Scorecard for summarising progress towards providing attributes that meet user needs in the Future Population and Migration Statistics Programme

Introduction

One of the aims of the Future Population and Migration Statistics (FPMS) programme is to develop approaches for providing more frequent and timely estimates of attributes than the existing system, to better meet user needs. Attributes, (also referred to as Characteristics by the FPMS programme) are the distributions of variables beyond age and sex, such as Ethnicity, Activity Last week, Qualifications, Tenure, Income and Unpaid Carers. These distributions can be required and used at different levels of geography. Users also need a selection of cross-classified attributes, such as Ethnicity by Self-defined Health. Attributes can be associated with various entities such as persons, addresses, households, Communal Establishments and families. The existing population statistics system is mainly served by the estimates that come from a decennial census, with some attributes coming from surveys or administrative data for attributes that the census does not include. Those included on the census can be updated in intervening years in the main through ongoing surveys.

Whilst the census provides estimates with a high degree of accuracy, including for low levels of geography, they become outdated relatively quickly, especially for population attributes that change rapidly. Thus the FPMS programme is exploring how to provide such statistics on a more frequent basis, focusing on the use of administrative data.

Progress on the approaches for each attribute is varied, and there are a lot of attributes to explore. Understandably, each must be examined in turn as the data landscape across them is heterogenous, as is the user needs and uses. This paper presents an attempt to use a scorecard type approach to objectively summarise where FPMS is in relation to meeting user needs across the array of attributes.

The intention would be to complete the scorecard for all attributes being considered by FPMS, and then use the scores to provide a summary analysis of the overall position. This can be repeated every few months to show change in the scores as progress is made. The process of completing the questions contributing to the scorecards is also of benefit to ensure that important aspects are considered on an ongoing basis.

This would be used for reporting to the UKSA board on both the overall position and change over time, to enable them to make informed decisions. The scorecard will be part of the overall decision making process for FPMS, as many other aspects of the programme will

need to be considered. These, and the decision making process itself, are beyond the scope of this paper.

Ask for the panel

The panel are asked for their views on the scorecard approach, in particular whether they feel it provides a good evidence-based summary of the current position for attributes which may come within scope of FPMS.

Qualitative summaries of approaches and progress

Currently, the approach and progress for each attribute is captured in a qualitative summary for each, completed by the team undertaking the work on each attribute. This includes a description of any acquisition, research or other work underway and potential sources and approaches where research has not concluded. The summary of each is provided via a Red-Amber-Green (RAG) rating. This does not really provide a useful summary of whether the approach will meet user needs, as sometimes the RAG status reflects whether the work is on track from a project perspective (rather than an output perspective). However, this provides useful background material for a scorecard, which can be used to help firstly fill it in and secondly as secondary qualitative and explanatory material to sit behind the scorecard.

The scorecard

Overall, the proposed scorecard has been designed to be relatively simple. Further complexity could be built in, but it would become unwieldy and probably more subjective, so for this first attempt simplicity was the objective. The teams doing the research and developing the approaches will use the scorecard, so all of the completion will be based on their judgement.

The scorecard has two sections – the first attempts to summarise the requirement and user need, based on whether there is FPMS work ongoing, whether the attribute is a protected characteristic/political/has high user engagement, the key geographical level required and the priority of the user need. The last of these is subjective, but should be based mainly on the outcomes of the 2023 consultation and any more recent engagement.

Weights are applied to these three areas such that it is scored out of 10, with the largest weight being on user priority (which can score 2 for high, weighted up to 4). The possible scores, weight and max score are shown in Table 1.

Table 1: Requirement scoring

Requirement	Possible scores	Weight	Max score
FPMS work ongoing	0,1	1	1
Attribute type	0,1	1	1
Key geographical level required	1,2,3,4	1	4
User priority	0,1,2	2	4
Total			**10**

The second part of the scorecard attempts to set out progress towards an approach which meets user need. It includes elements for:

- Design existence and maturity
- Data availability and sustainability
- Assurance by MARP
- Dimensions of quality

Whilst some of these are subjective to an extent, they should be able to be evidenced in some form so any challenge to the scoring must be able to be backed up.

These have been weighted so that it can score up to 10, but non-availability of data in a reasonable timeframe is penalized and thus in theory the lowest score is -5. The negative penalty (and potential score) is designed to reflect that even if there is a known solution, but it cannot be delivered for 5 years or more, overall it should have a low score (a good example might be an attribute which 'could' be added to existing administrative sources).

Table 2 shows the scores, weight and max score for the approach section.

Table 2: Approach scoring

Approach	Possible scores	Weight	Max score
Design and Data			
Is there a design	0,1	0.5	0.5
Data availability	0,1	2	2
No data - timeframe	0,1,3,5	-1	0
Data - sustainability	0,1,2	0.5	1
Maturity	0,1,2	0.5	1
MARP assurance	0.1	1	1
Quality evidence			
Timeliness	0.1	0.5	0.5
Geography	0.1	1	1

Frequency	0.1	0.5	0.5
Coherence	0.1	0.5	0.5
Coverage	0.1	1	1
Accuracy	0.1	1	1
Total			**10**

Each of the scoring aspects is discussed below:

Is there a design

Whether there is a written down, well explained and justified design with evidence to show that options have been considered and evaluated, given the user need. The design should be end-to-end and include a description of strengths and weaknesses.

Data Availability

Whether the data for the design is available to ONS at the current time. This may be administrative data, survey data or alternative sources. If the design requires survey data, but that survey has not been implemented then it is considered not available. Equally, if an administrative source is identified as needed for the design but if there is no supply agreement, or that data has not been provided to ONS then it is considered not available.

No data – timeframe

If the data is not available to implement the design (i.e. the answer to the previous question is no) then this question is about the likely timeframe for that data to become available. That may be the time to implement a survey or collect/acquire administrative data. The idea here is that a longer timeframe penalizes the score such that a well specified design may well have been determined, but if it cannot be implement for a long period of time then the scorecard should reflect a low score,

Data – sustainability

If the data is available, in order to meet user needs on an ongoing basis an element of sustainability should be considered. The judgement here is whether the data sources are sustainable over the long term. A survey could be considered more sustainable in some respects than administrative data, because ONS does not own that admin data. However, if there are agreements to limit changes to that admin source then sustainability would be better. Voluntary surveys on the other hand are suffering from reduced response, so efforts to make them sustainable should be considered.

Maturity

This score is meant to reflect the maturity of the design, data and any implementation. If the design has not been implemented then this should score low. If however the design and data have been used to produce outputs, and these outputs show promise and users are engaged in providing feedback then the design could be considered mature. This score can also reflect ONS maturity in using designs of the type for the attribute. New methods can be considered less mature than existing standard approaches.

MARP assurance

This simple question is whether MARP have been asked for advice and positive feedback has been provided.

Timeliness

Does the design meet user need in terms of timeliness – the length of time between the reference point and the production of the statistics. This may not be the case if it take 3 years to produce statistics, or 5 years worth of data is required to be pooled to obtain the output. Timeliness is important when users require up to date statistics that measure change.

Geography

Does the design provide outputs for the key level of geography required by users.

Frequency

Does the design provide the frequency of statistics that users require. This may be yearly or less frequently taking other quality aspects into account.

Coherence (i.e. definition)

Does the design meet the definition that users require. Users may require statistics for an attribute where quite often an administrative source (and occasionally a survey) may not quite provide that concept. For instance, users may require statistics about demand for health services which is not the same as health conditions recorded in administrative data (note to meet that need in the past there has been a self-complete census question on general health). Another example is the breakdowns required by users, for example for ethnic group. Users may require statistics for breakdowns that are not captured by administrative data. This can also be the case for surveys/censuses for small groups.

Coverage

Does the design provide high coverage of the population of interest, or are there likely biases which result in differential coverage across the attribute. The design could include

methods for mitigating such biases (e.g. nonresponse adjustments for a survey, imputation methods for administrative data) so then the question is about the existence of evidence that the biases will be sufficiently small.

Accuracy

Does the design result in levels of uncertainty that are in line with user expectations (i.e. is the coefficient of variation acceptable)? Does the design include methods for measuring and reporting that uncertainty? If a survey is the main part of the design, the sample size (and response patterns) will be the main driver of accuracy. For administrative data, the answer is more complex as uncertainty is harder to quantify but equally coverage will be more important to address.

Scorecard Implementation

The scorecard is implemented in Excel with simple drop downs and automated calculations. The blank scorecard is included in Annex A. Examples of the scoring will be provided through case studies, and the reader is encouraged to use the attachment in the Annex to explore the potential scenarios and scoring.

With the two elements for each attribute, once a large number of attributes have been scored, summary statistics could be used to show overall progress, a time series, frequency statistics and plots of user needs against progress. This will provide useful overall summaries and management information on progress and overall status.

To demonstrate how the scorecard works, here are three case studies. The first is an attribute that has had quite a lot of work, the second is one that has not and the third is an attribute not available on Census.

Case study 1 - Ethnicity

Attribute scorecard				
Description of attribute and solution	Ethnicity derived from linked administrative sources, using the ABC as the spine			
			Score	**Weight**
Programme	Being progressed actively	Yes	1	1
	Is it a protected characteristic	Yes	3	3
User Needs	Priority	High	6	3
REQUIREMENT SCORE			**10**	
Solution	Is there a specified, tested and implemented design for delivering this?	Yes	0.5	0.5
	Is the data for that design available now?	Yes	2	2
	If not, what is the timeframe for that data?	0	0	1
	If Yes, how sustainable is the data?	Medium	0	0.5
	How mature is the design?	Medium	0	0.5
	Has the design been assured by MARP?	No	0	1
Quality	**Is there evidence the solution meets the following quality criteria?**			
	Timeliness	Yes	0.5	0.5
	Geography	Yes	1	1
	Frequency	Yes	0.5	0.5
	Coherence (ie definition)	No	0	0.5
	Coverage	No	0	1
	Accuracy	No	0	1
APPROACH SCORE			**4.5**	

The ethnicity scorecard scores the maximum of 10 on requirement, as it is a protected characteristic and has a high user priority need. It is being progressed by the FPMS team. In terms of approach, there is published work using linked administrative data and the Statistical Population Dataset as the base population. Coverage is about 85%. Whilst it provides timely data with a lag of less than two years, and can produce low level outputs, there isn't not evidence that it would provide all the ethnicity subgroups or accuracy that users require, Thus it scores 4.5 out of ten.

Case Study 2 – Welsh Language

Attribute scorecard				
Description of attribute and solution	Welsh Language			
			Score	**Weight**
Programme	Being progressed actively	Yes	1	1
	Is it a protected characteristic	No	0	3
User Needs	Priority	High	6	3
REQUIREMENT SCORE			**7**	
Solution	Is there a specified, tested and implemented design for delivering this?	No	0	0.5
	Is the data for that design available now?	No	0	2
	If not, what is the timeframe for that data?	3-5 years	-3	1
	If Yes, how sustainable is the data?	Low	0	0.5
	How mature is the design?	Low	0	0.5
	Has the design been assured by MARP?	No	0	1
Quality	**Is there evidence the solution meets the following quality criteria?**			
	Timeliness	No	0	0.5
	Geography	No	0	1
	Frequency	No	0	0.5
	Coherence (ie definition)	No	0	0.5
	Coverage	No	0	1
	Accuracy	No	0	1
APPROACH SCORE			**-3**	

Welsh Language is a variable that is not available in administrative data. Whilst it has a high requirement score of 7, reflecting the importance of the variable to the Welsh Government, very little work has been done on this attribute. As a result the approach is unclear and any data would likely not be available for at least 3-5 years, and so the approach score is negative due to that likely timeframe and unclear approach.

Case Study 3 – Income

Attribute scorecard				
Description of attribute and solution	Person Income using linked administrative data and the ABC as the basis			
			Score	**Weight**
Programme	Being progressed actively	Yes	1	1
	Is it a protected characteristic	No	0	3
User Needs	Priority	High	6	3
REQUIREMENT SCORE			**7**	
Solution	Is there a specified, tested and implemented design for delivering this?	Yes	0.5	0.5
	Is the data for that design available now?	Yes	2	2
	If not, what is the timeframe for that data?	0	0	1
	If Yes, how sustainable is the data?	Medium	0	0.5
	How mature is the design?	Medium	0	0.5
	Has the design been assured by MARP?	No	0	1
Quality	**Is there evidence the solution meets the following quality criteria?**			
	Timeliness	Yes	0.5	0.5
	Geography	Yes	1	1
	Frequency	Yes	0.5	0.5
	Coherence (ie definition)	No	0	0.5
	Coverage	No	0	1
	Accuracy	No	0	1
APPROACH SCORE			**4.5**	

Income is not a variable collected in censuses, but has high user demand due to its use in predicting poverty. So it scores relatively highly in terms of requirement. There has been previous work on income, using PAYE data and various other, mainly DWP/HMRC sources to supplement income data. This linked the data to the SPD and produced individual level outputs, which have been published. The lag is about 1-2 years given the nature of the sources used, with self assessment data the one source which would improve coverage substantially. There was also a pilot looking at whether survey data could predict income, but this was paused. Thus the approach scores a respectable 4.5, with more work likely to lead to improvements.