Fractional counting for administrative based population statistics

(January 2021)

Table of Contents

Key Messages **2**

Purpose 2

Key Asks of MARP 2

Executive Summary **3**

Paper Title **4**

Introduction 4

Context and Background 5

Strategy for developing an initial hypercube prototype 7

Implementation 7

Discussion 13

Future work 16

References 17

Annex **18**

Key Messages of Paper

Purpose

- This paper introduces and sets the scene for the fractional counting approach for producing administrative based, multivariate population outputs. It explains how the approach works alongside (and extends on) other administrative based projects under the Population and Migration Statistics Transformation Programme. It demonstrates a simplified hypercube with some component models to show the potential for the use of fractional counting to produce administrative based population and characteristics estimates, at a more granular level. The aim is to provide proof of concept and to inform the direction of future work.

Key Asks of MARP

The panel are asked to comment on the paper and consider the following questions:

- Should we pursue this approach for estimating small area multivariate characteristics?
- Do you have any comments on the work undertaken?
- Do you have suggestions for the component methods/ models?
- Are you aware of any similar work?

Executive Summary

This project assesses the feasibility and benefits of a fractional counting approach for population statistics that are derived from integrated data sources (including administrative and survey data). The term "fractional counting" is used to refer to the process of calculating statistical outputs from a weighted data set, where the weights account for under or over coverage in the underlying administrative records and reflect other uncertainties such as the correct postcode address for that individual, as well as the reliability and timeliness of the data sources.

The main principle in constructing the weighted dataset is that the adjusted census database represents the true 2011 population and can be used as a comparator for the initial development of weights for the administrative-based dataset. For example, the probability that the postcode address recorded in an administrative source is the correct address for a person could be estimated based on the assumption that the census record represents the true address. The weighted administrative dataset could subsequently be rolled forward over time using updated administrative data, but clearly as time elapses from the census the relationships based on the Census responses will become less relevant and a mechanism will be required for updating the models on which these are based too.

It is thought that fractional counting may have potential to provide more robust estimates than can be achieved when using integer counts, which typically involves making assumptions in order to decide how to classify conflicting information across sources. The focus is to provide a framework for estimating detailed multivariate tabulations rather than estimating the overall population totals, for example gender/age/labour market status.

We are currently at an early stage of the work; this paper primarily sets the scene for fractional counting, presenting how the development of the fractional dataset works alongside (and extends on) other administrative based research outputs such as the Administrative Based Population Estimates (ABPEs). It demonstrates a simplified prototype and some component models to show the potential for the use of fractional counting to produce administrative based population and characteristics estimates, at a more granular level.

Progress to date looks promising, with each of the considered models being effective both in terms of pure classification metrics as well as in reproducing high-level aggregated weights, but has been limited as we have not been able to obtain all the key administrative sources for 2011, particularly 2011 Patient Demographic Service (PDS) and Customer Information System (CIS) data. Both these sources have proved to be useful in estimating the correct address of people registered on administrative sources in previous work. We have improvised with data currently available within the ONS secure systems in order to understand and train the component models but will need to align the reference dates of source data in order to fully evaluate the methods and properties of our weighted dataset.

Future work will consider parametric versus algorithmic methods for rolling forward the fractional datasets and how to assess the quality and wider properties of the outputs.

Fractional counting for administrative based population statistics

Daniel Ward, Jonathan Rees, Iva Spakulova, Greg Payne, Alison Whitworth

1. Introduction

This project investigates a fractional counting approach for population statistics that are derived from integrated data sources (including administrative and survey data). The approach is to build a weighted, record level dataset which can be used to estimate multivariate population outputs at any geographical level. The dataset can include characteristic information such as age, sex and labour market activity, where data are available in the administrative records for the attributes, and in this case would form a multidimensional cross tabulation or "hypercube". We refer to the dataset as the hypercube throughout this paper; by hypercube we mean a table that contains weighted counts of people for each combination of characteristics, for example gender/age/location. The fractional counting approach is used to account for uncertainty in the administrative records, so for example the probability that the address recorded on the administrative record is the correct address and the reliability of that source. A source might be considered more reliable if there is evidence of recent activity for that individual. The fractional counting has potential to provide more robust estimates than can be achieved when using integer counts, which typically involve deciding which address recorded on the administrative source is the correct address and allocating the individual, by estimating residency weights across available addresses. The focus of the methodology is estimating those detailed cross tabulations via the hypercube, rather than estimating the overall population totals. One key assumption is that robust population totals exist.

The approach was suggested by Professor Li-Chun Zhang (2019) "On provision of UK neighbourhood population statistics beyond 2021" (paper provided). This ONS project will produce a prototype of the weighted dataset suggested in his paper and investigate how updated administrative records can be used with regular survey data to continuously roll this forward over time. Without a future decennial census, a mechanism for periodic auditing or quality assurance of the rolled forward hypercube will also be essential. The project will investigate the extent to which a large-scale survey (such as the Integrated Population and Characteristics Survey (IPACS)) can achieve this and make recommendations for any specific requirements going forward. The project in its entirety provides a potential system for creating continuous estimates for the component population subgroups in the hypercube. However, it does not provide a solution for creating data where it does not exist in the administrative or survey sources, or for creating coherence between migration outputs derived independently and population totals based on the fractional hypercube. How to integrate these aspects will be considered at a later stage and are the subject of other projects under the ONS Population and Migration Statistics Transformation (PMST) programme.

We are currently at an early stage of the work; this paper primarily sets the scene for fractional counting, presenting how the development of the hypercube works alongside (and extends on) other administrative data based projects such as the Administrative Based Population Estimates (ABPEs). It demonstrates a simplified hypercube with some component models to show the potential for the use of fractional counting to produce administrative based population and characteristics estimates, at a more granular level.

2. Context and Background

ONS currently uses the cohort component method to provide local authority (LA) population estimates by age (single year of age) and sex, on an annual basis between census years. This method updates census estimates at an aggregate level by accounting for the components of population change (births, deaths and migration). People are counted at their census address and then administrative data is used to account for internal migration and lower level distributions of international migration, whilst survey data is used for higher level estimates of international migration (regional). The changes are applied to the census-based estimates at LA level and then these population totals are broken down to small area level using administrative data as indicators of recent small area population distributions/or change. A top down approach is used whereby the more detailed (lower level geography) estimates are constrained to higher level (more accurate) population totals creating coherent population size estimates at all levels.

Detailed population estimates by population characteristics such as ethnicity, or employment status are published as part of the decennial census outputs. Between census years some key characteristics are estimated at country or regional level using dedicated surveys and, where possible, small area estimation techniques are used to draw strength from administrative data for more detailed estimates. For example, LA level estimates of unemployment are estimated using the Labour Force Survey (LFS) and Benefit Claimant Counts in a model-based approach.

ONS's goal is to better meet our users' needs using firstly government-held administrative data, and other data. The census gives us a snapshot of society, but it only happens once every ten years; and over time the value of the data to decision makers decreases. The Office for National Statistics is taking forward work to transform the population and social statistics system. The National Statistician will make recommendations to the government in 2023 on the future of population statistics, including what measures will be necessary to support the new approach; this may include another census.

One of the benefits in making greater use of administrative data is the potential for more frequent small area statistics such as those produced currently only once every ten years. Users ideally require multivariate outputs i.e. population totals by age, sex, employment status, ethnicity etc, for small geographical areas. ABPEs have been produced and published as research outputs under this programme.

Version 1 and 2 of the Statistical Population Dataset (SPD) derived for the ABPEs were produced by linking administrative records. Version 1 (V1) counted people on a fractional basis where records were linked and the different sources for each individual allocated equal weighting. A person with records on both the Patient Register (PR) and Customer Information Service (CIS), for example, would be represented with a weight of 0.5 for each source and counted on a fractional basis if the addresses recorded on each straddled target geographies. Treating the sources with equal weighting was found to result in biased distributions however, so version 2 (V2) of the SPD adopted an approach that combined two methods for allocating persons to addresses; firstly to use the NHS Personal Demographic Service (PDS) movers extract to determine logically which address is correct. Where use of the PDS data did not resolve the conflict a modelling approach was used to determine the most likely address (the models are summarised in Appendix 3). The individual was then assigned, as a whole, to the address scoring the highest probability.

Version 3 (V3) of the SPD counts people on an integer basis using an activity-based approach to reduce over-coverage in the admin-based population dataset. The objective is to build rules specific to age groups that make best use of the data sources that provide the best coverage for that group. For example, the CIS is used to identify individuals ages 16-64 through evidence of economic activity, with the Higher Education Statistics Agency (HESA) dataset, which captures data on students in higher education, being used to confirm or add additional records. Similarly, the English and Welsh School Census' (ESC, WSC) are used alongside PDS data to identify school aged individuals. Those with activity on at least one administrative data source are included in the SPD and, to prevent those with activity on multiple sources receiving duplicate counts, the data sources are linked and only one record selected.

One of the key objectives of this version of the SPD is to drive out over-coverage from the administrative based estimates so all records included should have a sign of activity within the 12 months prior to the reference date of the ABPE or appear in the same address and have a relationship with an active person. This has resulted in intended higher levels of estimated under-coverage than seen in previous versions of the ABPEs, but by reducing the overcount it is thought to provide a more appropriate platform for combining with a Population Coverage Survey to produce coverage-adjusted population size estimates using Dual System Estimation in a similar way to the methods used to obtain Local Authority (LA) coverage adjusted estimates for the 2011 census.

In summary, methods for estimating population totals have primarily focused on counting individuals on an integer basis and sought evidence to classify characteristic values where there is uncertainty. The exception is an early iteration of the SPD where a fractional counting approach was tested making the simple assumption that records for an individual are equally valid. This project will assess the feasibility and benefits of counting entirely on a fractional basis using estimated probabilities that reflect uncertainty in the underlying sources. It will draw upon the findings of research undertaken for the ABPEs, particularly the modelling undertaken for V2 of the SPD to predict the probability of correct address. It will also draw upon the methods that make use of "signs of life" to eliminate over-coverage and also those for determining usual residence status.

A "simulation and estimation" framework has been developed as part of the PMST Programme in order to test, develop and compare the performance of different estimation methods. This involves simulating data which represent the characteristics of a "true" base population and then also simulating administrative and survey data for this population. The overall design of the simulated administrative and survey data relies on observed coverage patterns for administrative data, and assumes levels of non-response that are typical for a coverage survey. In essence, the framework creates a "level playing field" on which to compare proposed estimation methods and to allow broad "stress testing", to establish under what conditions estimators do well. The framework development has focused on population size estimation, but it could be potentially extended to include the estimation of multivariate characteristics.

To date the Fractional Counting project prototype has been developed with the original data sources outside of the simulation framework. As iterations of the framework develop and capture more of the underlying characteristics of the sources, we will investigate creating a simulated fractional database within the framework in order to test and compare its component methods. However, at present, the aim of the framework is to develop a methodology for estimating total population size, and any stratification of estimates remains at a high level only. In a top down approach, as currently used for population statistics, these higher level administrative based estimates could provide benchmarks or calibration totals for the more detailed multivariate outputs obtained from the fractional population dataset. Further detail of the simulation work is available in Archer, R. et al. 2020.

3. Strategy for developing an initial hypercube prototype

The main principles of the fractional counting framework are that the adjusted census database represents the true 2011 population and can be used as a comparator for the initial 2011 admin-based dataset. In its simplest form the ratios of the estimates from the census and combined administrative data for each multivariate population subgroup, form the initial weights for the hypercube. In order to obtain more accurate population outputs, the weights need to be refined to account for different sources of uncertainty such as the residential address where this differs across sources. For 2011, probabilities for the address could be estimated using the "correct" census records and relevant covariate data for linked records. The adjusted administrative dataset can be rolled forward over time, using updated administrative data, but clearly as time elapses from the census the relationships on which the weights depend will become less relevant and a mechanism will be required for updating the models too. Without another census, surveys (where the data collection is controlled and has known inclusion probabilities) will be required to refit the models. The estimation precision is determined by the size of the updated datasets but is likely larger than the initial census-based estimates so the new weights would need to take this into account too.

The advantage of the fractional counting approach is that it has potential to reduce bias in population estimates, providing the error in the administrative data sources can be correctly defined and modelled. Zhang (2019), for example, demonstrates the properties of fractional counting for local area population totals when accounting for error in the administrative residential address records. The weighted hypercube also has the potential for producing estimates by any multivariate tabulation where person level data for the component univariate characteristics are available and can be linked to the initial unadjusted population dataset. It is not a trivial task to estimate and account for all the underlying uncertainties and processes within the hypercube however, the properties of the prototype developed in this project will be fully evaluated and a method for capturing the remaining uncertainty (and underlying variations) in estimates will be investigated with a view to producing accuracy measures. This will include an auditing process integrated within a rolling system for continuous outputs.

The hypercube is also dependent on the ability to successfully link data sources, initially linking longitudinal administrative data, and then surveys and other sources as part of the rolling process. For our prototype we use the Demographic Index (DI), constructed within ONS to enable linkage between sources using only an ONS ID. The index allows data to be made accessible without retaining the personal identifiers and provides an anonymised database of longitudinally linked administrative data. Sources in the Demographic Index include administrative registers of different government departments and authorities to ensure broad coverage of the population. It uniquely identifies every person who has registered or interacted with the selected administrative systems. Matching error will propagate throughout the hypercube, the impact of this will be considered in the quality evaluation described above.

4. Implementation

The steps for implementing the fractional counter and rolling forward the hypercube are outlined below. A summary explanation of these (as described in Zhang 2019) is provided in Appendix 1.

Steps for developing the hypercube:

I. Develop the initial extended population database from administrative sources using 2011 data.
II. Estimate the probability that the addresses recorded on the administrative data are the correct residential address for that individual using census responses as the truth.
III. Develop an indicator that each individual within the database belongs to the target population, in our case this is usually resident population (i.e. has been or intends to be resident in the UK for a period of one year).
IV. Calculate the initial hypercube by summing the probabilities (of whether each address recorded by the administrative process is the correct address) and multiplying by (0,1) indicators that the person is part of the target, usually resident population and the address is within the target area of interest.
V. Extend the hypercube to include other key census "type" characteristics of the population.
VI. Develop methods for rolling forward the hypercube: Updating the initial administrative dataset and model parameters in a nearly continuous process over time
VII. Develop procedures for periodic auditing of the estimated population totals, perhaps using a purposely designed coverage survey in an audit-assisted approach.

To date we have worked on steps I to IV; constructed an initial extended administrative population database by age, sex and postcode of residence, and investigated different modelling approaches to obtain probabilities for weights that reflect uncertainty in the address. The tables are illustrative of the fractional counting approach and would need more work before providing valid population outputs.

Section 4.1 below provides a simple illustration of the hypercube demonstrating how the weight adjustments work. Section 4.2 describes our early work on developing residency models to obtain more informed address weights for improved distributions across geographies. It builds upon previous work within ONS to model the correct address of individuals recorded by administrative sources.

4.1. A simple illustration of the hypercube

The spreadsheet in Appendix 2 demonstrates how the weighted hypercube might be constructed. It starts with Census records for 20 fictitious respondents in a table and 3 admin data sources, appended in another table to represent an initial integrated administrative dataset. Each administrative source covers a subset of the population and contains some errors. Weights are calculated to 1) ensure that each individual is counted only once, 2) provide a crude representation of coverage error in the administrative sources, and 3) modelled weights to represent known uncertainty in the administrative records such as the correct address. The table representing the weighted hypercube contains a count of people for each combination of characteristics gender/age/location. The example is intended to represent how the weights might be constructed for the hypercube but does not show the full methods that will be used to derive the weights in our prototype hypercube.

4.2. Modelling address weights using linked administrative datasets

Our aim is to produce address weights for individuals in our hypercube, using covariate information from the range of available administrative datasets as well as taking into consideration the inherent reliability and usefulness of each individual dataset. These weights allow individuals to be fractionally counted at conflicting addresses instead of being assigned a single correct address,

The practical steps taken to accomplish steps I to IV of the hypercube development plan, can be summarised as:

I. Develop the initial extended population database from administrative sources using 2011 data (Section 4.1)
   a. Concatenate data from administrative sources linking through the DI (PDS, PR, HESA, ESC, WSC)
   b. Reduce the population database such that it contains a single row for each address linked to an individual (for cases where multiple sources contain the same address)
   c. Link to Census to create "truth" state, to model as dependent variable for each record

II. Estimate the probability that the addresses recorded on the administrative data are the correct residential address for that individual using census responses as the truth (Section 4.2)
   a. Sample prototype extended dataset (stratified random sample)
   b. Conduct feature selection
   c. Implement weight modelling using logistic regression, support vector machines and random forests
   d. Run hyperparameter optimizations for specified models
   e. Evaluate model quality based on ability to be generalised to unseen 'holdout' dataset

III. Develop an indicator that each individual within the database belongs to the target population, in our case this is usually resident population (i.e. has been or intends to be resident in the UK for a period of one year).

IV. Calculate the initial hypercube by summing the probabilities (of whether each address recorded by the administrative process is the correct address)
   a. Find sum of individual weights by age for comparisons with the 'true' population counts

As our hypercube was constructed using the DI linked to admin data sources and the 2011 Census, only including individuals identified on the Census, we don't currently consider step III and consider that each individual is a member of our target population (usually resident).

In order to cover a range of possible covariate relationships and modelling complexities, we initially consider the following models:

- Logistic regression (LOGR)
- Support vector machines (SVM) (Chang, C.-C. 2001)
- Random forests (RF) (Genuer, R. 2018)

The efficiency of fitting LOGR models will be beneficial when estimating the residency weights for larger sample / Census size populations, where machine learning based methods such as SVM and RF are more costly to fit on larger sets and can be more prone to over-fitting the available data. In order to quantify the fitting ability of the various models we measure the following metrics:

- Accuracy / Balanced accuracy
- Recall
- Specificity
- Precision
- F1
- AUC (Area under receiver operating characteristic (ROC) curve (Bradley, A. 1997))

These metrics are commonly used to determine the performance of machine learning / regression models, where the goal is to classify the data. Whilst we are not necessarily interested in the ability of the model to classify the data into "true" and "false", they can give us measures to compare between our models and to consider whether each model is over-fitting to the training sample when we apply the model to predict on the holdout dataset.

Due to the structure of the hypercube, in order to utilise the listed models, we also must make the assumption that an individual may be found on several records, and these records may have several conflicting addresses, the sum of the new weights across individuals may sum to more/less than 1.

The result should be that any estimates constructed from these weights should recover the higher-level population, with re-distributing LA level, with the LA level estimates able to subsequently undergo benchmarking to known totals.

In the following section we discuss in further detail how we constructed our hypercube samples, before deriving and selecting the covariate information.

4.2.1. Constructing hypercube sample for modelling

We constructed training samples (simulating a PCS, similar to the ABPE and ACID works) to be used in the initial feature selection and individual model hyperparameter optimisations as well as for the subsequent model comparisons when used to estimate population weights for independent holdout datasets. We take stratified samples (across all LAs) consisting of 1.5% of the hypercube, to represent our PCS. These samples are then linked to the census to be classified as either:

- Matching and thus flagged as a record containing the individuals' "true" address, or
- Non-matching and thus flagged as not being the individuals "true" address

These flags can then be used as the dependent variable for each of our models. Analysis of the initial model fits showed a propensity for allocating high weights for all records regardless of their true class, entirely mis-weighting the "false" addressed records. This is due to the records in the minority "false" record class having little influence compared to the much larger majority "true" class. Fitting to the heavily unbalanced (90% "true", 10% "false) samples results in high metric scores (>0.9) for all bar specificity (~0.2) and balanced accuracy (~0.6). As we are interested in distributing weights between the conflicting addresses, as opposed to simply classifying the addresses, we may be better served by balancing the classes within our sample. There are a number of potential ways to balance the sample, either over-sampling the minority class to produce simulated records, by under-sampling a subset of the hypercube removing overrepresented majority class records, or by restricting the sample to records with known conflicts.

We have initially implemented an over-sampling method to balance the classes and to give the models a better chance at weighting both the "true" and "false" records, having implemented a synthetic minority oversampling technique (SMOTE) on our training samples (Kovacs, G. 2019). SMOTE is a package within the imbalanced learn Python library, which can synthetically oversample minority class data in a binary classification data set. For example, if our class "0" has 1,000 entries and class "1" has 5,000 entries, SMOTE can synthetically generate 4,000 more "0" observations in order to balance out the data.

This is a particularly useful strategy when executing binary classification algorithms such as RF, which tend to underperform when we have an unbalanced data. It is worth noting, however, that not all classification algorithms require such balanced data.

SMOTE operates by synthesising new minority class instances *between* existing instances. In an instance where two class variables are used in the classification, this could simply be illustrated by drawing extra points along a straight line between two existing data points. In the below example (using the popular *iris* data set), the dark red circles represent the *real* instances of our minority class, with the light red circles being synthesised data points. This brings the class discrepancy up from 13:4 to 13:12, a ratio which will result in far higher accuracies when using RF in particular.

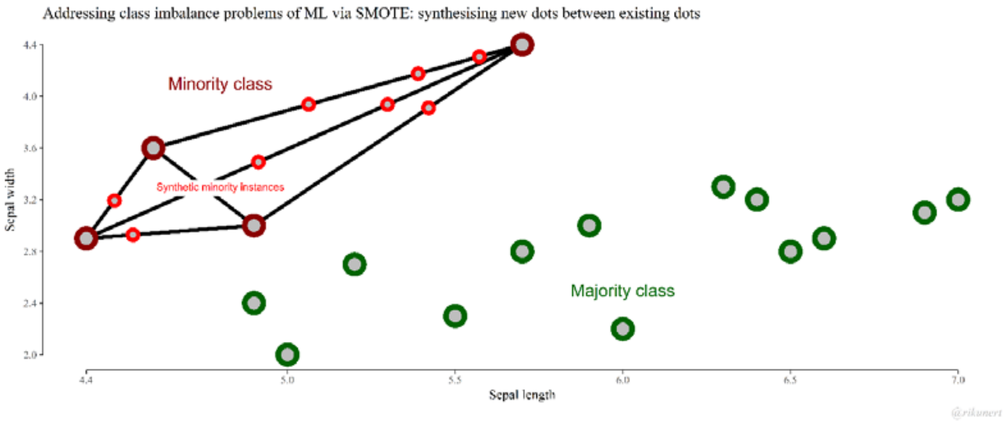


FIGURE 1. EXAMPLE OF SMOTE BALANCING METHOD

Our data has more dimensions than shown in Figure 1, but the same principle applies. New data points can be synthesised between real data points in *n* dimensions, resulting in a far more balanced dataset.

Whilst we used SMOTE to balance our sample classes, increasing the models' ability to correctly allocate weights to the "false" records, it is not the only method we could incorporate to improve the class representation. Alternatively, we can produce samples by under-sampling from a larger subset of the hypercube such that we arrive at a more even distribution of "true" and "false" records. It would also be possible to consider the subset of records where there is at least one conflict, however this has the potential to cause problems by creating a sample which actually contains more "false" records than "true", as the requirement is at least one conflict however with multiple admin data sources there is the possibility of including records with several conflicts.

4.2.2. Feature selection

This section describes the features available for inclusion in our modelling of address weights, both previously utilised and derived specifically, for details on covariate derivation see Appendix 4. For all individuals, and their potential address, we include seven covariates from the administrative datasets (age, sex, individual flags indicating source (PR, PDS, HESA, ESC, WSC)) as well as derive eight additional covariates.

The age of the record (i) both informs on an individual's status as usually resident, and acts as a measure for record reliability. The number of individuals registering at an address (ii) provides information on the turnover of population in an area according to each source. The count of consistent surnames at an address (iii) provides evidence of family potentially indicating reduced likelihood of misplacement. The fraction of conflicting records (iv) shows source address cohesion as a simple probability of finding each address across available records. The student status flag (v) aims to capture the behaviours of the student subpopulation, specifically with regards to conflicts due to incorrect placement according to the Census. The PDS activity flag has been shown to be highly indicative of a record containing the correct address according to ABPEs. However, as the PDS data used to construct our hypercube is more recent than the reference Census data we are missing data between these dates (2011 – 2016) and potentially introducing misplacement when constructing our "true" address flags by linking 2016 PDS to 2011 Census. The inclusion of more recent comparative data will increase the usefulness of this covariate and as such we continue with the development within our prototype. The flag indicating the presence of multiple individuals with the same surname (vii) can be used to determine the presence of a family or single/unmarried individuals. Finally, we derive the distance between subsequently updated addresses (viii), calculated by determining the haversine distance (as-the-crow-flies) between pairs of addresses in order of most recent update.

These particular covariates were selected for inclusion in our prototype modelling in order to broadly capture a range of features related to each admin database, in terms of each individuals' interactions with a specific source (frequency of interaction, recency of interaction), how they interact with a range of sources (source address cohesion), and the inherent design of each source (frequency of update, incorrect inclusion/removal of records).

Whilst additional covariates (such as income and benefits status) are planned to be included in the future and have been shown as effective as predictors of usual residence and correct address (ABPEs) we focus on these fifteen covariates in this paper. The potential impact of inclusion of these additional covariates is discussed further in Section 5.

4.2.3. Model selection, hyperparameter optimisation and specification

In prototyping the fractional counting methodology on our hypercube we are continuing to derive and add further covariate information to our model, as such the model specifications are tweaked so that we can understand the importance / usefulness of each variable and their interactions with other variables. Further to this, defining the importance of each introduced or removed variable is not trivial as the importance of a variable may be due to its (possibly complex) interaction with other variables. At this stage we consider relative rankings of the 'importance' of features as determined by

- RF impurity-based feature importance (normalized total reduction of fitting error attributed to each feature) (Louppe et al, 2013)
- Automatically generated LOGR model testing ('glmulti' (Calcagno & de Mazancourt, 2010))

Impurity-based feature importance involves measuring how the fitting error changes when individual variables are held consistent whilst another is varied, to determine the proportion of improvement attributable to each. To determine the importance placed in each variable when applying LOGR, we combine the results of multiple-model spawns, iterating through specifications until the model with the lowest fitting error is identified, with the results of comparing p and z-values. The ranking of variable importance for the different models can be found in Appendix 5. In brief however, it appears that features considered consistently important are:

- Age
- Postcode fraction

With the consistently least important shown to be:

- Distance
- PDS activity
- Student flag
- Source HESA

One thing to note from the importance analysis is how PDS activity is deemed of lower importance, as well as the absence of Source PDS as a variable of greater importance as suggested by the results of the ABPE work. This is likely due to the reduced PDS data we used to construct the hypercube, with data only available post 2016, the effects this may have and how the introduction of more up-to-date data may change the model specification will be discussed in Section 5. The results from the multiple-spawned LOGR models suggest that the optimal model includes all the currently available covariate information and as such we used all variables for each of the models (LOGR, SVM & RF).

To determine the optimal hyperparameters for both the SVM and RF models we conducted parameter sweeps. They covered the range of SVM cost (c), boundary dissipation (gamma) and error-tolerance (epsilon) values, to produce corresponding fitting metrics for comparison and to allow us to select parameters that result in model fits with different relative metrics. Similarly, for our RF models we again ran a parameter sweep to optimise 1) the number of trees, 2) the tree depth, 3) minimum samples per leaf and 4) minimum samples per split. The resulting hyperparameters for the SVM and RF models, used going forward with fitting our models and testing on the holdout samples are:

- SVM
    - C = 1
    - γ = 0.0625
    - ε = 0.1
- RF
    - # of trees = 100
    - Depth = 20
    - Minimum split = 2
    - Minimum leaf = 5

Using the resultant fits (to the simulated PCS) we test the performance of each model by predicting the address weights for individuals in separate holdout test samples. The holdout datasets act as intermediary proxies for the hypercube, before we implement the methods on the full-size hypercube. As such, our holdout datasets are currently formed by taking a 5% sample of the hypercube, stratified across LAs to maintain LA distributions. In the following section we discuss the resultant model fitted weight distributions, with their respective metrics, and how fractionally counting individuals affects the population estimates.

5. Discussion

In this section we discuss the resultant weight distributions of each of the available models (LOGR, SVM and RF) in Figure 2, a summary of the performance of each model by comparing their ROC curves, and comparisons between the predicted weights for each of the models when testing on our holdout hypercube proxy samples and the "true" population weights. In terms of pure classification potential, both the LOGR and RF models performed similarly well, with the fitted LOGR metrics <~3% of RF across all metrics, and with both outperforming the SVM both in terms of the raw metrics recorded as well as computational complexity (runtime). The runtime cost is also a key consideration, as the time per fit and prediction for each SVM simulation is significantly (~30-50x) greater than that for RF, with LOGR taking even less time again to complete each simulation. Even when considering the cumulative weight predictions (Figure 3) LOGR and RF perform similarly in capturing the higher-level population weights while SVM, although appearing to return weights on-average closer to the true weights, appears to be qualitatively worse.

Figure 2 shows the resultant address weight distributions for each model ((a) – LOGR, (b) SVM, (c) RF). Along the x-axis are the model predicted weights, with the y-axis indicating the density. For comparison, if we assumed an address was either entirely "true" or "false" we would expect the weight distributions to be centred at their "truth" state with zero deviance, with "true" (blue) records all weighted 1.0 and "false" records all weights 0.0.

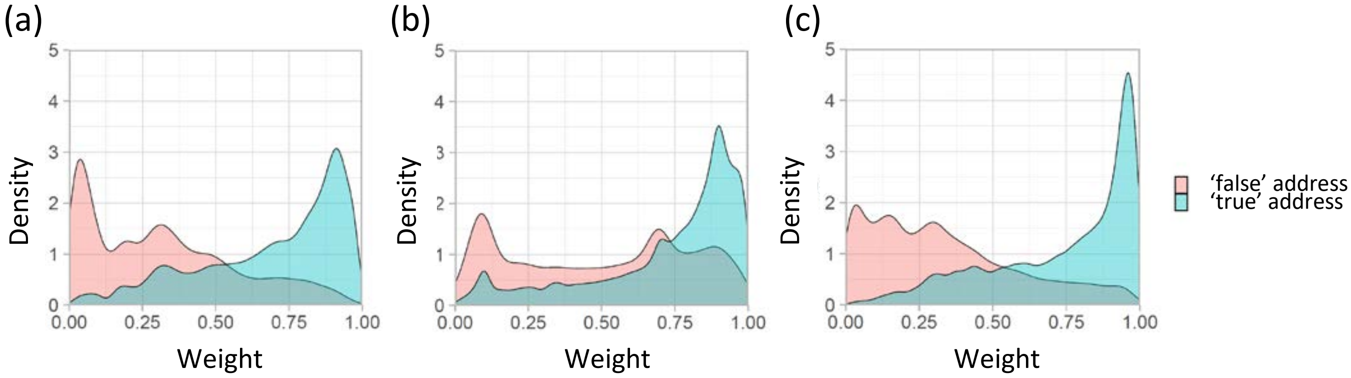


FIGURE 2: (A) LOGR (B) SVM (C) RF PREDICTED WEIGHT DISTRIBUTIONS WHEN FITTED TO SIMULATED PCS, COMPARING MODEL PROPENSITY FOR WEIGHTING RECORDS IDENTIFIED AS BOTH "TRUE" (BLUE) AND "FALSE" (RED) ACCORDING TO CENSUS SPECIFIED ADDRESS

We find, across all fitted models, skewed weight distributions with peaks towards their respective truth states (weight = 0 / 1), tailing off towards the opposite state, with the peak for "true" records greater (+~0.15 LOGR to ~2.2 RF) than the "false" records. However, each model has different propensities for allocating weights throughout the range for both "true" and "false" records, with the distributions for each respective state in the LOGR fit (Figure 2(a)) appearing to be qualitatively symmetrical, with similar densities for each record type towards each respective peak. Looking at the SVM fit (Figure 2(b)) we see similar peaks at the truth states; however, the peaks have shifted further from the extremes (at 0 & 1). Additionally, with the SVM fit there appears to be distinct peaks at both ends of the distribution, potentially as a result of the non-linear nature of SVM fitting, which is not present in either the LOGR or RF fits. The resultant weights distribution from the RF fits (Figure 2(c)) is qualitatively similar to that of the LOGR fit, however with the peak at 1.00 increased and the peak at 0.00 reduced. This suggests the model has a greater propensity for weighted the "true" records more heavily towards 1.00, with the weights for "false" records more uniformly distributed throughout the range.

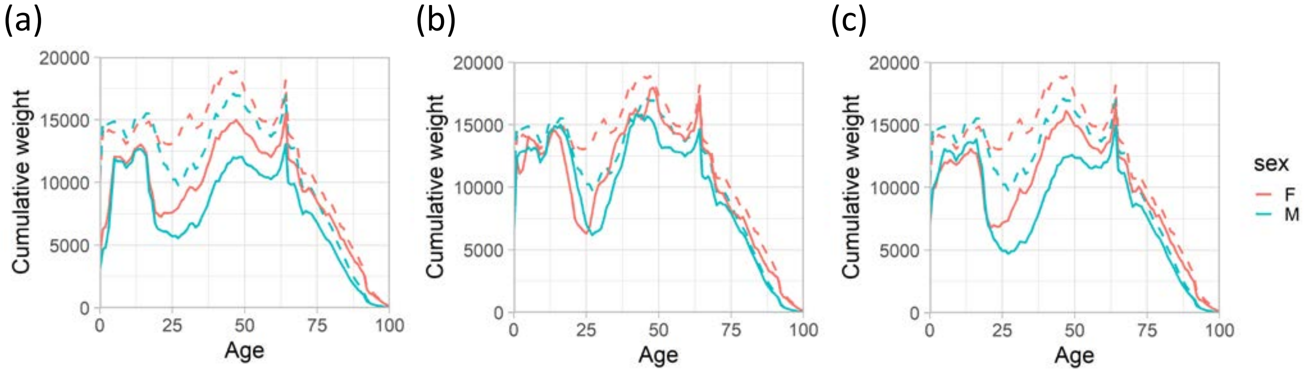


FIGURE 3: COMPARISON BETWEEN CUMULATIVE PREDICTED WEIGHTS (SOLID LINES) AND TRUE POPULATION WEIGHTS (DASHED LINES) BROKEN DOWN BY SEX (FEMALE – RED, MALE – BLUE) FOR (A) LOGR (B) SVM (C) RF

Figure 3 shows the resultant estimated population weights (solid lines), broken down by sex (female – blue, male - orange), compared to the "actual" weights when considering only the "true" Census address records (dashed lines). Figure 3(a) shows the predicted weights using LOGR, Figure 3(b) the weights using SVM, and Figure 3(c) the results of RF fitting. Each of the models appears to under-weight the population, by approximately 20% for all ages except those in the range 20 – 40 years, where the under-weighting is approximately 45%. The broad underweighting may be due in part to the dilution of the dataset through balancing with simulated data. Effectively introducing a broad propensity to underweight which becomes apparent when the models are used to predict on the holdout dataset where the "true" address record class is the majority again. The discrepancy for individuals 20 – 40 years old may also be explained by the weaknesses in our current hypercube prototype created using the DI. As the earliest available linked PDS dataset was from 2016 we are missing those records created in the meantime, whilst also potentially misplacing individuals who are present on the 2011 Census but at a different address than that indicated in the PDS dataset. The potential for greater churn and misplacement on administrative records for individuals within this age range could potentially explain the underweighting seem here, especially as a major benefit of the PDS dataset is its increased frequency of update compared to the PR. It could be expected that if the delay between the reference date (Census 2011) and admin data (PDS, 2016) was reduced we would also reduce this weighting discrepancy. Additionally, the introduction of CIS data, previously shown as a good indicator of correct address especially for working age individuals, would likewise be expected to improve this discrepancy.

Comparing the different models, it appears that qualitatively the LOGR and RF fits produce broadly similar weight estimates, broadly allocating lower weights with greater emphasis for ages 20 – 40. However, whist they appear to respond worse than SVM in terms of reproducing population weights, when considering how well they qualitatively capture the true weights they appear to perform better than SVM. The SVM estimated weights appear more oscillatory, less capturing the qualitative dynamics and instead potentially over-fitting to the training sample. The resultant estimates from LOGR and RF appear like they might be good candidates for benchmarking to the true weights whereas the SVM weights might not be so easily benchmarked. As previously mentioned, fitting LOGR and RF and subsequent prediction on a test dataset was significantly more efficient than SVM, with runtimes being longer for smaller training sets (~0.1% Census) and intangible at the sample sizes we would want to use it for. Whilst there is evidence that SVM modelling does manage to produce estimates with similar accuracy to the other methods, occasionally being second to RF in training and close behind in testing, the issue in scaling up combined with the relative accuracies of the alternative methods suggests we may be better off focusing on LOGR and RF for future model development.

In summary, whilst we could consider each of the models for further development based on the pure metrics and resultant predicted aggregated weights, with each potentially outperforming the others when considering different subsets of the qualitative and quantitative measures we applied. However, prefaced by our future aim to apply these methods to an extended hypercube of administrative data with the potential to continue to grow larger with each additional dataset added and covariates derived, continuing with SVM becomes increasing intangible. This is in large part due to the increased runtime relative to both alternative models, especially when scaling up the size of the datasets involved, compounded by the lacklustre metrics and potential for overfitting. Taking this into consideration, potential future development would focus on LOGR and RF as candidate models for predicting address weights for use in fractional counting.

6. Future work

Stage 1: Proof of concept

To date we have focused on producing a prototype of the hypercube for the key population characteristics (age, sex and location of residence), as outlined in steps I to IV in Section 4 and Appendix 1. This has involved constructing an initial extended administrative dataset for 2011 and estimating initial weights for the hypercube based on comparisons with 2011 adjusted census (representing the "true" population). The aim with this initial stage of the project is to provide proof of concept and to inform the direction of future work plans.

Progress to date looks promising but has been limited as we have not been able to obtain all the key administrative sources for 2011, so whilst we have been able to explore and train potential models, we have not been able to fully evaluate these. To complete these stages, we will also need to consider the best methods for classifying "usual residence", the definition used for official population outputs. We will continue to shape our research according to the research findings under the PMST programme particularly in developing the ABPEs and ACID projects.

Stage 2: Rolling and expanding the hypercube

Assuming proof of concept, the second stage of the project will investigate how we roll the hypercube forward on a continuous basis over time, this will include the use of continuous ONS surveys with updated administrative data via an incremental process and assess the relative benefits of parametric versus algorithmic approaches.

As more administrative data becomes available and our understating of their coverage and potential use for capturing census like population characteristics develops, the hypercube can be extended to include additional dimensions. In doing this, we will consider the benefit from making use of wider combined administrative and survey data as well as other open data sources.

Stage 3: The methods for auditing

We will need to consider how to measure uncertainty in the hypercube and the stability of the models over time. It is envisaged that a purposely designed coverage survey will provide a periodic quality review (audit) of the hypercube; key questions are whether a purposely designed coverage survey will be sufficient and the required sample size of this. Alternatively, would a large-scale collection more similar to a full or partial census be needed.

Additionally, we will need to consider how we measure and communicate the uncertainty in the outputs derived from the hypercube. Can we estimate the main underlying variance through the model-based components of the hypercube, and can we make use of a simulation set up to understand this.

Further considerations

We will consider how we ensure the coherence of all population statistics. This will include methods to benchmark the more detailed multivariate outputs from the hypercube to higher level population totals; how we produce household outputs (for example, household size and structure); and how we encompass the components of population change (births, deaths and migration).

References

Archer, R. et al. (2020) "Estimating population size without a census" paper for external Assurance Panel, External Assurance Panel Papers, EAP129, ONS.

Bradley, A. (1997) "The use of the area under the ROC curve in the evaluation of machine learning algorithms". *Pattern Recognition.*

Calcagno, V., & de Mazancourt, C. (2010). "glmulti: An R Package for Easy Automated Model Selection with (Generalized) Linear Models". *Journal of Statistical Software*.

Chang, C.-C. & Lin, C.-J. (2001). "LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, detailed documentation (algorithms, formulae, etc) can be found at http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.ps.gz

Genuer, R. et al. (2017). "Random Forests for Big Data". *Big Data Research.* p 28-46

Kovacs, G. (2019). "smote-variants: a Python Implementation of 85 Minority Oversampling Techniques". *Neurocomputing*.

Louppe, G. et al. (2013). "Understanding variable importances in forests of random trees". *Advances in Neural Information Processing Systems, p 431-439*

Zhang, L, (2019) "On provision of UK neighbourhood population statistics beyond 2021" provided alongside this report

Appendix 1: Steps in developing the initial hypercube (summary from Zhang 2019)

I. Develop the initial extended database from administrative sources using 2011 data and compare the administrative counts with 2011 census counts for each population subgroup to provide weights representing coverage of the combined administrative dataset.

II. Estimate the probability that the addresses recorded on the administrative data are the correct residential address for that individual. These probabilities are obtained using linked census and administrative data. The census address is assumed to be the correct address, so for this core (linked) population predicted probabilities may be obtained using auxiliary characteristic information and an indicator of whether the address on the admin source is correct or not. If we treat the core population subset as a non-probability sample from the admin population data set and assume the selection is non-informative, we can use this core data set to make predicted probabilities for unlinked individuals in the extended admin data base. Note that the fractional counter is unbiased provided that the probabilities for the true address can be correctly determined by the indicators and covariate information (i.e. the model fit is good).

III. Develop an indicator that each individual within the database is usually resident (i.e. part of the target population). Zhang (2019) suggests three possible methods, using a combination of census, census coverage survey and/or sampling from linked administrative and census data. These suggestions draw upon methods used, and the experience of other countries (e.g. Latvia and Estonia). We will consider these alongside the methods applied for the ONS, Admin Based Population Datasets for England and Wales.

IV. Using the fractional counter, the population estimate for each area of interest is the sum of the probabilities (of whether each address recorded by the administrative process is the correct address) multiplied by the (0,1) indicator that the address in the administrative source is within the target area of interest and is part of the resident population. The probabilities are summed for all individuals within the admissible admin database.

V. Characteristics of the population can be easily estimated using the fractional counter where the (correct) value for the variable of interest is available in the administrative dataset. Where the administrative sources record different values for the same individual across sources, the probability of each value being correct may be estimated (as for address) provided there are good covariate data available (which are correlated with the variable of interest). Methods to reconcile other issues with the characteristic information (e.g. definitional issues, coverage issues) are being investigated as part of the transformation program.

VI. Rolling forward the initial extended database and hypercube over time: Updating the model parameters will be nearly continuous over time, similar to incremental learning in machine learning. It is likely that new data would be available for a subset of individuals in the extended administrative dataset over time. The exact properties of the new data would determine their use in the rolling process. For example, depending on the sample design, survey data (with known inclusion probabilities) may be used to refit the model with updated indicators for erroneous enumeration or misplacement whilst administrative data on the other hand may provide updated covariate information. The estimation precision is determined by the size of the updated datasets and is likely smaller than the initial census-based estimates. Potential approaches for the rolling process include:

> Parametric – e.g. logistic regression models where it is assumed there has been no change for individuals without new data and Empirical Bayes Prediction allowing this assumption to be avoided.
>
> Algorithmic – e.g. Decision tree in which part of the updated observations are used for training the model and part for validation.

VII. Periodic auditing of the estimated population totals: It is envisaged that a purposely designed coverage survey could be used in an audit-assisted approach that validates the underlying extended population database periodically whilst the continuous ONS surveys are used with the admin data in the incremental rolling process. Zhang (2019) notes that previous investigation of the audit sampling inference approach for statistics based on big data, demonstrated negligible variance of the point estimates compared to their potential bias and therefore rendered the conventional hypothesis testing inadequate. He proposes a novel accuracy measure which also has the advantage that the audit sample can be smaller than usually envisaged in a coverage sample.

Appendix 2: Example hypercube provided alongside this paper.

Appendix 3: Summary of work previously undertaken in ONS on modelling residency

ABPE V2

Version 2 of the administrative population dataset for the ABPES included use of a modelling approach to predict the most likely address where this could not be determined using logical processes. The approach was to use a logistic regression model with simulated population coverage survey data, and administrative covariate information (for both the individual and other residents at the same address) to calculate probabilities that the CIS and PR address records were correct.

The simulated survey data was obtained by taking a 1% random sample of 2011 Census records from each local authority. The address reported in the Census sample was assumed to be the correct address for that individual.

Covariate data included:

- number of persons registering at the PR and CIS addresses in the years subsequent to their own registration
- number of persons who the individual shares the same surname with at the PR and CIS addresses
- difference between the CIS address start date and the PR modification date
- a flag indicating evidence of benefit 'activity'3 at the CIS address

This last covariate was obtained from benefits data supplied by DWP and included: the National Benefits Database (NBD), Single Housing Benefit Extract (SHBE) and Tax Credits dataset.

The regression coefficients generated by the model were then used with the relevant covariate values for individuals on the SPD dataset to resolve address conflicts. The individual was assigned to the address scoring the highest probability. A description of methods is available at here.

The analysis reported individuals were up to three times more likely to be resident at the CIS address where they received a state benefit, depending on which benefit this was. It was also found that 85% of conflicting SPD records in 2011 were assigned to the same address as found on the 2011 Census when using the model, although this percentage was found to vary across age groups.

Address Centric Admin Combined Intelligence Dataset (ACID).

The ACID project has been initiated to investigate contingency plans for lower than expected response to the 2021 Census. It aims to model an indicator for correct address from linked admin data and census rehearsal data in order to test potential methods for making adjustments to census data for non-responding households.

The analysis models 2019 Census rehearsal responses linked to the PDS and other administrative data, at the record level. The dependent variable is a binary indicator of whether a link was made between the rehearsal response data and the PDS records. A positive link is taken to indicate that the address information in the PDS data can be considered correct for that individual. The indicator is modelled against covariate data representing the respondent's administrative attributes using a logistic regression model.

Covariate data included:

- indicator where a request that record is removed from PDS data (can be from patient or practise) has been made
- similarity of surname with others associated with UPRN between PDS and Council Tax data (similarity score using Levenshtein edit distance)
- similarity of surname with others associated with UPRN between PDS and English School census (similarity score using Levenshtein edit distance)
- Flag that PDS and ESC have child of same age associated with address
- Difference between dates of data source records

The analysis is at an early stage, but initial results are reported as promising. Further detail of the ACID project is available in the internal working paper presented to the Census Research Assurance Panel, November 2020.

Appendix 4: Modelling covariate definitions and variable names

Covariates currently considered

- Individuals age (*age*)
- Individuals sex (*sex*)
- Source flags (*source_pds, source_pr, source_hesa, source_esc, source_wsc*)
    - True if individuals address identified in admin dataset
- Time since record creation/most recent update (*record_age*)
    - Taken to be time between latest record update & reference date (Census 2011)

- Number of persons registering at the admin address in the years following their own registrations (*new_moves_count*)
    - Count of individuals with more recent updates to admin records at each address for each individual at the address
- Number of persons who the individual shares the same surname with at the admin address (*surname_count*)
    - Count of matching surnames at each address
- The fraction of all records and individual is identified on which contain each conflicting address (*postcode_fract)*
    - Fraction of identified admin datasets that each individual's address in found on
- A flag indicating 'student' status (*student*)
    - True if individual has a recent HESA/ESC/WSC record
- A flag indicating evidence of activity on the PDS database within 12 months (*pds_activity)*
    - True if PDS record update within 12 months
- A flag indicating multiple residents with the same surname (*family*)
    - True if *surname_count* > 1
- Distance between conflicting addresses (km)
    - Distance between subsequently updated addresses
    - Haversine distance (as-the-crow-flies) between pairs of addresses in order of most recent update

Appendix 5: Covariate importance analysis results

The most 'important' features are shown in Table 1, with the apparent least 'important' in Table 2.

TABLE 1: MOST IMPORTANT FEATURES IDENTIFIED FOR EACH MODEL

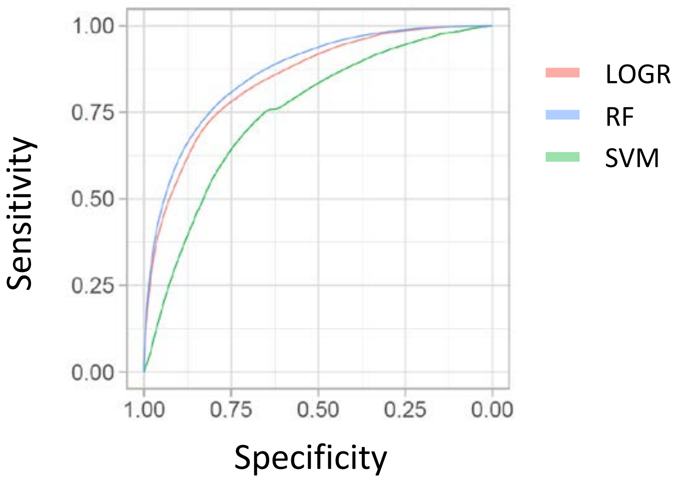
Rank	Logistic regression	Random Forest
1	source_pds	age
2	age	postcode_fraction
3	sex	surname_count
4	postcode_fraction	source_pr
5	source_esc	new_moves_count

TABLE 2: LEAST IMPORTANT FEATURES IDENTIFIED FOR EACH MODEL

Rank	Logistic regression	Random Forest
1	distance	source_wsc
2	pds_activity	pds_activity
3	record_age	distance
4	student	source_hesa
5	source_hesa	student

The output from the automated model selection ('glmulti') running logistic regression shows that the current optimal model includes all the available covariate information, suggesting that there may be inter-variable relationships which are not necessarily clear from the output importance metrics. Additionally, as we will continue to update the covariate selection as further data is acquired, we will continue to retest the model specifications as and when this occurs.

Appendix 6: ROC curves for fitted logistic regression (LOGR), support vector machines (SVM) and random forests (RF) models to holdout dataset



APPENDIX FIGURE 1: ROC CURVES FOR FITTED LOGISTIC REGRESSION (LOGR), SUPPORT VECTOR MACHINES (SVM) AND RANDOM FORESTS (RF) MODELS TO HOLDOUT DATASET

Appendix 6 Figure 1. shows the resultant receiver operator characteristic curves (ROC curves) for the models tested on the holdout population sample, along with the respective AUCs. According the these curves the most effective model for classification purposes would be the RF model, with the greatest sensitivity achieved for a corresponding specificity compared to the alternative models. The SVM returned the least effective and useful fit, with both the sensitivity and specificity of the test fit suffering. This could potentially be a result of overfitting by the SVM model on the training PCS that does not generalise to the holdout set, resulting in unexpected dynamics along the classification boundary (which is shifted to produce the ROC curve). The LOGR fit produced a ROC broadly similar to the RF model's ROC, but with a distinct dip at approximately sensitivity = 0.6 and specificity = 0.8 and so within this region of the metric space the RF model would be more effective than the LOGR model.