International Journal of Population Data Science

Journal Website: www.ijpds.org

Data resource profile: a guide for constructing school-to-work sequence analysis trajectories using the longitudinal education outcomes (LEO) data

Shivani Sickotra[1*]

Submission History

Submitted:	08/01/2025
Accepted:	06/02/2025
Published:	25/03/2025

[1] Sheffield Methods Institute, School of Education, University of Sheffield, The Wave, 2 Whitham Road, Sheffield, S10 2AH

Abstract

Introduction
Sequence analysis is a powerful methodology for examining longitudinal school-to-work trajectories. Despite its growing use, there is limited guidance on preparing suitable datasets. This resource details the creation of a dataset specifically designed for sequence analysis, capturing yearly education and employment activity states for 556,182 individuals from England's 2010/11 school-leaver cohort.

Methods
The dataset was constructed using the Department for Education's Longitudinal Education Outcomes (LEO) data. SQL was used to extract relevant variables, and data linkage and preprocessing was performed using R. Data processing was tailored to sequence analysis, including reducing the number of activity states and applying a hierarchy to integrate education and employment data.

Results
The resulting dataset spans activities from the first non-compulsory state in 2011/12 until 2018/19, tracking trajectories from ages 16/17 to 23/24. The dataset was designed with the ability to subset school-leavers by their initial Combined Authority residence to aid in regional analysis of school-to-work trajectories. Individual-level socio-demographic characteristics that can be linked to the longitudinal activity histories were also built, alongside longitudinal geographic locations and employment earnings data. Additionally, the limitations of the developed data are discussed.

Conclusion
This resource provides crucial guidance for researchers and practitioners who may require experience preparing input datasets for sequence analysis, addressing the current gap in available resources. By offering step-by-step instructions and shared code, it empowers users to recreate or adapt the dataset for their specific research needs. Its ability to subset by region further supports localised and comparative studies of school-to-work trajectories, making it a valuable tool for advancing existing research. The LEO data can be accessed by application through the Office for National Statistics Secure Research Service.

Keywords
education data linkage; administrative data linkage; sequence analysis; longitudinal education outcomes; school-to-work; data development; data pre-processing; big data

Key features
- **Purpose and Unique Dataset Design** – This dataset is uniquely tailored for sequence analysis, addressing the lack of accessible resources and guidance on creating input data for this methodology. By providing a transparent account of its development process, it lowers barriers to conducting sequence analysis and promotes reproducibility.
- **Population and Scale** – The dataset includes activity states for 556,182 individuals from the 2010/11 school-leaver cohort in England, tracking their education and employment histories from ages 16/17 to 23/24.
- **Comprehensive Data Linkage** – Data from the Department for Education's Longitudinal Education Outcomes (LEO) database was extracted, linked, and preprocessed using SQL and R. The detailed methodology related to this is outlined.
- **Categories of Data** – Covers longitudinal school-to-work activity histories, socio-demographic characteristics, longitudinal residential geographic locations and earnings dataset creation, to support diverse research questions on educational and employment trajectories using sequence analysis.
- **Access** – Researchers can access the dataset by applying through the Office for National Statistics Secure Research Service. The code related to this data resource can be accessed at https://github.com/sickotra/Developing_SchooltoWork_Trajectories_for_Sequence_Analysis_LEO_Data.git.

*Corresponding Author:
Email Address: ssickotra1@sheffield.ac.uk (Shivani Sickotra)

https://doi.org/10.23889/ijpds.v8i6.2953

Background

Sequence analysis methods have been widely used to examine school-to-work trajectories both in the UK and internationally [1], with an increasing focus on teaching these techniques in recent years [2–4]. The typical workflow starts by creating linear sequences that specify the education or employment 'state' for each individual in the sample after they leave school. These sequences are usually recorded in monthly or yearly intervals, depending on data availability and computational capacity. Next, cluster analysis groups individuals with similar activity histories into distinct typologies. Finally, multinomial logistic regression is used to understand the socio-demographic characteristics likely to lead to these typologies [5]. A key advantage of sequence analysis is its ability to examine activity states collectively, offering a holistic view of long-term pathways rather than treating each life stage as an independent event. This allows for a deeper understanding of how life experiences interconnect over time. For a more detailed explanation of sequence analysis and its use, see [6].

The UK Longitudinal Education Outcomes (LEO) administrative dataset collates education and employment data for approximately 39 million individuals [7]. Hence, the LEO data is ideal for analysing school-to-work trajectories, particularly using sequence analysis methods once prepared appropriately. It integrates data from multiple sources, including the National Pupil Database (NPD) [8] and the Higher Education Statistics Agency (HESA), with employment and benefits data from HM Revenue and Customs (HMRC) and the Department for Work and Pensions (DWP) [7]. Analysing education and employment activities using this linked dataset offers the potential for many rich insights [9, 10].

Despite the rise in sequence analysis research and the growing emphasis on teaching the method, there is still limited guidance on preparing the input data needed for its application. Source data, whether administrative or survey-based, often requires significant preprocessing and the construction of longitudinal trajectories before sequence analysis can be conducted. This data preparation process is typically complex and time-consuming, involving numerous analytical decisions by the researcher. Indeed, the creation of longitudinal data often requires as much, if not more, time than the subsequent sequence analysis and is arguably the most critical phase as it directly influences the results.

In sequence analysis teaching materials, the preprocessing stage to link and transform the data into an appropriate format is usually overlooked and a prepared dataset is used [2]. Although this does not pose an issue for methods teaching, the datasets used do not reflect the complex nature of administrative or survey data and the level of preparation required. This then limits the accessibility of the sequence analysis method for interested researchers if an 'ideal' dataset is not immediately available. Moreover, the lack of appropriate guidance to create such a dataset could ultimately discourage the use of the technique.

Within existing sequence analysis literature [5, 10, 11], researchers are required to either summarise or omit detailed descriptions of the data creation process so a greater emphasis can be placed on the results and research implications, largely due to journal word count restrictions. Moreover, while Department for Education [12] utilised the LEO data and sequence analysis to analyse post-16 pathways, the report provided very minimal explanation of the data preparation process. The same applied for the LEO sequence analysis research by Bowyer et al. [10] on young people who experienced custody. Although Anderson and Nelson [13] did include a technical report related to their research on post-16 education and labour market pathways using LEO, the data development was not related to sequence analysis. An exemplar of producing longitudinal data has been created by Wright [14], although again this was not tailored to sequence analysis.

As a result, there remains a lack of comprehensive technical reports outlining the steps required to create a bespoke dataset specifically for sequence analysis that are both accessible to researchers learning the method or wider non-academic audiences interested in leveraging the data analysis technique. This data resource aims to address this knowledge gap by presenting an in-depth LEO data development guide specifically intended for sequence analysis research. The sequence analysis input dataset created extracted the yearly education and employment activity states for 556, 182 individuals in the 2010/11 English school-leaver cohort and linked them to form longitudinal activity histories. Individual-level socio-demographic characteristics for the regression stage in sequence analysis, as well as longitudinal geographic locations and employment earnings data were also created.

This guide serves as a blueprint for researchers in academia or industry interested in conducting longitudinal school-to-work research using sequence analysis or exploring the capabilities of the LEO data. The intention is not to teach or produce sequence analysis results, but rather to outline the process of creating a sequence analysis input dataset. It aims to enhance the accessibility and transparency of the data development process, addressing the gaps in the existing literature. To date, no detailed input data methodology for sequence analysis has been shared, making this contribution unique.

The dataset was built as part of PhD research focused on comparing post-16 trajectories in Combined Authority (CA) regions using sequence analysis. A working paper on young people facing difficult school-to-work pathways in the South Yorkshire and Greater Manchester CAs has been produced using the data. CAs are administrative areas in England designed to facilitate regional economic development [15]. CA-level analysis can be beneficial for place-based insights as English devolution is applied at this administrative level. Although there is an emphasis on CA subsets within this data resource, alternative geographic levels are also available within the developed data for wider place-based applications [16].

Methods

This data resource profile has been created to accompany the 'LEO Sequence Analysis Data Development' code file [17]. It closely follows the structure of the code and aims to explain succinctly and diagrammatically the data preprocessing undertaken to aid others in similar research pursuits.

LEO data extraction using structured query language (SQL)

The LEO data is held in a SQL database comprised of several tables and so the relevant variables from specific tables were extracted to create the bespoke sequence analysis dataset using the R programming language. All data were from the LEO Standard Extract Iteration 1 [18]. Data access for only specific variables was obtained through the Office for National Statistics Secure Research Service (ONS SRS) [19].

Table 1 provides a summary of the component tables used from the LEO data. The data created used the Spring School Census from the NPD, the National Client Caseload Information System (NCCIS) from the NPD, HESA, HMRC Employment, Self-employment and DWP Benefits tables to create a longitudinal record of the yearly activity histories for the 2010/11 (aged 15/16) school-leaver cohort in England. The CA that an individual was residing in at school-leaving age was linked to their corresponding activity history.

The cohort was initially extracted from the NPD School Census data and linked to socio-demographic characteristics. The activity histories created began from the first non-compulsory observed state in 2011/12 until the 2018/19 academic year. This corresponded to ages 16/17 to ages 23/24. A longitudinal history of employment earnings per tax year and residential geographic locations was also created covering the length of the study period. This used the LEO Employment Earnings and Geography tables. Other tables in LEO were not utilised as these were not required or approved for access.

Individual characteristics

There are 3 census collections in the NPD - Autumn, Spring and Summer. The Spring school census was used as the ethnicity and Free School Meal (FSM) variables cleared for access were only available for this collection period. Private-schooled and home-schooled data was not available.

The school census table was inner joined to the KS4 table to retrieve the variables in Table 2. In the LEO education data, one person may have multiple or no records in the LEO employment data. Therefore, bridging lookups are provided during data access which allow one-to-one linking between the education and employment tables. These are referred to as 'resolved' or 'LEO matched' [22]. When selecting the 2010/11 school-leaver cohort from the School Census, the SQL query inner joined these bridging lookups to ensure that all individuals selected had some record available within the DWP/HMRC activities data, employment data and geography data. The match rate was 98.3% (see Table 5) which aligned with the 95% average match rate indicated in the LEO User Guide [22].

The records were then filtered using the birth month and year to extract the relevant cohort. A filter to retain only pupils *on roll* (registered students) was also implemented since demographic data was missing for all pupils not on roll. In the NPD School Census, there is one 'main' record and other

Table 1: Summary of the component tables used from the LEO data

LEO data table	Summary
NPD Spring School Census	Collects detailed information on pupils in state-funded schools in England, including demographics, attendance, and academic performance. Data is gathered through termly school censuses submitted by schools to the Department for Education. The Spring term refers to a census collection between January and February.
NPD Key Stage 4 (KS4)	Contains data on student performance in Key Stage 4 assessments, including General Certificate of Secondary Education (GCSE) results. Key Stage 4 refers to the final two years of compulsory schooling in England (ages 14-16). GCSEs are a set of exams taken at age 16 in subjects like Math, Science, and English, similar to a high school diploma in other countries. Data is collected from awarding bodies and matched with prior attainment and school census records.
NPD National Client Caseload Information System (NCCIS)	Contains data on young people's participation in post-16 education and training in England. Local authorities collect and submit this information to track the education, employment, or training status of 16 to 18-year-olds.
Higher Education Statistics Agency (HESA)	Gathers comprehensive data on students enrolled in higher education institutions across the UK, covering enrolment numbers, courses, qualifications and demographics. Universities and colleges submit this data annually to HESA.
LEO: Benefits, Employment, Self-Assessment, Earnings, Geography	Integrates data from various government departments to provide insights into individuals' employment status, benefit claims, earnings, self-employment income and geographic information. This data is collected through administrative records from HMRC and DWP. Further details on specific data sources can be found in the LEO Variable Request Form [18].

Sources: [7, 20, 21].

Table 2: Unpreprocessed individual characteristic variables extracted from LEO using SQL

LEO data	Variable	LEO data variable name	Description – retrieved from (ONS, 2021)
NPD Spring_Census_2011	ID	Pupilmatchingrefanonumous_spr11	ID to link to other NPD tables and link to other LEO tables using LEO bridging lookups
	Gender	gender_spr11	Gender of the individual, possible values either male or female
	Ethnicity	ethnicgroupmajor_spr11	Pupil's major ethnic group based on extended ethnicity code. Allowed values: Any Other Ethnic Group, Asian, Black, Chinese, Mixed, Unclassified, White
	Special Educational Needs (SEN)	senprovision_spr11	Provision types under the SEN Code of Practice. Allowed values: No SEN; School Action or Early Years Action (up to 2014/15); Statement (up to 2017/18); SEN support (since 2014/15); Education, health and care plan (since 2014/15)
	FSM	everfsm_All_Spr11	Flag to indicate if pupil has ever been recorded as eligible for free school meals on Census day in any Spring Census up to the pupil's current year (not including nursery). A student qualifies if their parent or guardian receives certain government benefits, such as Income Support.
	2001 Lower Super Output Area	llsoa_spr11	National Statistics Postcode Directory Lower Layer Super Output Area derived from the pupil's postcode (based on 2001 Census)
NPD Key Stage 4	GCSE Attainment	KS4_LEVEL2_EM	Flag to indicate whether pupil achieved 5 or more GCSE and equivalents at grades A*-C (Level 2) including GCSE English and Maths.

duplicate records for some pupils. Therefore, a filter to retain only the 'main' record was used.

Activity states

Table 3 lists the variables used from the NCCIS, HESA, DWP Benefits, HMRC Employment and Self-employment SQL tables in LEO. To obtain the NCCIS and HESA variables, the Spring Census was inner joined to the respective data table with the same cohort filtering used to extract the individual characteristics. During HESA extraction, additional filters to keep only active records that had a start date at any point within the study period were used. For the DWP Benefits, HMRC Employment and Self-employment data, the Spring Census was first inner joined to the LEO bridging lookups and then inner joined to the respective table. The cohort filtering was subsequently applied to each of the data tables to obtain the variables listed in Table 3. When extracting the DWP Benefits data, additional filters to select only Out of Work (OfW) benefits for all start dates following the 2011/12 tax year were also applied. For the HMRC Self-employment data, a self-employed filter had to be used, and information was only available in tax years unlike the benefits and employment spells data using start and end dates.

Employment earnings and geography states

Similarly, for the employment earnings and geographic data, the Spring Census was first inner joined to the LEO bridging lookups and then inner joined to the respective table to obtain the variables in Table 4. The same filtering as the individual characteristics was applied to extract records for the correct cohort.

Import Extracted SQL Data into R

After extracting the required data as *.csv* files from the LEO Standard Extract using SQL, these were imported into

Table 3: Unpreprocessed activity state variables extracted from LEO data tables using SQL

LEO data table	Variable	LEO data variable name	Description – retrieved from (ONS, 2021)
NCCIS			
NPD Spring_ Census_2011	ID	pupilmatchingrefanonumous_ spr11	ID to link to other NPD tables and link to other LEO tables using LEO bridging lookups
NPD NCCIS_2011_ to_2019	Current activity	NCCIS_Current_Activity_ Code	Indicates Current Activity of the Young person
	Start date	NCCIS_Current_Activity_ Start_Date	Date Current Activity Started
	Verification date	NCCIS_Current_Activity_ Verification_Date	Date Current Activity last confirmed
HESA			
NPD Spring_ Census_2011	ID	pupilmatchingrefanonumous_ spr11	ID to link to other NPD tables and link to other LEO tables using LEO bridging lookups
HESA	Level of Study	he_xlev301	Level of study - 3 way split
	Start Date	he_comdate	Date of Commencement of Programme mm/dd/yyyy
	End Date	he_enddate	End date of instance
DWP Out of Work Benefits			
NPD Spring_ Census_2011	ID	pupilmatchingrefanonumous_ spr11	ID to link to other NPD tables and link to other LEO tables using LEO bridging lookups
LEO_Benefit	Start Date	startdate	Start date of benefit spell
	End Date	enddate	End date of benefit spell
HMRC Employment			
NPD Spring_ Census_2011	ID	pupilmatchingrefanonumous_ spr11	ID to link to other NPD tables and link to other LEO tables using LEO bridging lookups
LEO_ Employment	Start Date	startdate	Start date of employment spell
HMRC Self-employment			
NPD Spring_ Census_2011	ID	pupilmatchingrefanonumous_ spr11	ID to link to other NPD tables and link to other LEO tables using LEO bridging lookups
LEO_ Self_ Assessment	Self Employed	Self_Employed	Self-employed indicator

RStudio for preprocessing using the R programming Language. R Version 4.0.2 and R Studio Version 2022.07.2 Build 576 were used.

The data imported were:

- NPD Spring school Census 2010/11 school-leaver cohort with individual characteristics matched to LEO
- Publicly available Geographic lookups (static by definition) [17]
    - Lower Super Output Area (LSOA) 2001 to LSOA 2011 to Local Authority District (LAD) 2011
    - 2011 LAD to CA region
    - 2011 LAD to Government Office Region (GOR)
- 2015 Income Deprivation Affecting Children Index (IDACI) deciles at LSOA level (static for research purposes)
- 2011 Urban and Rural indicator at LSOA level (static for research purposes)
- NCCIS activity states from the 2011/12 -2017/18 academic year
- HESA activity states from the 2013/2014 – 2018/19 HESA academic reporting year
- DWP OfW Benefits from 2011/12 – 2018/19 tax year

Table 4: Unpreprocessed employment earnings and residential geography variables extracted from LEO data tables using SQL

LEO data	Variable	LEO data variable name	Description– retrieved from (ONS, 2021)
HMRC Employment Earnings			
NPD Spring_Census_2011	ID	pupilmatchingrefanonumous_spr11	ID to link to other NPD tables and link to other LEO tables using LEO bridging lookups
LEO_Earnings	Earnings	earnings	Summed earnings (per person) for each available tax year
DWP Residential Geography			
NPD Spring_Census_2011	ID	pupilmatchingrefanonumous_spr11	ID to link to other NPD tables and link to other LEO tables using LEO bridging lookups
LEO_Geography	Local Authority District Code	LAUA	Local authority code the learner resides in
	Local Authority District Name	LAUANM	Local authority name
	Government Office Region Code	GOR	Region code the learner resides in
	Government Office Region Name	GORNM	Region name the learner resides in

- HMRC Employment activities from 2011/12 – 2018/19 tax year
- HMRC Self-employment activities from 2013/14 – 2018/19 tax year
- HMRC Employment Earnings from 2011/12 - 2018/19 tax year
- DWP Residential Geographic states from 2011/12 - 2018/19 tax year

Figure 1 shows the activity states coverage for the 2010/11 English school leaver cohort. This shows the years that any amount of data was observed based on the cohort selection in the SQL queries prior to any preprocessing and linkage. For example, in Figure 3 the NCCIS Current_Activity variable used began from 2011/12 and no data relating to the 2010/11 school leaver cohort was observed for 2018/19. Where the term 'observed' is used in this report to describe available data, this means it is cohort specific. Therefore, if concerned with the 2017/18 school-leaver cohort, many NCCIS activities data would be observed in the 2018/19 academic year since pupils would be aged 16/17. Where the term 'observed' is not used, data availability is structural within the LEO dataset. The NCCIS data beginning from 2011/12 academic year determined the 2010/11 cohort chosen for this analysis as this maximised the longitudinal data available. No data was observed for HESA activities in academic reporting years 2011/12 and 2012/13. No Self-employment activities data was available in LEO for the 2011/12 – 2012/13 tax years and no data was observed for 2013/14 for the selected 2010/11 school-leaver cohort.

The Education and Skills Act 2008 in England required young people to remain in some form of education or training until age 17 in 2013 and then until age 18 in 2015 [23]. This did not apply for the selected 2010/11 cohort, which meant individuals could potentially enter the labour market directly after leaving school. Therefore, DWP and HMRC data that covered the entire span of the study period was utilised.

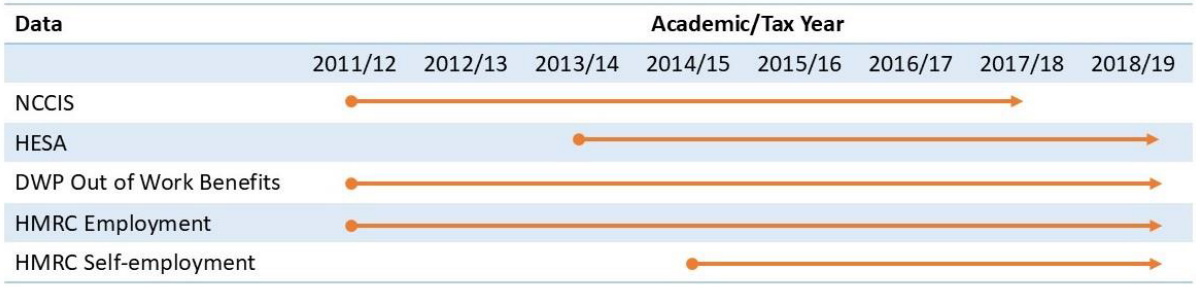
Preprocessing

Individual characteristics

The first stage in preprocessing the individual characteristics was to ensure that all Pupil Matching Reference IDs were unique and there were no duplicate records of individuals within the selected 2010/11 school-leaver cohort. In the LEO User Guide, it is highlighted that the KS4 Pupil table mostly contains a single record per pupil, but some may have additional records if they attended multiple education establishments within the academic year [22]. As the spring School Census 2010/11 school-leaver cohort was linked to GSCE attainment from the KS4 table, this meant there was a risk of duplicate values. Duplicates included two records where the GCSE criteria was either obtained or not obtained for <10 individuals. The record where GCSE was obtained was retained, leaving one record per individual. A SEN indicator variable was also created, where any type of SEN provision was coded as 1 and no SEN was coded as 0. Where SEN and FSM were missing, individuals were assumed not to have SEN or not be eligible for FSM.

To extract school-leavers residing in specific CAs, the geographic lookup files were linked to the individual characteristics. First, the 2001 LSOA to 2011 LSOA to 2011 LAD lookup file was left joined. The LEO LSOA variable extracted using SQL relates to the National Statistics Postcode Directory LSOA derived from the pupil's postcode based on the 2001 Census and available from the

Figure 1: Unpreprocessed Activity States Coverage for the 2010/11 School-leaver Cohort



2001/02 – 2013/14 academic years [22]. Since this was the only LSOA information available covering the 2010/11 academic year, the lookup provided more up to date geographic boundaries for the 2010/11 school-leaver cohort. The 2011 LAD to 2023 CAs and 2011 LAD to GOR lookups were also left joined to the individual characteristics. Greater Manchester was the first CA introduced in 2011, followed by the creation of 9 others from 2014 – 2018. Therefore, 2023 boundaries were used to facilitate CA-level analysis relevant to the current economic and political landscape. Once the lookups were joined, only records for individuals residing in England were retained. The 2011 Urban Rural classifications data was left joined, and recoded into an indicator with values either Urban, Rural or Unknown. The 2015 IDACI deciles variable was also left joined and a flag was created to indicate whether the LSOA was in the top 10% most deprived nationally. While these variables were pertinent to the working paper discussed in the Introduction, researchers may define any characteristics relevant to their specific interests.

NCCIS activity states

The NCCIS activity states for each academic year were extracted and imported in separate files. There was a possibility that more than one activity could be recorded per academic year. Therefore, the modal activity within the academic year was calculated. If there were multiple or no categorical modes, the first appearing activity was used.

To link these together, the activity states for the initial 2011/12 academic year was first left joined to the IDs from the preprocessed English 2010/11 school-leaver cohort from the section above. This meant that the activity states were retained only for those in the LEO matched cohort. The NCCIS activity states for the remaining academic years were successively left joined to this to form a longitudinal data frame of education histories.

The sequence analysis methodology which the data is being developed for has optimal performance when there are a relatively smaller number of activity state types, also known as the sequence analysis alphabet. This alphabet is the first analytical choice that is required within the sequence analysis workflow and is largely determined by the data used and the research purpose. Sparse activity states provide little insight during sequence analysis and clutter the visualisations produced; hence a more detailed alphabet does not necessarily equate to a better analysis. The ideal alphabet should balance parsimony and detail, which is best achieved by beginning with a comprehensive state alphabet and reducing this by collapsing states together where deemed appropriate [2 p113]. An iterative approach was taken to reduce the NCCIS activity states from 46 states to 11. Note that this reduced alphabet refers to NCCIS activity states only. The final sequence analysis alphabet used within the research data is presented in Figure 10 after integrating the NCCIS, HESA, HMRC and DWP data.

The relative frequency of the NCCIS activities, iterative generation of sequence analysis figures, and the 2010 NCCIS Data Catalogue which consisted of pre-grouped activities [24 pp93-94], were collectively used to optimise and reduce the alphabet. The results produced during the iterative sequence analysis process were not permitted ONS clearance. Figure 2 summarises the reduced NCCIS alphabet.

HESA activity states

Similar to the NCCIS activity states, HESA activity states for each academic year were also extracted and imported in separate files. All files were checked to ensure there were no duplicate IDs and they were successively full joined to retain every individual who ever had a record of Higher Education (HE) in HESA. This linking created a longitudinal data frame with one record per individual and the possible activity state values were either 'Postgraduate', 'Undergraduate' or 'Further Education'. In this analysis, 'Undergraduate' was relabelled as 'Higher Education' since this was not limited to first degrees and contained other types of HE level qualifications [25]. It also meant this could be grouped with the NCCIS Higher Education state within the sequence analysis alphabet. It should be noted that the activity state used referred to any period of any length in HE. For example, a 2-month or 9-month spell in HE were both marked as HE for the full academic year.

There were a small number of outliers in the first two academic years of the study period (2011/12 and 2012/13). The HE activities observed within these years had no end date and instead had a second start date in 2013/14. Although it is unlikely school-leavers would enter directly into HE, it was possible these data were relating to the HESA 'Further Education' activity state. However, the level of study was not further education and so these years were discarded. Hence, the HESA data used in the analysis ran from the 2013/14 – 2018/19 academic year. This longitudinal HESA data was left joined to the IDs from the English 2010/11 school-leaver cohort, selected in the Individual Characteristics Preprocessing Section, to retain only HESA records related to the cohort.

Figure 2: Reduced NCCIS activity states informed by observed total state frequencies across all years for the selected 2010/11 English school-leaver cohort, an archived Department for Education data catalogue and iterative sequence analysis visualisation

NCCIS Activity Code	NCCIS Activity State Description	Reduced Alphabet (11 NCCIS States)
210	School Sixth Form	School Sixth Form
220	Sixth Form College	Sixth Form College
230	Further Education	Further Education
240	Higher Education	Higher Education
310	Apprenticeship	Apprenticeship
320	Full time employment with study (regulated qualification)	Employment
330	Employment without training	
340	Employment with training (other)	
350	Temporary employment	
360	Part Time Employment	
410	Education Funding Agency (EFA) and Skills Funding Agency (SFA) funded Work-based learning (~ESFA from 2017)	Government Supported
430	Other training	
440	Training through the Work Programme	
450	Traineeship	
460	Supported Internships	
510	Personal Development Opportunity in receipt of allowance or wage	Not in Education, Employment or Training (NEET)
520	Other Personal Development Opportunities	
530	Reengagement provision	
610	Those not yet ready for work or learning	
615	Start date agreed (other)	
616	Start Date agreed (RPA compliant)	
619	Seeking employment, education or training	
620	Not available to labour market Young carers	
630	Not available to labour market Teenage parents	
640	Not available to labour market Illness	
650	Not available to labour market Pregnancy	
660	Not available to labour market on religious grounds	
670	Not available to labour market those who are currently unlikely to be economically active	
680	Not available to labour market Other reason	
380	Self-employment	Self-employment
381	Self-employment combined with study (regulated qualification)	
110	Registered at a school or other educational establishment	Other
130	In a Custodial Sentence	
140	Not registered at school or educational establishment	
250	Part time Education	
260	Gap Year students	
270	Other post 16 education	
280	Independent specialist provider	
540	Working not for reward	
550	Working not for reward combined with part time study	
710	Custody - young adult offender	
720	Refugees/Asylum seekers	
150	Current Situation not known	Unknown/NA
810	Current situation not known	
820	Cannot Be Contacted	
830	Refused to disclose activity	
Key: High frequency, Medium frequency, Low frequency, Very low frequency		

DWP OfW benefits, HMRC employment and self-employment activity states

Unlike the yearly NCCIS and HESA activity states, the DWP OfW Benefits and HMRC Employment LEO data were available in only a spells format with start and end dates. The respective spells extracted were from the beginning of the 2011/12 tax year and covered the full length of the study period until the 2018/19 tax year in a singular file. Both the OfW Benefits and Employment start and end date of the spells were converted into tax year start and end dates so that these could eventually be integrated with the yearly NCCIS and HESA academic data. Where start and end dates were the same tax year, these were marked as the full tax year. The data were transformed into two separate longitudinal data frames and linked to the Individual Characteristics Preprocessing Section cohort ID's to retain only relevant records. The first had yearly columns to indicate whether the individual was

claiming OfW benefits and the second had columns to indicate whether the individual was in employment at any point in a given tax year.

The OfW benefits were [22]:

- Jobseekers Allowance
- Jobseekers Training Allowance
- Employment and Support Allowance
- Incapacity Benefit
- Income Support
- Passported IB
- Severe Disablement Allowance
- Pension Credit
- State (Retirement) Pension
- Carers Allowance (Invalid Carers Allowance)
- Attendance Allowance
- Universal Credit – Searching for Work
- Universal Credit – No Work requirements
- Universal Credit – Preparing for work
- Universal Credit – Planning for work

The Self-employment data in LEO were available only in a yearly format. The activity states for each tax year were extracted from SQL and imported into R as separate files. Successive full joined for the 2014/15 to the 2018/19 tax year data were performed to create the required longitudinal self-employment histories. Again, only the records relevant to the 2010/11 cohort were retained using the individual characteristics IDs.

Integrate NCCIS, HESA, DWP and HMRC activities

Figure 3 enables a visual representation of the longitudinal trajectories intrinsic to sequence analysis. It shows the observed data across the study period in the preprocessed NCCIS, HESA, DWP and HMRC activity states from the Preprocessing sections. This largely reflected the activity states coverage in Figure 1, but now all yearly data have been linked per component dataset and each filtered to retain only the records relevant to the preprocessed cohort in the Individual Characteristics Preprocessing section. Therefore, each component dataset subplot in Figure 3 consists of 557,171 individuals and each horizontal line represents a longitudinal trajectory.

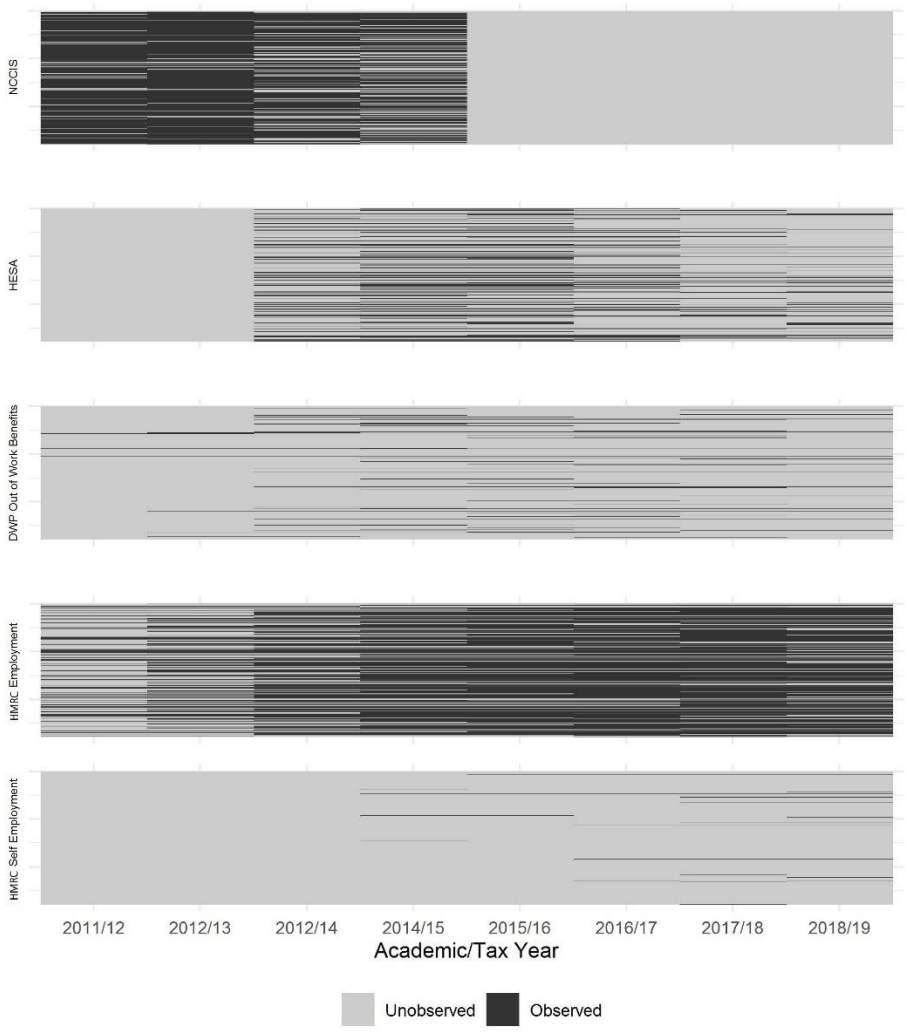
In order to integrate these activity states data, a decision regarding the use of academic and tax years was necessary as NCCIS and HESA data were based on academic years, but DWP/HMRC data were based on tax years. The employment and benefit data in LEO are provided as spells, hence there is an ability to preprocess these into academic years. However, since the self-employment and earnings data were only provided in tax years, I chose to preprocesses all DWP/HMRC data as tax years and integrate these with the education data in academic years.

Figure 4 shows the calendar date, academic year, tax year and the age of individuals in the 2010/11 academic year. The blue line represents the beginning of the study period in the 2011/12 academic year corresponding to the first post-16 non-compulsory activity for the selected 2010/11 school-leaver cohort. It can be seen there are two possible tax years that run during a given academic year. Either the tax year in the first half or the tax year in the second half of the academic year can be used to integrate the NCCIS and HESA data with the DWP/HMRC data. The decision was made to align with the tax year in the first half of the academic year, for example the 2011/12 tax year data was integrated with the 2011/12 academic year data. This alignment was chosen as it enabled a tax year to cover a greater proportion of an academic year (7 months). Additionally, on a data development level, it allowed a more straightforward integration process as the academic and tax year labelling remained the same. Furthermore, it meant that more DWP/HMRC data was available near the end of the study period where a greater number of OfW benefits, employment and self-employment states were observed. If the tax year in the second half of the academic year was used for alignment, there would have been no 2019/20 tax year data in LEO to align with the 2018/19 academic year.

Figure 5 illustrates the integration procedure taken. The NCCIS, HESA, OfW Benefits, Employment and Self-employment panel data were effectively 'overlaid' successively. Since the purpose of the research was to investigate education and labour market activities, the NCCIS was used as the base data. This enabled more detailed education states to be present. Where NCCIS data were missing, HESA was overlaid. Where data were still missing, the OfW benefits were filled. Benefits can be claimed whilst being either employed or self-employed in the LEO data [22]. The OfW Benefits was given priority over the employment and self-employment data to highlight any individuals who claimed benefits at any point in the year, regardless of whether they were (self) employed. The research that the dataset was built for was intended to identify individuals who may experience difficult trajectories into the labour market. Therefore, prioritising OfW Benefits over the employment and self-employment data helped to better achieve this aim. Individuals who were self-employed could also be employed [22]. Employment data was integrated first since this typically includes more stable labour market participation than self-employment.

A limitation in the integrated activities was that degree/HE-level apprenticeships in HESA were recorded as 'Undergraduate' [18]. Note that during the HESA preprocessing, 'Undergraduate' was relabelled to 'Higher Education'. Nevertheless, this meant that apprenticeships from NCCIS could not follow through into subsequent HESA integrated years since they could not be distinguished. An apprenticeships flag was present in HESA, but this was only available from 2016/17 academic year onwards and hence was not utilised [26]. An 'initiatives' variable distinguishing apprenticeships was also available in HESA but access to this variable for the project was not permitted [27]. This could be used in any future developments of this integrated data.

Figure 3: Observed data in the 2010/11 English school-leaver cohort preprocessed activities prior to integration, n = 557, 171



Note: Figure 3 includes some overplotting due to the large data sample size.

The LEO HMRC data did not include apprenticeship flags and so apprenticeship courses at any level were likely recorded as 'Employment'.

Employment earnings and residential geographic states

The employment earnings for each tax year were extracted and imported in separate files. They were successively full joined and linked to the preprocessed individual characteristics IDs to retain records for only the 2010/11 English school-leaver cohort. This created a longitudinal history of known employment earnings per tax year. The geographic states for each tax year were also extracted and imported in separate files. The LAD to CA lookup file was used to create a new variable which specified the individuals' geographic location as CAs for each tax year. Where the geography was not within a CA, the GOR was used. These data frames were also full joined and linked to the individual characteristics IDs to keep the relevant cohort records.

The longitudinal earnings and geographic histories were created but not used within the working paper mentioned in the Introduction. Future sequence analysis work could incorporate employment and self-employment earnings data alongside employment sector data from LEO Iteration 2 [28] to gain further insights. An early research intention was to use the school-to-work histories created in conjunction with the geographic histories for multichannel sequence analysis. Multichannel sequence analysis is a relatively new extension of traditional sequence analysis which accounts for multiple trajectories simultaneously [2]. The *TraMineR* sequence analysis package is continuously developing to incorporate more functionality for such techniques [29]. This route was not followed since the approved data included residential geographic location only and did not include term-time or temporary addresses. Future research could adapt the linked geographical data to incorporate such information and create a more accurate history of movements. Geographic locations at the end of the study period could also be used to understand whether young people remained in their original CAs or relocated.

Figure 4: Academic year and tax year data alignment

<table>
<tr><th rowspan="2">Calander Date</th><th rowspan="2">Academic Year</th><th rowspan="2">Tax Year (From 6th April)</th><th colspan="2">Age of youngest/oldest individual in the 2010/11 academic year</th></tr>
<tr><th>August Born</th><th>September Born</th></tr>
<tr><td>Sept-2010</td><td rowspan="12">Academic Year 2010/11</td><td rowspan="6">Tax Year 2010/11</td><td rowspan="11">15</td><td rowspan="11">16</td></tr>
<tr><td>Oct-2010</td></tr>
<tr><td>Nov-2010</td></tr>
<tr><td>Dec-2010</td></tr>
<tr><td>Jan-2011</td></tr>
<tr><td>Feb-2011</td></tr>
<tr><td>Mar-2011</td><td rowspan="6"></td></tr>
<tr><td>Apr-2011</td></tr>
<tr><td>May-2011</td></tr>
<tr><td>Jun-2011</td></tr>
<tr><td>July-2011</td></tr>
<tr><td>Aug-2011</td><td></td></tr>
<tr><td>Sept-2011</td><td rowspan="11">Academic Year 2011/12</td><td rowspan="6">Tax Year 2011/12</td><td rowspan="11">16</td><td rowspan="11">17</td></tr>
<tr><td>Oct-2011</td></tr>
<tr><td>Nov-2011</td></tr>
<tr><td>Dec-2011</td></tr>
<tr><td>Jan-2012</td></tr>
<tr><td>Feb-2012</td></tr>
<tr><td>Mar-2012</td><td rowspan="6">Tax Year 2012/13</td></tr>
<tr><td>Apr-2012</td></tr>
<tr><td>May-2012</td></tr>
<tr><td>Jun-2012</td></tr>
<tr><td>Jul-2012</td></tr>
<tr><td>Aug-2012</td><td rowspan="5">17</td></tr>
<tr><td>Sept-2012</td><td rowspan="4">Academic Year 2012/13</td><td rowspan="4">18</td></tr>
<tr><td>Oct-2012</td></tr>
<tr><td>Nov-2012</td></tr>
<tr><td>Dec-2012</td></tr>
<tr><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td></tr>
<tr><td>Sept-2018</td><td rowspan="12">Academic Year 2018/19</td><td rowspan="6">Tax Year 2018/19</td><td rowspan="11">23</td><td rowspan="11">24</td></tr>
<tr><td>Oct-2018</td></tr>
<tr><td>Nov-2018</td></tr>
<tr><td>Dec-2018</td></tr>
<tr><td>Jan-2019</td></tr>
<tr><td>Feb-2019</td></tr>
<tr><td>Mar-2019</td><td rowspan="6">Tax Year 2019/20 NOT AVAILABLE IN LEO ITERATION 1</td></tr>
<tr><td>Apr-2019</td></tr>
<tr><td>May-2019</td></tr>
<tr><td>Jun-2019</td></tr>
<tr><td>Jul-2019</td></tr>
<tr><td>Aug-2019</td><td></td></tr>
</table>

Results

Table 5 details the attrition of records throughout the data linkage procedures in the Preprocessing sections. There were 556, 182 individuals remaining in the final English 2010/11 school-leaver cohort after preprocessing the data. The CA that individuals were residing in at school-leaving age from the preprocessed individual characteristics were left joined to the integrated activity states. This enabled the education and employment activity histories of school-leavers initially residing in a CA or Greater London to be subset from the full English cohort for use in the working paper described in the Introduction. Other geographic levels from the individual characteristics could also be left joined to the integrated activity states based on researcher needs.

Figure 7 shows the final observed data after integrating the NCCIS, HESA, DWP and HMRC data. More intuitively, this is the result of effectively collapsing the visual representations of the observed data in Figure 3. It is common to impute missing data within activity histories during the sequence analysis workflow, especially when using survey data [2 pp17-20]. However, since LEO provided such a large sample size with very little unobserved data seen in Figure 7, missing activity states were not imputed. Instead, these were included in the sequence analysis alphabet as a 'missing' state (Figure 10).

Figures 8 and 9 show examples of the school-to-work histories created and the individual-level socio-demographic characteristics using fabricated data. These figures show the structure of the data designed for subsequent sequence analysis research. The ID variable can be used to link the individual characteristics to the sequence analysis outputs to explore the results.

Figure 10 summarises the final sequence analysis alphabet of education and employment activity states possible within the English longitudinal activity histories. The sequence analysis methodology aims to reduce the complexity of

Figure 5: Integration procedure used to combine the NCCIS, HESA, DWP and HMRC preprocessed activities to create full education and employment activity state histories

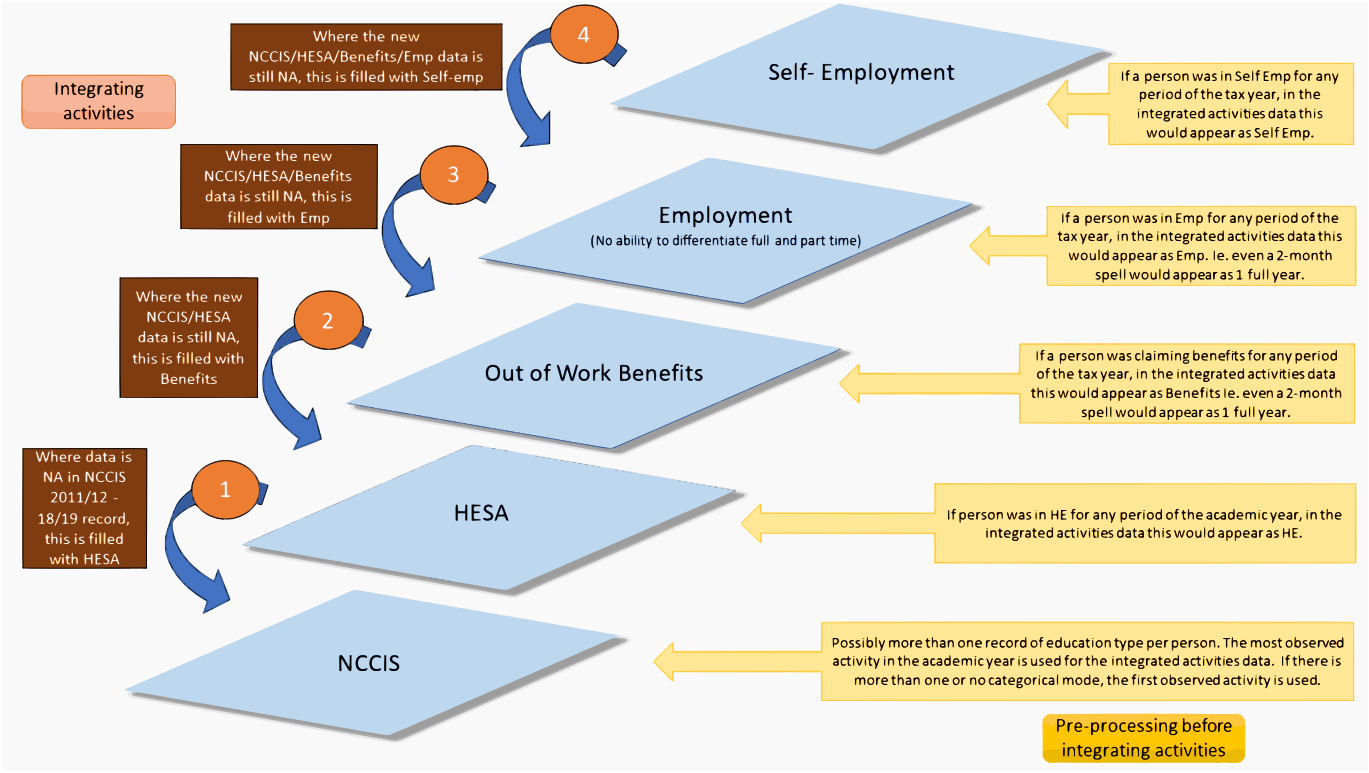
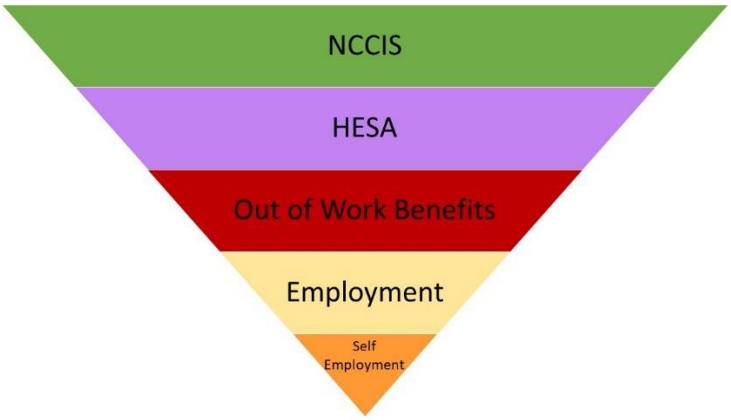


Figure 6: Hierarchy imposed by the integration of the education and employment activity data



longitudinal data and hence outputs are best summarised visually. The colours used to represent the activities in the alphabet are an important aspect in sequence analysis visualisations which are often overlooked. It is advised that researchers should use contrasting colours to emphasise certain states and use different shadings to help visualise commonalities between states [1]. For this developed dataset, I chose to represent early post-16 activities as shades of blue; apprenticeships and other government supported activities as shades of green; HE and postgraduate activities as shades of purple; Not in Education, Employment or Training (NEET) and OfW benefits as shades of red; and self (employment) as shades of orange. Other and missing activities were represented in greys to have minimal distraction in the sequence analysis visualisations. The contrasting colour groups and shadings used helped to better highlight patterns within the data as well as reduce cognitive load.

Figure 11 summarises the developed school-to-work trajectories of the 2010/11 school-leaver cohort split by CAs and Greater London, corresponding to procedure No. 7 in Table 5. These subplots are called 'sequence index plots' and were generated using the *TraMineR* sequence analysis R package [2, 29]. Within a subplot, each line represents an individual's trajectory from age 16/17 to age 23/24 and the colours represent the activity state they were in each year. This enables a longitudinal view of trajectories for a specific region to gain an idea of inherent patterns. Figure 11 is intended to provide a basic overview of the input data that has been created within this data resource, which can

Table 5: Attrition of individuals throughout the data development process

Procedure no.	Procedure	Number of records	Cumulative number of records lost
1	Select cohort from NPD School Census with on roll and main record filtering	568, 586	
2	Inner join this with KS4	563, 875	4711
3	Inner join this with LEO bridging lookups (i.e. is resolved/LEO matched)	559, 164	9422
4	Drop duplicate IDs caused by KS4 linkage (LEO match rate $558{,}810/568{,}586 \times 100 = 98.3\%$)	558, 810	9776
5	Retain only those who had geography data recorded in 2010/11 academic year and only those in England	557, 171	11, 415
NCCIS	NCCIS records already left joined to the above cohort before preprocessing	557, 171	
HESA	Preprocessed yearly HESA records *before* left joining to the above cohort	Suppressed	
OfW Benefits	Preprocessed yearly benefits records *before* left joining to the above cohort	119, 576	
Employment	Preprocessed yearly employment records *before* left joining to the above cohort	494, 470	
Self- Employment	Preprocessed yearly self-employment records *before* left joining to the above cohort	44,000	
6	Integrated activity states linked to procedure No. 5 cohort (i.e. those who had observed activity histories – see Figure 7)	556, 182	12, 404
7	Subset individuals residing in CAs and Greater London at school-leaving age	230,756	
	South Yorkshire	15, 026	
	Greater Manchester	29, 574	
	Cambridgeshire and Peterborough	8157	
	Liverpool City Region	17,339	
	North East	12,647	
	North of Tyne	8215	
	Tees Valley	8050	
	West Midlands	31,936	
	West Yorkshire	24,784	
	West of England	4883	
	Greater London GOR	70,145	

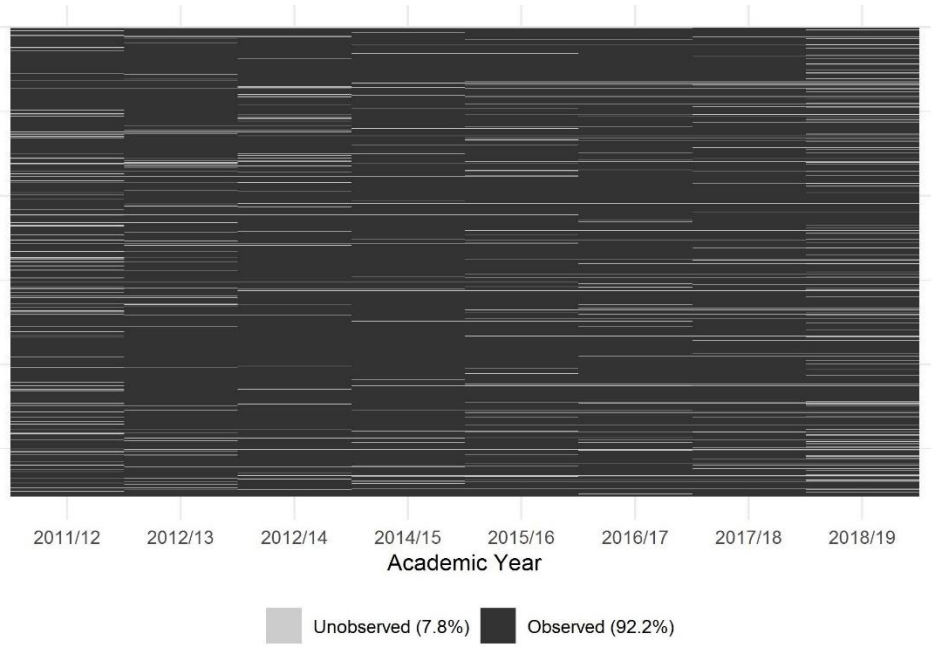
now be interrogated in future research using sequence analysis techniques.

Discussion

This research aimed to provide a comprehensive guide for constructing longitudinal school-to-work datasets suitable for input into sequence analysis algorithms. It was assumed that researchers either (1) had already selected sequence analysis as their methodological approach and required guidance on dataset construction or (2) sought to understand the practical manipulation of the LEO data. Notably, the study did not seek to perform sequence analysis or interpret empirical findings but rather to detail the development of the final dataset. As a result, this discussion does not engage in a conventional comparison of research results with existing sequence analysis literature.

A key challenge arising from the novelty of this contribution is the limited prior work against which to contextualise the methodological choices made in this study. Anderson and Nelson [13] provided one of the few technical reports using the LEO data for school-to-work analysis. Certain aspects of the data preprocessing, such as the establishment of a hierarchy between education, employment, and benefit activity states drew from this existing report. However, Anderson and Nelson's [13] work primarily described the rationale behind their decisions rather

Figure 7: Observed data in the 2010/11 English school-leaver cohort integrated activities, n = 556,182



Note: Figure 7 includes some overplotting due to the large data sample size.

Figure 8: Example of integrated activities with fabricated data

ID_PMR	2011/12	2012/13	2013/14	2014/15	2015/16	2016/17	2017/18	2018/19	MCA_Name
1	School 6th Form	School 6th Form	Higher Education	Higher Education	Higher Education	Postgraduate	Employment	Employment	South Yorkshire
2	Further Education	Further Education	NEET	NEET	NEET	Out of Work Benefits	Out of Work Benefits	Out of Work Benefits	Greater Manchester

Figure 9: Example of preprocessed individual characteristics with fabricated data

ID_PMR	Gender	Ethnicity	SEN	FSM	GCSE_attainment	Urban_Rural	IDACI_Decile_MostDep	LSOA_code_2011	LAD_code_2011	LAD_Name	MCA_Name	Region
1	Female	Asian	No	Yes	Yes	Urban	Yes	E01007854	E08000019	Sheffield	South Yorkshire	Yorkshire and the Humber
2	Male	White	Yes	Yes	Yes	Urban	Yes	E01005488	E08000005	Rochdale	Greater Manchester	North West

than outlining a step-by-step, reproducible methodology for dataset construction. Furthermore, because their research was not intended for sequence analysis, it did not address important considerations such as building a well-defined sequence analysis alphabet. In contrast, the dataset developed in this study was detailed thoroughly and designed explicitly to optimise sequence analysis research.

Beyond providing a practical resource for researchers, this work also establishes a methodological foundation for future discussions on best practices in sequence analysis dataset construction. Subsequent studies can build upon this work, critically evaluating and refining the data development process in relation to the approach presented here.

Limitations

The school-to-work activities were prepared in a yearly format and assumed the activity for the full year regardless of the amount of time spent in that state. This was true for the HE, Employment or Benefits data. However, the order of integration (Figure 6) helped to correct some of the states, for example a 1-month period in employment would have been marked as an education state if applicable for that year since education states took precedence over employment states. Even though the opposite case, e.g. a 1-month period in education and a longer spell in employment being marked as education is possible, the hierarchical ordering designed is sufficient for this research. Although monthly data could have been used to build the activity histories, it is highly likely that the resulting dataset would have been too large for *TraMineR* to accept. Furthermore, due to project time constraints, it was not feasible to dedicate further time into developing the data. It is acknowledged that this guide is only one method of building such a dataset and many other methodologies exist and hence should be adapted to researcher's specific needs.

The structural limitations of the LEO data can be found in the LEO User Guide [22]. Particularly relevant limitations

Figure 10: Final sequence analysis alphabet of the possible activity states in the full English 2010/11 cohort, n = 556,182

Number	Colour Representation	Activity State	Frequency	Percentage
1		School 6th Form	433,642	9.7
2		6th Form College	132,301	3.0
3		Further Education	502,381	11.3
4		Apprenticeship	115,075	2.6
5		Government Supported	41,907	0.9
6		Higher Education	716,487	16.1
7		Postgraduate	62,697	1.4
8		NEET	126,013	2.8
9		Out of Work Benefits	302,471	6.8
10		Employment	1,604,779	36.1
11		Self Employment	48,014	1.1
12		Other	17,734	0.4
13		missing	345,955	7.8
Total			4,449,456	100.00

were that there was no information regarding the hours worked and so part-time and full-time employment could not be distinguished, there was no information on the types of employment in iteration 1 of the data, and low-income individuals who earnt below the minimum tax threshold may not have been included if they worked for a small employer. Also, apprenticeship activity states were labelled as 'Higher Education' in HESA and as 'Employment' in LEO. Therefore, they were not distinguishable in later years. A limitation of administrative data is that missing activity states could be due to reasons such as migration or death which cannot be distinguished from labour market inactivity.

Future research

Since the data was developed to undertake CA focused sequence analysis, only a small proportion of the prepared activities and individual characteristics of the 2010/11 English school-leaver cohort were utilised. Therefore, there is untapped potential regarding the full English cohort or analysis at lower geographic levels. The LAD information is available in the prepared data which could be used to investigate school-to-work trajectories in any specific location in England. To add to this, the data resource could be used to create trajectories for multiple cohorts for cross-cohort sequence analysis.

In future developments, the proportion of the year in the state could be used to increase the accuracy of the yearly activities discussed in the limitations. The NCCIS data provided the necessary FE activities information required for this research. However, the Individualised Learner Records data in LEO could be used to incorporate qualification levels and course subject information. This could be particularly interesting and provide new insights if used in combination with the additional Inter-Departmental Business Register employment sector data now available in LEO Iteration 2 [7, 28]. Moreover, there is scope to explore the employment earnings and geographical trajectories data. The developed longitudinal data could also be extended past the 2018-19 academic year using LEO Iteration 2. This could lead to research understanding the long-term effect of COVID-19 on school-leaver trajectories. It should be noted that in the second iteration of LEO, LAD areas are pseudonymised, therefore researchers may need to use the Output Area data available and a lookup file to aggregate up to LAD-level, subject to data owner's approval.

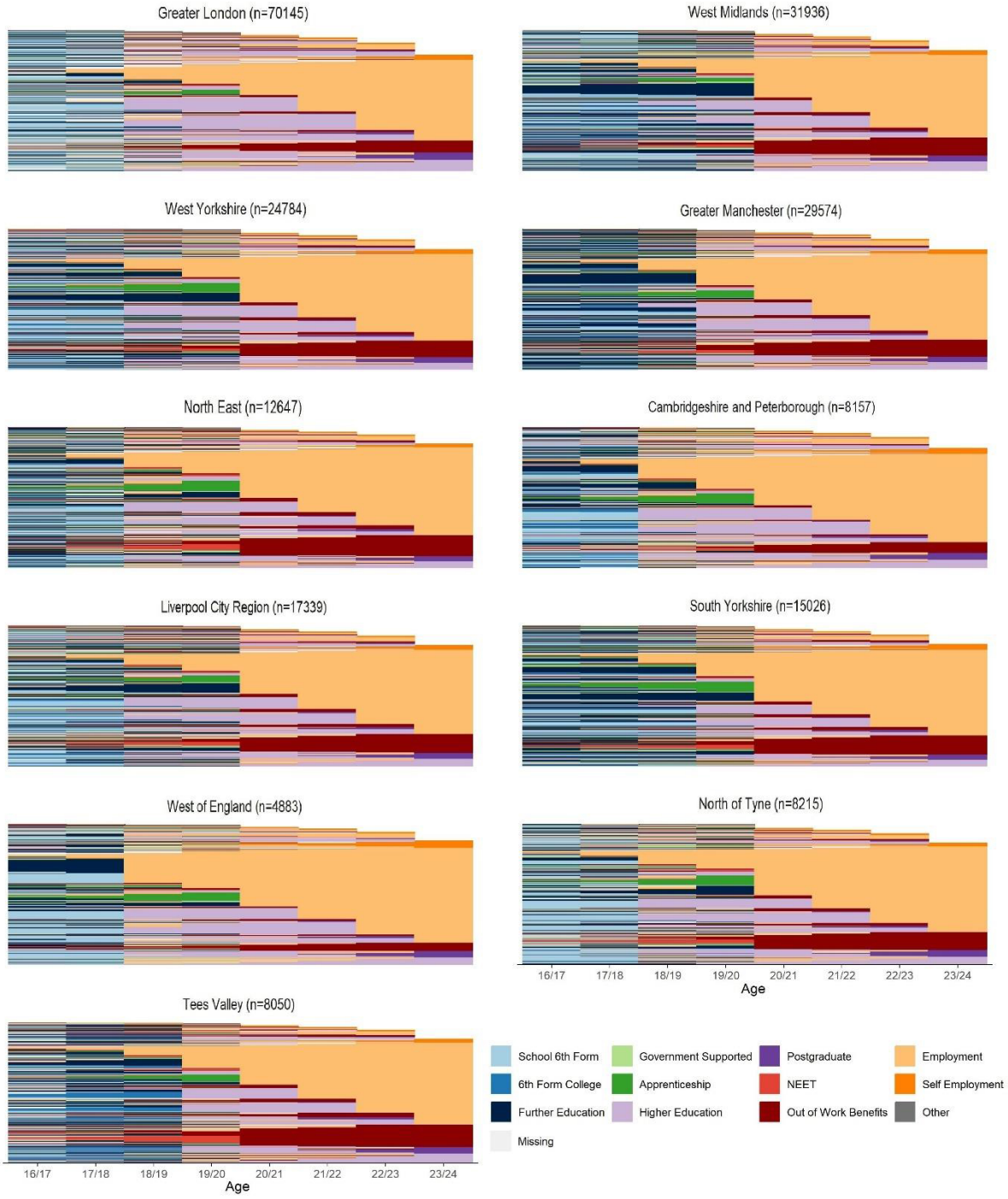
Although the dataset was developed for sequence analysis with LEO data, there is potential for the dataset to be utilised as input for other data analysis and data science techniques. The overall data development procedure used in this data resource could also be adapted to other administrative datasets based on researcher requirements.

Conclusions

Despite the growth of school-to-work sequence analysis research and the recent emphasis on teaching the method, there has been limited guidance on preparing the input data required for its application. Existing journal articles do not provide the full data development process and teaching materials tend to use an 'ideal' dataset that fails to address the complexity of administrative or survey data. Understanding how to prepare sequence analysis input data involves several analytical decisions by the researcher and is important since this directly influences the results.

This data development guide aims to address the lack of technical reports outlining how to develop an input dataset for sequence analysis. Longitudinal school-to-work trajectories for 556,182 individuals in the 2010/11 English school-leaver cohort were created using LEO, with the ability to subset individuals by their CA geography. Individual-level socio-demographic data were also prepared which could be linked to the respective school-to-work trajectory. Longitudinal

Figure 11: Sequence Analysis Index Plots Showing School-to-work Trajectories for the 2010/11 School-leaver Cohort per Combined Authority (n = 230,756)



Note: Figure 11 includes some overplotting due to the large data sample size.

geographic data and employment earnings data was also created.

A key strength of this data resource is the comprehensive methodology section which is intended to be accessible to both academic and non-academic audiences. For researchers interested in conducting school-to-work sequence analysis using LEO, the developed data can be recreated using this in-depth guide and open-source code [17]. It also acts as a practical blueprint for those new to the sequence analysis method who may need to develop their own data. Moreover, the work demonstrates the capabilities of the LEO data and how this can be leveraged for impactful analysis. This unique contribution is well-positioned to facilitate future research and innovation in the field.

Acknowledgements

I would like to thank the ONS and the Department for Education for providing access to the LEO dataset. I would

also like to thank the ONS Secure Research Service Customer Support, Operations and Statistical Support staff for their assistance and time to clear research outputs. Additionally, a thank you to the University of Sheffield Department of Economics for allowing me to use their resources.

I would also like to thank my PhD supervisors Andy Dickerson, Gwilym Pryce and Philip McCann for their valued support and guidance during the creation of the dataset. I send my regards to Grace Simmons at the South Yorkshire Mayoral Combined Authority who helped shape my data development approach through our regular discussions. Moreover, I am grateful for the funding provided by the Economic and Social Research Council through the Data Analytics and Society Centre for Doctoral Training, as well as funding from my project partner, the South Yorkshire Mayoral Combined Authority.

Ethics statement

Ethical approval was gained through the ONS SRS during data access for project 2022/042 'Education and Labour Market Trajectories in UK City Regions'. Ethics approval was also gained through the University of Sheffield Research Ethics Committee.

Conflict of interests statement

None declared.

Publication consent

Publication clearance granted by the ONS SRS.

Funding statement

This work was supported by the Economic and Social Research Council [grant 2433665, https://gtr.ukri.org/projects?ref=studentship-2433665] and the South Yorkshire Mayoral Combined Authority. The funding sources were not involved regarding the preparation of the article in study design; in the collection, analysis and interpretation of the data; in writing of the report; and in the decision to submit the paper for publication.

Data availability statement

Department for Education; HM Revenue and Customs; Department for Work and Pensions; Higher Education Statistics Agency, released 29 September 2022, ONS SRS Metadata Catalogue, dataset, Longitudinal Education Outcomes SRS Iteration 1 Standard Extract - England, https://doi.org/10.57906/pzfv-d195.

The LEO data is owned by the Department for Education. Access to the data is granted through requests via the ONS SRS and applicants must be fully accredited researchers under the Digital Economy Act 2017 [19]. For a more in-depth explanation of the way LEO was collected and constructed by the Department for Education, the LEO User Guide can be requested from the ONS [22]. All ONS security cleared R code, SQL code and other files related to this data resource can be found on GitHub [17].

References

1. Liao TF, et al. Sequence analysis: Its past, present, and future. Soc Sci Res. 2022;107:102772. https://doi.org/10.1016/j.ssresearch.2022.102772

2. Raab M, Struffolino E. Sequence Analysis. Thousand Oaks, CA: Sage; 2022. https://doi.org/10.4135/9781071938942

3. Saqr M, et al. Sequence analysis in education: principles, technique, and tutorial with R. In: Saqr M, López-Pernas S, editors. Learning analytics methods and tutorials. Cham: Springer; 2024. p. 321–54. https://doi.org/10.1007/978-3-031-54464-4_10

4. Sequence Analysis Association. Sequence Analysis Association. 2024. Available from: https://sequenceanalysis.org/.

5. Dickerson A, McDool E, Morris D. Post-compulsory education pathways and labour market outcomes. Educ Econ. 2023;31(3):326–52. https://doi.org/10.1080/09645292.2022.2068137

6. Cornwell B. Theoretical foundations of social sequence analysis. In: Granovetter M, editor. Social sequence analysis. Cambridge: Cambridge University Press; 2015. p. 21–56. https://doi.org/10.1017/CBO9781316212530

7. Department for Education. Transparency data: Longitudinal Education Outcomes (LEO) data. GOV.UK. 2024. Available from: https://www.gov.uk/government/publications/longitudinal-education-outcomes-leo-dataset/longitudinal-education-outcomes-leo-data.

8. Jay MA, Grath-Lone LM, Gilbert R. Data resource: The National Pupil Database (NPD). Int J Popul Data Sci. 2019;4(1). https://doi.org/10.23889/ijpds.v4i1.1101

9. Anderson O. Post-16 education and labour market activities, pathways and outcomes (LEO). Int J Popul Data Sci. 2023;8(2). https://doi.org/10.23889/ijpds.v8i2.2186

10. Bowyer A, Dorsett R, Thomson D. The school-to-work transition for young people who experience custody. Longitud Life Course Stud. 2023;14(3):339–57. https://doi.org/10.23889/ijpds.v8i2.2186

11. Anders J, Dorsett R. What young English people do once they reach school-leaving age: A cross-cohort comparison for the last 30 years. Longit Life Course Stud. 2017;8(1):75–103. https://doi.org/10.14301/llcs.v8i1.399

12. Department for Education. Post-16 pathways at level 3 and below: Experimental statistics on young people's transitions from education to work in England. London: Department for Education; 2020.

13. Anderson O, Nelson M. Technical report for education and labour market pathways of individuals (LEO). London: Department for Education; 2021.

14. Wright L. Producing working-life histories in the BHPS and UKHLS. Essex: UK Data Service; 2020. https://doi.org/10.17605/OSF.IO/C3V9F

15. Local Government Association. Combined authorities. 2024. Available from: https://www.local.gov.uk/topics/devolution/devolution-online-hub/devolution-explained/combined-authorities.

16. Overman HG, Xu X. Spatial disparities across labour markets. Oxf Open Econ. 2024;3(Suppl 1):i585–i610. https://doi.org/10.1093/ooec/odae005

17. Sickotra S. Developing School-to-work Trajectories for Sequence Analysis Using the Longitudinal Education Outcomes (LEO) Data. 2023. Available from: https://github.com/sickotra/Developing_SchooltoWork_Trajectories_for_Sequence_Analysis_LEO_Data.git.

18. Office for National Statistics. LEO I1SE variable request form v1.3. Newport: Office for National Statistics; 2021.

19. Department for Education. Apply to access the Longitudinal Education Outcomes (LEO) dataset. GOV.UK. 2021. Available from: https://www.gov.uk/guidance/apply-to-access-the-longitudinal-education-outcomes-leo-dataset.

20. GOV.UK. Find and explore data in the National Pupil Database. 2025. Available from: https://www.find-npd-data.education.gov.uk/categories.

21. Higher Education Statistics Agency. Higher education student data. 2025. Available from: https://www.hesa.ac.uk/data-and-analysis/students.

22. Department for Education. LEO Data: A guide for users. London: Department for Education; 2019.

23. UK Parliament. Education and Skills Act 2008. 2009. Available from: https://bills.parliament.uk/bills/202.

24. Department for Education. Client Caseload Information System Data Catalogue 2010–11 [Archived]. Sheffield: Department for Education; 2010.

25. Higher Education Statistics Agency. XLEV301_1.13.1. 2023. Available from: https://www.hesa.ac.uk/collection/c17051/derived/xlev301.

26. Higher Education Statistics Agency. XAPP01_1.1.1. 2017. Available from: https://www.hesa.ac.uk/collection/c18051/derived/xapp01.

27. Higher Education Statistics Agency. Student 2021/22 – Initiatives. 2021. Available from: https://www.hesa.ac.uk/collection/c21051/a/initiatives.

28. Office for National Statistics. LEO I2SE Variable Request Form V2.1. Newport: Office for National Statistics; 2023.

29. Gabadinho A, Ritschard G, Müller NS, Studer M. Analyzing and visualizing state sequences in R with TraMineR. J Stat Softw. 2011;40(4):1–37. https://doi.org/10.18637/jss.v040.i04

Abbreviations

LEO:	Longitudinal Education Outcomes
NPD:	National Pupil Database
HESA:	Higher Education Statistics Agency
HMRC:	HM Revenue and Customs
DWP:	Department for Work and Pensions
CA:	Combined Authority
SQL:	Structured Query Language
ONS:	Office for National Statistics
SRS:	Secure Research Service
NCCIS:	National Client Caseload Information System
FSM:	Free School Meal
SEN:	Special Educational Needs
LSOA:	Lower Super Output Area
GCSE:	General Certificate of Secondary Education
KS4:	Key Stage 4
LAD:	Local Authority District
GOR:	Government Office Region
IDACI:	Income Deprivation Affecting Children Index
NEET:	Not in Education, Employment or Training
HE:	Higher Education
OfW:	Out of Work