim and Newtownabbey
3. North Down and East Antrim
4. Lisburn and Castlereagh
5. South Down and Armagh
6. Causeway Coast and Glens
7. Derry City and Strabane
8. Fermanagh and Mid-Ulster

Recommendations to data owners

Larger sample: To obtain a more comprehensive dataset, it would be advantageous if more businesses were included as census businesses over all industries and employment bands. Although potentially cumbersome, those businesses willing and able to provide data should be encouraged to do so and possibly incentivised. This would enhance the availability of consecutively reported data for individual businesses over time.

Missing observations and labelling of imputed observations: The current dataset, contrary to expectations, has no missing data as the current approach seems to automatically populate uncompleted fields with zeroes. While this provides a 'clean' dataset, it would be preferable to distinguish true zeroes from missingness. In addition, there is no indication within the dataset

[6] The NUTS classification, known as nomenclature of territorial units for statistics, is an EU and UK hierarchal system for dividing up the economic territories for the purposes of collecting, developing and harmonising statistics at three levels. The NUTS 1 classification reflects major socio-economic regions, NUTS 2 reflects basic regions used for regional policy application and NUTS 3 reflects small regions that are used for specific diagnoses (Eurostat, (2023), NUTS – Nomenclature of territorial units for statistics, available at: https://ec.europa.eu/eurostat/web/nuts/background.

[7] Though it could be argued that North Down and Ards are closer to Dublin port so may behave differently; we tested for this and found no significant difference in GVA per employee between the two areas.

whether the data is returned or imputed. An additional variable with this information would enhance transparency.

Automated survey: Better use of automated functionality to check/query data (anomalies) when it is being input by businesses would potentially increase the reliability of the data and reduce errors in the published dataset.

Review: Additional resources should be allocated to checking the validity of the original entries and correcting anomalies. Use of growth and ratio trends over time should be used to highlight errors. In addition, checks to other sources such as HMRC tax data, VAT return data, annual financial statements and PAYE returns may provide assurance over underlying data entries. This resource could also prompt engagement or chasing-up missing entries.

Identification of 10X priority cluster businesses: To gauge the progress of the 10X strategy, it would be helpful if businesses that fall within each of the priority sectors were identified. This is not being done. We recognise such a classification is difficult. SIC codes focus on what the product is, not how it is made, and so distinguishing the constituent companies of 10X is problematic. Nevertheless, better identification is critical to meaningful analysis.

Make more data available: Trust the system and procedures in place to protect the identity of the underlying businesses. Researchers using the data have to become ONS SRS Accredited Researchers. This involves being trained on data presentation and use that protects the identity of data sources. In addition, the ONS SRS Statistical Support Team performs a double review process on all outputs from the data to ensure the data does not contain any disclosive material before it is cleared for use. This is a rigorous process. If more data were made available, such as postcode data, then more meaningful analysis would result.

Additional data which would help to further develop the research

Financial and insurance activities: As data is not collected/published on financial and insurance activities, there is a lack of data for the Fintech/financial services 10X priority cluster. Data on this sector should be collected and published as it is important in the policymaking context for Northern Ireland.

Farming: Farming businesses make up a large part of Northern Ireland trade, yet many are not included in the survey. Consideration of how to track and reflect (on a comparable basis to NIABI/BESES) farm trade to and from GB and ROI would provide a more comprehensive indication of trade for Northern Ireland. This data would be of interest to policymakers.

Additional variables:

10X identification: At present, businesses are allocated to industry SIC codes, typically based on a self-assessment by businesses of what their principal product or service is. A problem, however, from the point of view of evaluating aspects of policy in Northern Ireland, is that 10X priority clusters are defined rather more in terms of **how** products and services are produced (the technologies being used) rather than **what** is produced.

This implies a problem in identifying which businesses are within 10X priority areas and which are not. From the point of view of producing metrics to evaluate the implementation of the 10X strategy, the boundaries of what constitutes a 10X priority cluster[8] should be clearly defined and then used to identify businesses from within the population. As these businesses are of strategic importance, they should be surveyed each year, irrespective of size and location. In addition, as the performance of 10X businesses will be closely monitored over the coming decade, there should be clear guidance on how a business starts or stops to qualify as 10X. This means that shifts in reported statistics are not a consequence of changes to the definition, but the result of changes to trading behaviour and economic performance. Of course, we recognise that a product-based approach to standard industrial classifications is very long-established. It therefore may not be reasonable to expect ONS/NISRA to provide data consistent with the defined boundaries of the priority clusters in Northern Ireland.

We recommend an additional data entry point in the survey that gathers information on **how** products or services are produced.

Measuring innovation and technological progress: The dataset we had access to did not have an appropriate measure of innovation and technological progress. Measuring innovation and technological progress is difficult due to variations in definition. Studies examining international or inter-regional competitiveness often use expenditure on research and development as a proxy. However, this is an imperfect proxy because it is an "input" measure. Another possible input measure of innovation and technological progress is the percentage of the labour force with advanced qualifications (e.g. PhDs in science or technology). An alternative and more "output" measure of innovation and technological progress could be the number of patents issued.

Most analyses of the determinants of economic growth (growth accounting) have concluded that technological progress has been the main cause of growth in economies, and is more important than additions to the labour force or capital stock. Given this context, we recommend that additional data on the levels of spending on research and development at the business level be collected. Information on this relationship would be of interest to policymakers who could design interventions (such as tax credits or enhanced capital allowances or grants) to promote research and development in priority sectors.

Other data for researchers: There are several areas of supplementary data that would be helpful for researchers in the analysis of trade, including:

- financial-type indicators, e.g. assets to give an indication of business size and capital expenditure to give an indication of investment propensity.
- whether a business has started or has failed.
- information to enable head office combined data to be apportioned to regions (employment numbers or sales figures or proportions per predefined region).
- an indicator of when a business changes regional location, so it is not confused with growth in current business activity.
- it would be beneficial to have HMRC-matched data for certain key variables to provide some confidence on the weighting process (assuming HMRC has data for the population, such as total sales).
- data at postcode level. This would enable more refined analysis of trade, e.g. activity of businesses operating close to the Northern Ireland-Republic of Ireland border. Additional checks to ensure anonymity would be required but we recommend leaving this to the ONS

[8] Agri-tech; life and health sciences; advanced manufacturing and engineering; fintech/ financial services; software (including cyber); screen; low carbon.

SRS trained researcher. The ONS SRS output audit checks are diligent and from our experience are unlikely to allow any issues to pass through.
- social data, e.g. access to data on deprivation.
- an extended dataset to include 2021 and 2022. This (when available) should provide better evidence on the impact of Covid-19 and the UK's exit from the EU.
- detailed data on export destination country, export duty, and distance would help with the design of a trade model for Northern Ireland.
- separate identification of one-off interventions, such as Covid-19 type supports (subsidies).

Please include code files used in your analysis

Please contact NISRA Research Support Unit: rsu@nisra.gov.uk for code files.

Feedback on metadata, synthetic data and other documentation provided

Supporting documents: NISRA staff have been helpful in clarifying the procedures used in the collection and analysis of the NIABI/BESES data and links to other information were provided within the ONS SRS. However, the provision of more comprehensive written documentation would add to the usefulness of the data.

We recommend:

- A dedicated support webpage containing links to documents that explain clearly how the sample is selected each year, the response rate, checks carried out on the data to that point, and summary descriptive statistics on the returns received.
- A link to a data dictionary that defines the individual variables included within the dataset and how they have been determined (if calculated from the data).
- A detailed summary of the processes involved in calculating weights, identification of outliers and the stratification of businesses within the survey. This would ideally provide access to the weighting formulas adopted and the data used to calculate the weights for each year.
- Details of how the geographical apportioning of returned data from large businesses is performed. This would provide more clarity on both the overall state of the NI economy and geographic variations.
- Reviewing the online provision of comparable resources by the ONS to see if there is any learning that could be used by NISRA (to enhance the usability of the NIABI/BESES data).
- Links to pages hosting publications that have used the dataset, publications on methodology, and complementary datasets that researchers might want to use when analysing the data, such as deflation indices.

Imputed data variables ("missing data"): Some gaps in the data are expected due to incomplete returns and non-response. Our problem, as data users, is that these gaps and the nature of the gaps are not clearly identifiable. Moreover, NISRA informed us that they imputed some data for

the period post-2018. The problem this presented to us as data users is knowing whether the data as presented is actual data or had been imputed by NISRA.

Due to the way the database is presented, we encountered difficulties when trying to impute data for the period 2014–2018. Identifying prior and future returns for a single business is problematic as there is no information on whether a business has been selected but has not responded or when they are no longer in operation. Though we did attempt imputation, we considered it unreliable, and so did not use it to amend the dataset for our analyses.

We recommend that details of the imputed variables and the underlying methodology be clearly set out. With this information, data users may consider the reliability of any imputation process or, where imputation has not been undertaken by NISRA, the data user may generate imputed variables.

Any other feedback

Working with large databases that are not clean means that there can be many draft versions of potential output as anomalies are uncovered that highlight issues.

The process of extracting output for discussion means that a highly controlled approach has to be taken, which introduces lengthy delays to the process.

Acknowledgements

Administrative Data Research Northern Ireland (ADR NI) takes privacy protection very seriously. All information that directly identifies individuals will be removed from the datasets by trusted third parties, before researchers get to see it. All researchers are trained and accredited to use sensitive data safely and ethically, they will only access the data via a secure environment, and all of their findings will be vetted to ensure they adhere to the strictest confidentiality standards. The help provided by the staff of Administrative Data Research Centre Northern Ireland (ADRC NI) and the Northern Ireland Statistics and Research Agency (NISRA) Research Support Unit is acknowledged. ADR NI is funded by the Economic and Research Council (ESRC). The authors alone are responsible for the interpretation of the data and any views or opinions presented are solely those of the author and do not necessarily represent those of the ADR NI. NISRA's data has been supplied for the sole purpose of this project.

About ADR UK

ADR UK (Administrative Data Research UK) is a partnership transforming the way researchers access the UK's wealth of public sector data, to enable better informed policy decisions that improve people's lives. By linking together data held by different parts of government and facilitating safe and secure access for accredited researchers to these newly joined-up and de-identified data sets, ADR UK is creating a sustainable body of knowledge about how our society and economy function – tailored to give decision makers the answers they need to solve important policy questions. ADR UK is funded by the Economic and Social Research Council (ESRC), part of UK Research and Innovation.

Contact

Anne Marie Ward

Tel: 028 95365973

Email: am.ward@ulster.ac.uk

SFR/SB: 05/2023