Distr.: General

11 October 2022

English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Migration Statistics

Geneva, Switzerland, 26−28 October 2022
Item A of the provisional agenda
Improvements in use of administrative data for migration statistics

A machine learning approach to classifying UK long-term international migrants using administrative data

Note by Office for National Statistics

Abstract

The Office for National Statistics (ONS) is transforming population, migration, and social statistics using a combination of administrative and survey data. Estimating international migration using administrative data has brought significant challenges, including producing timely estimates. Administrative based estimates have inherent time lags, due to time needed for collecting and processing data, as well as interacting with services, e.g. for 12 months or more, in order to align to the UN definition of a long-term migrant.

ONS has developed Admin-Based Migration Estimates (ABMEs). The ABMEs currently use a rule-based approach to classify LTIMs. This requires up to five years of data to complete the classification for a whole cohort of migrants. We propose to assess how a predictive machine learning model may provide timely prediction at the aggregate level before the rule-based classification. We would also like to assess the use of machine learning to classify UK returning migrants and emigrants, and its robustness to significant changes in migrant behaviour, for example during the coronavirus (COVID-19) pandemic. We are developing these methods with experts in ONS Methodology using Home Office's Exit Checks data as a proof of concept, assessing accuracy of early predictions of eventual LTIM status and how those predictions can be used as early indicators of provisional LTIM numbers to make better provisional migration estimates. In this paper, we will discuss the progress and challenges in this project. As a work in progress, we welcome feedback on the methods.

*Prepared by Mingqing Wu and Michael Hawkes

I. Introduction

1. Migration is one of the three fundamental components of change in any population. Improving UK international migration statistics is high priority for the Office for National Statistics (ONS). With the ongoing impact of the coronavirus (COVID-19) pandemic and ongoing changes in migration policy following Britain's exit from the European Union (EU), the latest ONS estimates of net international migration suggest that around 239,000 more people came to the UK than left in in the year ending June 2021, driven primarily by non-EU immigration (ONS, 2022a). To support this effort there is a need for accurate and timely estimates of long-term international migrants (LTIMs) to and from the UK to inform the measurement and impact of migration on the UK population, to facilitate government policy and decision-making.
2. Current ONS estimates of international migration are experimental and use a combination of administrative data and statistical modelling. These experimental estimates are based on the best data that are available, and employ rules-based definitions to classify LTIMs in administrative data. Due to the inherent time lags in administrative data and the necessary time needed for confirmation of LTIM status, to complete the full picture of a whole cohort of migrants can take up to 2 years. To produce more timely estimates of long-term international migration, new data and methods are needed to provide more timely classifications of migrants as either long-term or short-term.
3. This paper describes initial research into machine learning approaches for classifying LTIMs recorded in administrative data alongside short-term international migrants (STIMs), and migrants who have not yet clearly met rule-based criteria to be defined as STIMs or LTIMs. The key goal of such a method would be to predict the probability that a migrant recorded in admin data will be a LTIM before their true short-term or long-term status is known. These probabilities, predicted for all potential LTIMs, could then be used to produce provisional aggregate estimates of LTIMs that include known LTIMs (i.e. those that meet rule-based criteria) and predicted LTIMs whose true status is not yet known. We are using the border-control based Exit Checks data as our initial dataset to develop a proof-of-concept method for immigrants only. This paper will also provide information on our progress towards transforming population and migration characteristics, and the challenges we face.

II. The need for new methods

A. International migration is changing

4. Recently there have been observations of changing patterns in UK international migration, driven by impacts of COVID-19 pandemic, the UK's exit from the EU, and the introduction of a new immigration system at the end of 2020 (Home Office, 2020b).
5. For example, the COVID-19 pandemic has presented a significant challenge to the UK and to the statistics we rely on. We have compared IPS migration estimates in the same period in previous years with the numbers of passenger flows from air, train and ferry data in March 2020 and found that travel by all routes declined steeply. Air travel declined sharply in Quarter 2 2020, as strict lockdown measures and travel restrictions were introduced. On the other hand, ferries and Eurotunnel did not fall as much and therefore made upa corresponding proportional increase in travel via ferries and Eurotunnel (ONS, 2021).

6. Visa processing centres also had to be closed in March 2020. Our previous analyses show a marked decrease in work, study and family visas applied for and granted to non-EU nationals during Quarter 2 2020 (ONS, 2021).

7. Following the UK's exit from the EU, there have been ongoing changes in migration policy. For instance, as of 1 January 2021, EU citizens who wish to move to the UK will be subject to the same rules as citizens from the rest of the world, except Irish citizens who can continue to move to the UK without restrictions under separate arrangements. EU citizens already living in the UK had to apply to the EU Settlement Scheme if they want to continue to live in the UK.

8. Due to changes in international migration such as described above, there is a high level of interest in understanding how migration patterns are changing and what this means for society and the economy. Robust and timely population and migration statistics are needed for policymaking in response to these changes, and they underpin a wide variety of other statistics that support decisions and inform public debate. While NS has undertaken a broad programme of research on the transformation of migration statistics, the COVID-19 pandemic accelerated our need to innovate data and methods, including administrative data and modelling methods to produce timely measures of international migration to meet a broad range of user needs (ONS, 2022a).

B. ONS statistics are transforming

9. ONS aims to provide the best insights on population and migration, working with other government departments and using a range of new and existing data sources to meet the needs of our users. This is increasingly important in a rapidly changing policy and societal context, where we know our users need better evidence to support decision-making at both national and local levels.

10. Current ONS population statistics rely heavily on the decadal census. While this provides granular data at the lowest levels of geography every 10 years, it delivers less detail for the interim years and the quality of population estimates declines as we move further away from the census year.

11. The historical source for UK international migration was the International Passenger Survey (IPS), which recorded migrants' intentions to remain in or out of the UK in the next twelve months. The IPS has some limitations with respect to measuring immigration and emigration, as it is a sample survey and not every migrant to or from the UK is interviewed. IPS also does not capture all asylum seekers who may be entering or leaving the UK. IPS is based on initial intention of stay period and does not take into account the changing intentions of passengers. This also does not capture those who are crossing the land border between the UK (Northern Ireland) and the Republic of Ireland. There is broad consensus that the IPS has been stretched beyond its original purpose, and ONS is now considering additional data and methods to transform migration statistics (ONS, 2018).

12. Additionally, the IPS stopped in March 2020, due to the COVID-19 pandemic. Although re-launched in 2021, IPS was more focused on travel and tourism statistics. Due to the restrictions on time and locations leading to some low numbers of contacts, the new IPS data can obscure the migration patterns, especially patterns (e.g. seasonality) in time series data (ONS, 2022b). Such changes to the IPS accelerated the need to use administrative data sources and new methods to fill the evidence gap (ONS, 2022a).

III. Classifying LTIMs in Exit Checks data

A. The Exit Checks data

13. The initial administrative data we are using to develop our proof of concept machine learning classifier are derived from the data matching system and analytical capability built by the UK Home Office's Exit Checks Programme, the Initial Status Analysis (ISA). We use data from the Home Office ISA system, which combines visa and travel information to link an individual's travel movements into and out of the country. This dataset is known as the Exit Checks dataset (Home Office, 2020a).
14. These data include Advance Passenger Information (API) and Travel Document Information (TDI). API is passenger data submitted in advance of travel for most scheduled aviation journeys, while TDI is passenger data collected at the point of departure for other modes of transport.
15. There are also data from case working systems related to (out-of-country) entry clearance visa application casework and (in-country) casework e.g. on extensions of leave to remain, as well as biometric details submitted prior to visa applications (we do not have access to such information) and passport examinations data collected upon entry into the UK. This data helps mitigate any gaps in inbound API coverage.
16. These data are matched in order to produce an 'identity' and to determine the current compliance status of an individual. Each individual within the ISA system is allocated a unique identifier which consists of biographic details (such as name, passport or travel document number, date of birth, nationality and gender) and associated events (such as travel in or out of the country or periods of leave granted). The resultant dataset is termed the 'Initial Status Analysis'.
17. As with all large complex data collections based on administrative data, the data received may not always be complete and fully accurate. There are also a number of ways in which a traveller's departure may legitimately not be recorded by the system, for example where outward travel is by the Common Travel Area (CTA), or when a traveller unfortunately dies in the UK.
18. For more information of Exit Checks data, please refer to the Home Office documentation (Home Office, 2020a).

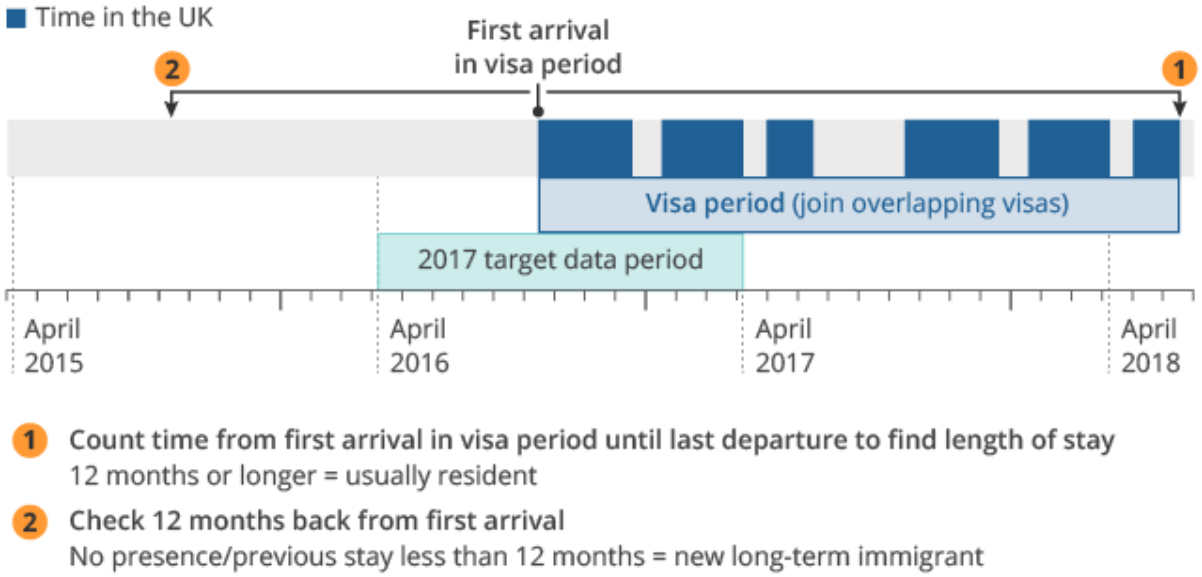
B. The current classification method

19. The current classification method for non-EU LTIMs adopts a static deterministic approach based on the actual outcomes of past activities in the Exit Checks data. We adopt the UN definition for this rule-based process for LTIMs: "A person who moves to a country other than that of his or her usual residence for a period of at least a year (12 months), so that the country of destination effectively becomes his or her new country of usual residence. From the perspective of the country of departure, the person will be a long-term emigrant and from that of the country of arrival, the person will be a long-term immigrant" (UN, 1998).
20. A migrant's time in UK is calculated on the basis of the first arrival date and last departure date in the visa period (If the visas are overlapping, they are joined together). If the time in UK is longer than 12 months, each period of time out of UK is shorter than 12 months and there is no previous stay in less than 12 months prior to

the first arrival date, then the migrant is classified as a new LTIM. This is illustrated in Figure 1 (ONS, 2020). A separate paper has been submitted to the UNECE Migration Statistics Expert group on the current estimation method: "Using administrative data to produce timely estimates of migration for the UK".

Figure 1

Rules to label the LTIMs in Exit Checks data



Source: ONS (2020)

C. The usefulness of supervised machine learning to classify UK LTIMs

1. Timeliness

21. Due to the time lags in the current classification method, we propose to assess how a predictive supervised machine learning model may provide timely prediction for individual immigrants recorded in administrative data before the rule-based classification can be applied. These individual predictions could then be aggregated to contribute to provisional estimates of LTIMs.

22. Supervised learning starts with manually labelled data. We use the current rule-based classification method to label the LTIMs as described above (Figure 1). Labelled historical data are then used to train the supervised learning model, so that predictions can be made on the LTIM status of migrants whose arrival date is too recent to apply the data definitions. ONS are investigating implementing alternative definitions to be applied in Exit Checks data, and we may investigate using these for generating training labels in future (ONS, 2020).

23. The supervised methods learn the relationship between input features that describe migrants (e.g. visa type, visa end date, days since arrival) and their labelled outcome as an LTIM or a non-LTIM. The trained models then predict of the probabilities of new migrants becoming LTIMs based on their input features. Therefore, machine learning models can potentially make predictions based on individual characteristics

without the need to wait for at least 12 months of data to apply rule-based data-definitions.

24. Using a supervised machine learning approach has the additional benefit of not needing to develop and maintain new assumptions and rule-based data categories for recent immigrants to achieve the classification goal before we have confirmation based on arrival and departure dates. In principle, where we might try and manually define certain categories of migrants with particular characteristics as being more or less likely to be LTIMs, in learning the relationships between the input features and the outcome the models will approximate a function that uses the available information to make that prediction i.e. the models could allow predictions to be made without explicit human specification (Arthur Samuel, 1959). The models may also adapt to changes in behaviours if they are updated over time (i.e. re-trained) with newly-labelled data that capture those changes in behaviour.
25. Predictions from machine learning models would be subject to error and uncertainty, and it is likely this error and uncertainty will be larger for more recent immigrants, but they could provide the basis for provisional LTIM estimates with increasing certainty over the life of a cohort (until they are able to be confirmed with a rule-based approach). They could also be used as indicators for trend exploration and early planning.

2. Fractional weighting and calibration

26. Calibration is an important aspect of training machine learning classifiers. It gives insight into model uncertainty, which can be later communicated to end-users or used in further processing of the model outputs.
27. Technically speaking, the goal of classification is to assign predicted classes to unlabelled data. However, it is also useful consider the model-predicted probabilities underlying these predicted classes. Such probabilities could be interpreted as fractional weights, and are useful for analysing the shortcomings of the model, and potentially for presenting uncertainty to end-users. Using fractional probabilities may also be beneficial if they result in reduced bias in provisional LTIM estimates compared to classification (Zhang, 2020).
28. In machine learning, a model is considered well-calibrated if the prediction probabilities output by a model correspond with the actual probability that the model is correct on that sample. For example, if a model predicts a sample to be in the positive class with a probability score of 0.8, it can be expected that the model would be correct in that prediction about 80% of the time.
29. The calibration of a model can be visualised by plotting a calibration curve – dividing the model's predictions into discrete bins based on the prediction probability, and plotting the proportions of each bin correctly predicted against the mean prediction probability of each bin. A perfectly calibrated classifier would thus be plotted as a straight line between (0,0) and (1,1).

IV. The machine learning approach

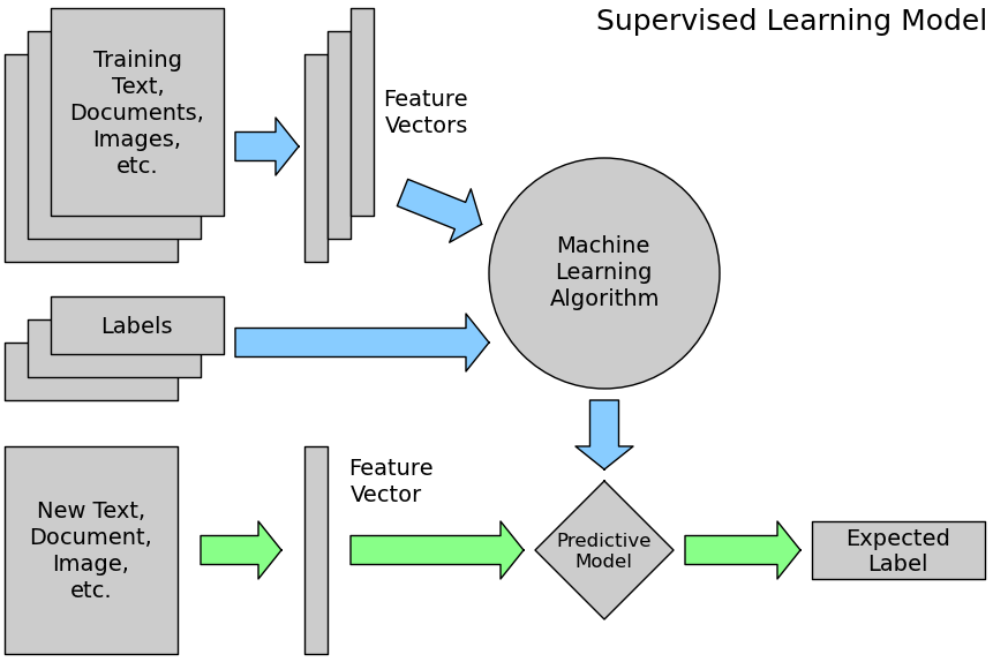
30. This project will use Exit Checks data as a record of migrants who are non-EU nationals to develop proof of concept methods. Pending the outcome of this initial research, future work will also consider a similar task using other administrative data

sources such as the Registration and Population Interactions Database (RAPID) from the Department for Work and Pensions (DWP) for EU nationals.

31. The classification uses supervised learning, which requires labels of both positive and negative examples in training the model. In our case, the positive label will be the LTIM. The machine uses the insights it gains from studying the LTIM and non-LTIM data to consider the features in the new data to help make predictions of probability of migrants becoming LTIMs. The machine learning approach enables us to predict the outcomes of migration statuses before the confirmation reaches the end of 12 months. The simplified supervised learning process is illustrated in Figure 2 below.

Figure 2

Supervised Learning overview



Source: Pedregosa et al. (2011)

A. The development practices

32. We follow UK Statistics Authority's data ethics guidance and work closely with the ONS Data Ethics team and completed an ethics assessment for this project. We only access the personal information in the ONS secure development environment and only aggregate outputs are output of this environment, subject to approval of the Disclosure Control Officers. No personally-identifiable information (PII) are used in the modelling work.

33. We follow the guidance of the UK Government Analysis Functional Standard and the quality assurance of government models (AF, 2021). We log our assumptions and decisions throughout the project to ensure quality of the development, reviewing with the Quality Assurance team as well as more independent expert panels. We also use the quality assurance framework to guide our reproducible coding and adopt reproducible pipelines to avoid human errors (GSS, 2022).

B. The base population

34. The model focuses on the long-term immigration of non-EU nationals as a proof of concept.
35. We start with the immigration component only, restricting our study to those who have a visiting history before the COVID-19 pandemic (April 2015-April 2019). This is because those are the majority of the population in the data, and it will allow us to develop baseline methods across the period where Britain exited the EU before we tackle the subsequent changes in migrant behaviour brought on by the COVID-19 pandemic.
36. Pending the development of the initial proof of concept method we will also investigate predicting LTIM status for immigrants without a visit history, and emigrants.

C. Training and testing

37. A sample of the migrants recorded in Exit Checks data are labelled within as either LTIMs or non-LTIMs based on the UN definition of LTIMs as described above. A set of supervised learning algorithms have been tested on the samples of covariates and labels in the non-EU data to learn the patterns in the data, before making predictions on the holdout data. At present, we are randomly splitting our dataset into a training set used to fit the model, and a testing set used to evaluate it, but we are also looking into stratifying the data such that LTIMs and non-LTIMS are equally represented in the training data. Additionally, we are looking at the possibility of dividing the dataset into training and testing splits based on time (ie. Training the model on all people who interacted with the Exit Checks system before a certain date, and testing on those who interacted after), in order to better evaluate how well a model trained on historical data performs on future data.

D. Algorithms

38. As no single machine learning algorithm will be useful for all use cases, we are testing three supervised learning algorithms to see which algorithm can provide the best results and interpretation.
39. We are treating LTIM classification as a binary classification problem to predict which migrants will become LTIMs, and are initially considering logistic regression (LR), random forest (RF), and XGBoost (XGB). LTIM classification could also be interpreted as a survival analysis problem with censored data, and we will also consider whether alternative methods may be more appropriate.

E. Model optimisation and performance assessment

40. We intend to optimise the trained machine learning models by iteratively tuning the hyperparameters using an automated hyperparameter optimisation technique such as Grid Search, with Accuracy (the number of correct predictions divided by the number of samples) chosen as the scoring metric to maximise in this process.

41. It should be noted that accuracy can be a poor metric to use when there is a significant class imbalance in the training data, as it can lead to scores that look impressive while not actually being a skilled classifier - for example, if 90% of samples in a dataset belong to the positive class, then a baseline model that only predicts samples as belonging to that class would have an accuracy score of 0.9, making it look intuitively like a skilled model.
42. Upon exploration, we concluded that Accuracy would be a useful metric for our project, as the classification classes are not too imbalanced and both the positive and negative class are of close to equal importance.
43. Additionally, we will be recording the Recall, Precision, and F1 scores of each model, in order to gain a fuller understanding of the model's performance, but these were not chosen to be directly optimised for, as they tend to be more useful in cases where only the positive class in the classification problem is of interest, or where there are different costs associated with misclassifying members of the positive or negative class. Similarly, Area under Receiver Operating Curve (ROC-AUC; the area under the curve plotting True Positive Rate against False Positive Rate for various model thresholds) will also be recorded in order to evaluate the discriminatory power of the model.
44. We will also consider using the log-loss of the model as the metric to optimise, as this could improve the prediction probabilities associated with each prediction, which would be useful when using probabilistic weighting to determine immigration statistics (rather than counting the number of people predicted in each category. However, this is dependent on further research to determine whether this would be beneficial.

F. Feature selection

45. The Exit Checks data are rich in information, not all of which is specifically useful in predicting whether a person is a Long-Term Migrant or not. Including every feature available to us could thus lead to models which are over-fit to statistical anomalies in the training data, and result in worse performance. We are initially building baseline models using a small number of attributes from the data and engineered features and we will pursue more complex feature engineering as necessary. To evaluate the importance of each feature, we will be using the feature importance scores output by the models – the average gain in purity from splits on that feature for XGBoost and Random Forest models, and the coefficients for each feature in the Logistic Regression models. We will then be able to manually inspect the features to identify those with little to no predictive power, as well as using Recursive Feature Elimination techniques to iteratively train models with fewer features, removing those features with the least predictive power.

G. Monitoring and retraining

46. Machine learning model performance needs to be monitored on an ongoing basis to detect possible model drift, where model performance degrades over time, especially when there are significant changes in data, patterns and concepts used in the model, such as changes in visa definitions or migration patterns. In such cases, models may need to be retrained over time where necessary. We will have to consider this given the changing nature of migration behaviour we've outlined above.

H. Lessons learnt

47. We are currently investigating the performance of the first baseline models that we have trained, and hope to publish initial results in a progress report later this year.
48. One of the lessons we have learnt from this project is that understanding of the data is essential. Datasets can be limited in size and they might not be representable for the full population, or the data capturing process might have not accounted for potential biases. Biases often only become apparent after thorough data analysis or when the relation between model predictions and the model input is analysed. The understanding of the data also guides the feature engineering and selection, as well as the choices of algorithms and performance metrics.
49. Supervised learning models also require a good level of expertise to structure appropriately.
50. We have realised the need of testing different algorithms, samples e.g. of different periods, characteristics/behaviours and sizes. These can all make an important impact on the outputs.
51. We also realise that we may need a suite of models, rather than just one model, to train and predict on different sub-populations. For instance, we may need separate models to predict LTIM status for different categories of migrants e.g. those with visit histories and those on their first visit.

V. Next steps

52. Using supervised machine learning to classify LTIMs with administrative data has some unique potential advantages to other classification methods, but it also presents some challenges that need to be handled with care.
    - Feature engineering to make the most of the information available in the Exit Checks data
    - Detailed analysis of the performance of the three algorithms for immigrant cohorts, including investigation of accuracy and bias of the predictions across and within cohorts.
    - Consideration of extensions to supervised machine learning that incorporate aspects of survival analysis.
    - Assessment of the need to retrain models over time to address model drift, and potential monitoring and retraining approaches.
    - Development of a proof of concept model for emigration.
    - Investigation of how to integrate machine learning LTIM predictions into the wider ONS ABME methodology, and how uncertainty from these predictions can be propagated and communicated to users.
    - Application of similar methods for classifying EU national LTIMs in RAPID data
53. We would also like to explore some areas that are currently understudied using other machine learning methods. For instance, unsupervised machine learning algorithms may be able to identify patterns within administrative data and help shape a heuristic

ruleset for circular migration and UK LITMs where no definitive LTIM labels can be applied. This project is still at its early stage and we welcome comments and suggestion to take this work forward.

VI. References:

54. Analysis Function (2021) Government Functional Standard GovS 010: Analysis, available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1011798/CO_Govt_Functional_Std_GovS010_Analysis_v2_Final_WEB.pdf
55. Government Statistical Services (2022) Reproducible Analytical Pipelines
56. Available from: https://gss.civilservice.gov.uk/reproducible-analytical-pipelines/
57. Home Office (2020a) Home Office statistics on exit checks: user guide, available from: https://www.gov.uk/government/publications/home-office-statistics-on-exit-checks-user-guide/home-office-statistics-on-exit-checks-user-guide
58. Home Office (2020b) Impact Assessment for changes to the Immigration Rules for Skilled Workers, available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/936121/Revised_Impact_Assessment_for_the_Skilled_Worker_Route_signed.pdf
59. Office for National Statistics (2022a) Long-term international migration, provisional, year ending June 2021, available from https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/bulletins/longterminternationalmigrationprovisional/june2021#main-points
60. Office for National Statistics (2022b) Estimates of overseas residents' visits and spending in the UK, available from https://www.ons.gov.uk/peoplepopulationandcommunity/leisureandtourism/datasets/estimatesofoverseasresidentsvisitsandspendingintheuk
61. Office for National Statistics (2021) Using statistical modelling to estimate UK international migration, available from https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/usingstatisticalmodellingtoestimateukinternationalmigration
62. Office for National Statistics (2020) Exploring international migration concepts and definitions with Home Office administrative data, available from:
63. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/exploringinternationalmigrationconceptsanddefinitionswithhomeofficeadministrativedata/2020-02-14
64. Office for National Statistics (2018) Report on international migration data sources
65. Pedregosa, Fabian; Varoquaux, Gaël ; Gramfort, Alexandre; Michel, Vincent; Thirion, Bertrand; Grisel, Olivier; Blondel, Mathieu; Prettenhofer, Peter; Weiss, Ron; Dubourg, Vincent ; Vanderplas, Jake; Passos, Alexandre; Cournapeau, David; Brucher, Matthieu; Perrot, Matthieu; Duchesnay, Édouard (2011) Supervised Learning overview, in Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011.; 12(85):2825−2830, 2011. Available from:

https://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/tutorial/astronomy/general_concepts.html

66. Sapon, Muhammad & Ismail, Khadijah & Suehazlyn, Zainudin & Ping, Chew & Malaysia, Nasional & Lumpur, Kuala. (2022). Diabetes Prediction with Supervised Learning Algorithms of Artificial Neural Network.
67. United Nations (1998) Recommendations on Statistics of International Migration, Revision 1, available from: http://data.un.org/Glossary.aspx?q=long-term%20migrant
68. Zhang, LC. (2020) On provision of UK neighbourhood population statistics beyond 2021. arXiv [Preprint]. Available from: https://arxiv.org/abs/2111.03100