How accurate is Origins?

FAQ 2

September 2022

Purpose of this note

Many people ask how accurate the Origins classification is and how it handles different types of names. This note addresses the more commonly asked questions.

A: What do we mean by accuracy?

When a person asks, "How accurate is Origins" we need to ask the question "accurate by comparison with what?". What Origins is attempting to infer is the genetic inheritance of an individual, what part of the world did their ancestors originate from.

Clearly this will not necessarily coincide with the answer an individual would give on a self-identification questionnaire. Nor will it match their nationality, nor the country or countries of origins of their parents. As a rule, when compared with self-identification, we believe that Origins is a better indicator of a person's distant ancestry, self-identification a better indicator of their aspiration or identification.

For example, we know that there are very many residents of Glasgow whose Irish surnames mark them out as descendants of former Irish immigrants. Many of these people support Glasgow Celtic and a number would still describe themselves as Catholic. But in the 2011 census they would describe themselves as Scots. Origins would describe them as Irish. The fact that the two classifications give different results is not evidence of "error" – they are different because one is primarily a backward-looking classification, the other a forward looking one.

Whilst the objective of some users of Origins is to establish the ancestry of individual named people, a more common use is to establish the distribution of ancestries across a population of named individuals, in other words, to profile a population as a whole rather than to accurately code a series of individual names. With respect to this use of Origins, the relevant consideration is whether discrepancies between Origins and self-identification are random or systematic. Where they are random, then it is likely that these discrepancies will cancel each other out, where they are systematic, they will not. In our experience, discrepancies between Origins and self-identification are mostly random, the result of which is that they are unlikely to affect the accuracy of a profile. For example, the practice of women taking the surname of their partners is likely to cause a discrepancy between the Origins code of an individual and their self-identification, but this is unlikely

to affect the accuracy of the profile of a population of women, so long as there is no systematic tendency of women of Origins type "W" to be more likely to be partners of men of Origins type "M" than for men of Origins type "M" to live with women of Origins type "W". What is important for profiling is that errors are not systematic, which is why the issue of married women changing their surname, though superficially concerning, does not cause significant error since its impact is random.

B: What is Origins coverage rate?

Providing your files are free of data capture errors, you should be able to code 99.5% of your customer records by Origins type. The residue consists of either names which the system does not recognise, because they are rare, or ones which the system cannot allocate to any particular Origins type because we have been unable to research it.

C: What is Origins' level of consistency with self-identification?

The level of consistency with self-identification varies from one Origins type to another. Origins achieves accuracy rates in excess of 90% in identifying South Asians and Muslims and 70% in identifying Black Africans, Greeks, Armenians and people from East and South East Europe. It achieves accuracy rates of 50% with Hispanics. Lower accuracy rates are achieved with people of Nordic or French origin, with Jews and Black Caribbeans.

As would be expected, the system is more accurate when coding names to general categories, such as South Asians or Greeks, than to specific sub-categories, such as Sri Lankans or Greek Cypriots.

We know errors exist with self-reporting and people can identify differently in different situations. It is not uncommon for people of mixed background to decide not to divulge their ethnicity or to respond inaccurately. For example, our analysis of Kent patient data showed that a statistically significant proportion of people with the surname "Patel" recorded themselves as "White British".

The importance many organisations are now attaching to removing names from CVs, due to potential bias on the part of the assessor, is a salutary reminder that it is as much a person's name as it is their physical appearance or ethnicity that contributes to discrimination. Disparities in the treatment of individuals are more likely to be based on assumptions about their ethnicity than their self-identification; a form of discrimination which Origins is particularly well suited to predict.

D: How does Origins handle persons of mixed ancestry?

Origins can be used to identify persons whose names come from more than one tradition, for example a person with an English personal name and a Finnish family name.

The confidence score given to each name combination can also be used to select or deselect people who are most likely to be of mixed ancestry. Restricting a communication to names with high confidence scores is an effective way of not including individuals who are least likely to belong to the selected target group.

E: How does Origins handle marriage?

Many women adopt their husband's surname when they marry. What are the implications of this for the accuracy of Origins?

From analysis of national data, it is evident that most people marry within their own community and that when a woman marries a man from a different community, it is most commonly, but obviously not always, a man from an Origins type that is relatively culturally close, for example, a woman with a white British surname marrying a man with a French one. As a result, a woman from a white British background will acquire a French surname and hence be mis-classified as French.
This is clearly a source of error if Origins were being used to classify individuals. However, most uses of Origins are to profile populations rather than to code individuals and we assume (rightly or wrongly, we do not know) that the number of women with a white British surname changing their surname to that of a French partner is roughly equivalent to the number of women with a French surname changing their surname to that of a white British partner. If that is the case this source of error will cancel itself out.

We might want also to bear in mind that behaviour is often based on the culture of the family unit, not necessarily that of the individual. A woman of Irish ancestry married to an Armenian man may well pick up many of the consumption habits of Armenians and vice versa. This situation is not dissimilar to the issues market researchers face when they attribute a social class to individuals on the basis of the social class of the "Head of Household" rather than the individual.

Origins can be used to identify instances where the Origins code of the personal name differs from the Origins code of the family name, so it is possible to analyse these individuals as a "mixed marriage" group or indeed to restrict analysis to those people whose personal and family names belong to the same Origins group or sub-group.

It is worth remembering that many instances of difference between the Origins of the personal and family name will occur among children whose parents have originated from different cultures.
For more information on identifying names from mixed Origins codes refer to FAQ 6.

F: How does Origins handle double-barrelled names?

The Origins classification recognises the identity of 4 million different family names. Many of the names that it does not recognise are what we refer to as "double-barrelled" names. These come in two forms, two separate surnames separated by a hyphen and two separate surnames separated by a space.

If the Origins software is presented with a surname which it does not recognise, it will search for a hyphen or a space. If it finds one, it will identify the text string before the hyphen or space and check whether it has a record of that text string in its reference file. If it does, it will return the code for that name. For example, the surnames "Lloyd-Webber" or "Lloyd Webber" would be coded as though the surname were "Lloyd". Note that if the text string "Lloyd" could not be found the software does not search for "Webber".

G: How does Origins handle people of Black Caribbean descent?

The two minority communities which are most difficult to identify on the basis of names are the Jewish and Black Caribbean communities.

Many Jewish refugees deliberately Anglicised their surnames on arrival in Britain and among the Liberal Jewish community there are many instances of people who bear common British or even German surnames and personal names, such as Rachel or Hannah, which are of Jewish origin but which have been become common choices of non-Jewish parents. As a rule Origins is better at identifying members of the Orthodox Jewish community than it is the Liberal Jewish community.

Most members of the Black Caribbean community bear family names which originated in the British Isles, very often the names of the owners of the plantations on which they worked. We are able to identify a number of distinctively Caribbean family names and personal names. When using Origins to profile geographic areas we deliberately upweight the proportion of names that are coded as Black Caribbean, reducing the proportion that are classified as English, Scottish or Welsh.

The effectiveness of this process is evident when one examples the Postcode Origins dominant ward map which highlights very accurately the areas of London where the Black Caribbean community is most strongly represented.